Economics Department
University of Southampton
Southampton SO17 1BJ, UK

Discussion Papers in
Economics and Econometrics

The Expert and the Charlatan: an Experimental Study in Economic Advice

Theodore Alysandratos (Nottingham Trent University), **Aristotelis Boukouras** (University of Leicester), **Sotiris Georganas** (City University, London), **Zacharias Maniadis** (University of Southampton

No. 2003

This paper is available on our website
http://www.southampton.ac.uk/socsci/economics/research/papers

ISSN 0966-4246

The Expert and The Charlatan: an Experimental Study in Economic Advice

Theodore Alysandratos, Aristotelis Boukouras, Sotiris Georganas and Zacharias Maniadis*

Abstract

How do people choose economic advice? We develop a set of validated questions on economic policy to examine the persuasiveness of expert versus populist advice. Populists, in our context, conform to commonly held beliefs, even when wrong. Two computerized advisers suggest answers to each question, and experimental participants are incentivized to choose the most accurate adviser. Do participants choose the high-accuracy adviser ('the Expert'), or the low-accuracy one ('the Charlatan'), whose answers are designed to be similar to the modal participant's? Our participants overwhelmingly choose the Charlatan. Revealing the Charlatan's exact modus operandi to debias behavior scarcely helps. Sequential feedback on the correct answer improves choices only slowly and partially. Bayesian choice models fail to explain behavior; a naive choice model akin to reinforcement learning with high inertia fits behavior well.

Keywords: Democracy, Economic Literacy, Expert Advice, Populism.

JEL Codes: C91, A11

*Alysandratos: Department of Economics, Nottingham Trent University, UK. Boukouras: Department of Accounting, Finance and Economics, University of Leicester, UK (Corresponding Author). Email: Aris.Boukouras@leicester.ac.uk. Georganas: Department of Economics, City University London, UK. Maniadis: School of Economics, Social and Political Sciences, University of Southampton, UK.

1 Introduction

τότ' ἔφη τὰς πόλεις ἀπόλλυσθαι, ὅταν μὴ δύνωνται τοὺς φαύλους ἀπὸ τῶν σπουδαίων διακρίνειν

Cities fail when they cannot distinguish fools from great men

Antisthenes

Public debate is crucial for the functioning of modern democratic societies. However, as Pericles prominently pointed out in his Funeral Oration, democracy requires citizens who are informed about public affairs, not "idiots", i.e. individuals who only acquire information about their private business. In modern democracies, newspapers and news channels offering curated content ensured some level of information among educated people. Recent technological developments, however, have upset information delivery. Citizens today have access to a wide range of opinions and advice via traditional media as well as via social media, podcasts and websites. In this cacophony, citizens may risk becoming systematically misinformed. Being presented with so much uncurated content, they may be attracted to 'charlatans', i.e. low quality sources of information that are pandering to the audience's preconceptions. This form of populism is of high interest to social science today. The key questions are to what extent laypeople can be manipulated by such communication tactics, whether they are able to distinguish 'experts' from 'charlatans' in the absence of perfect feedback about the accuracy of past advice, and how do they avail of feedback when it arrives.

In this paper we adopt the experimental approach as especially suited to uncover causal relationships and focus on a concrete question: whose advice do the public *choose* to heed? In particular, who does the public regard as an expert when different people juxtapose their views on an important topic? Experts are individuals who dedicate their lives to understanding a topic deeply, and outperform the laypersons in their ability to assess the relevant issues. In real life, scientific credentials, the exhibited confidence (conveyed in body language) and the ability to justify one's point of view in a debate affect perceived expertise. However, in the era of social media some of these factors take a secondary role, since there is (usually) no physical presence and mostly no direct debate. Other metrics, such as the numbers of followers and likes, and the popular appeal of the message itself play a role. In this study we abstract from issues like acquired authority as expressed in social media metrics, choosing to concentrate on a simple key factor that plays a role in digital communication: the public appeal of different views, by themselves, on important topics. To summarize, we ask: when everyone has a voice, whose voice do people listen to and who do they elevate to the status of an 'authority'?

The findings from our experiments indicate that a charlatan espousing popular beliefs can lead laypeople to follow her advice rather than the advice of a genuine expert. This is true even in the face of increasing negative evidence regarding the accuracy of the charlatan. In particular, we introduce a questionnaire on economic policy issues and validate the correct answers. For validation we use a sample of academic economists and select only questions that exhibit a high level of agreement on the correct answer. We also conduct a pilot study among laypeople in order to gauge the most popular answer to each question, i.e. the answer given by most people.

We then employ this questionnaire in a set of three-stage experiments with UK participants. In Stage 1, the participants answer each question on their own, receiving points for each correct answer. After answering each question, they see the suggested answers of two computerized 'advisers' on the same questions. One adviser is the 'Expert', who is designed to give the answer deemed correct by the academic consensus with probability 85% and a random answer otherwise. The other adviser is the 'Charlatan' who always proposes the most popular answer from the pilot study with layperson participants. In Stage 2, each participant sees a summary of the recommendations by both advisers from Stage 1 and they are asked to pick one adviser, who then answers all of the questions on behalf of the participant. In Stage 3, participants go through the questionnaire one last time, with their selected adviser answering the questions for them. This time the participants are given immediate feedback on whether their adviser was correct and they can switch between advisers if they wish to do so.

Our design allows to examine the fundamental questions that we raised above. Note that we deliberately abstract from all other aspects affecting perceived expertise and assume that the only thing people know is the content of the advice. This is done in order to isolate the appeal of the message itself from the way it is conveyed. We demonstrate that, in economic matters, people have a strong tendency to follow the adviser who suggests similar answers to the people's own priors (an insight consistent with Gentzkow and Shapiro, 2006). This leads to frequent incorrect choices, at the cost of a significant foregone payoff, a loss of approximately 40% of the maximum possible earnings in our experiments.

The inability of the participants to choose the Expert does not result mechanically from their lack of knowledge. The majority of our observations come from an experimental environment where participants are fully informed of how the two advisers choose their answers. Participants knew that the Charlatan chooses answers that are popular, but frequently wrong. Sophisticated individuals, even if they do not know much about economics, should realize that the Charlatan's answers are representative of people with their own level of knowledge, which implies that the correct adviser to choose is the one with the *least* common answers with themselves. This rational inference requires no feedback or experience. Yet, the vast majority of participants fail to infer correctly who the Expert is. As Kahneman (2011) argued "Our comforting conviction that the world makes sense rests on a secure foundation: our almost unlimited ability to ignore our ignorance." In our experiments, it seems that participants ignore their ignorance.

Even receiving feedback on an adviser's past performance does not correct the 'ignoring ignorance' bias fully; updating is substantially slower than what a Bayesian model would suggest. While it is known that humans are not good in Bayesian updating (Tversky and Kahneman, 1980), our results are obtained in an environment that exhibits novel characteristics that expand the scope of the findings. First, we do not use an abstract numerical task, but policy-related questions within a natural context. Second, feedback is very strong since correct answers are fully revealed. Third, in our questionnaire participants should be able to infer the appropriate answer using economic logic instead of a complicated mathematical formula. Our results are driven by a failure to disregard previously held beliefs and rationally assess the advisers on their merit given the feedback, not a failure to do maths.

Our primary contribution is to show that in the domain of economic expertise, charlatans can be effective - even if their strategy is public knowledge - because individuals cannot identify with the laypeople that are being pampered by this strategy. This points to a fundamental strategic failure to account for the pampering strategy. To our knowledge, no prior work has examined this phenomenon. Chakraborty et al. (2020) study theoretically the role of experts in electoral competition and contrast it to a populist alternative. Unlike their setting, in our own the interests of the expert and the participants are aligned and there is no potential benefit from picking the charlatan. Ronayne and Sgroi (2018) and Schotter (2003) examine how individuals respond to advice, while we are interested in a different question, employing a design where advisers effectively compete for attention. The literature on fake news is also indirectly relevant. Previous studies show that fake news spread faster than real ones (Vosoughi et al., 2018) and people share these news even though they can tell they are likely not accurate (Pennycook et al., 2019, 2020). Fake news are shared more because they are more interesting or surprising, suggesting interesting potential mechanisms at play when people choose among experts. Jerrim et al. (2019) identify 'bullshitters', individuals who claim knowledge or expertise in an area where they have little to none, as having increased overconfidence, persevering at tasks, and being popular among their peers. Extensions of our design with human advisers could yield insights as to how exactly the 'bullshitters' persevere. In this sense our results also relate to Pennycook et al. (2015) who study receptivity to pseudo-profound bullshit.

Our results also inform the psychology literature. The latter teaches us that citizens who lack the capacity to think deeply about a topic are likely to believe theories and analyses that sound intuitive to their ears. In fact, the well-known Cognitive Reflection Test (Frederick, 2005) purports to distinguish between people who solve problems using the automatic and effortless 'system 1' vs. the slower and more costly and analytical 'system 2' (Kahneman, 2011). If laypeople address matters of expertise on the basis of system 1, populists who pander to them by giving intuitive, but wrong, answers are likely to be more successful. According to Kahneman (2011), when faced with an unknown domain, people are unable to ask the question "What type of information would I need in order to answer this question?" Instead, they use a System-1 heuristic: "Can I make a convincing-sounding story about this?" Our findings seem to validate the predictions of the above literature.

The influential concept of overconfidence is also related to our participants' inability to view themselves as belonging to the population of laypersons in economic matters. Moore and Healy (2008) present a Bayesian foundation for two concrete types of overconfidence: the tendency to believe that one performs better - at an absolute level - than one truly does (overestimation) and the tendency to predict that one is at a higher quintile of performance than one really is (overplacement). While our model does not explain the origin of these two biases, it does predict that they exacerbate the tendency to choose the Charlatan over the Expert.

The paper is organised as follows. Section 2 elaborates further on the importance of economic expertise and the divergence of the views of the laypeople and economists. Section 3 briefly introduces the experimental design and the benchmark theoretical predictions, while Section 4 describes the conducted experiments in detail. Section 5 presents the empirical results and Section 6 concludes.

2 The Importance of Economic Expertise and Economics as a Science

In the USA, public distrust towards experts is considered a strong cultural characteristic (Nichols, 2017; Wills, 2002), while in the UK conservative politician Michael Gove popularized the famous slogan that the public "has had enough of experts". We argue that distrust in economists both exemplifies the general problem of distrust in experts and institutions, and at the same time it has unique features. It is important to first emphasize that the manner in which the public perceives economics problems can have enormous consequences for society (Blendon et al., 1997). In general, economic thinking is central for key institutions and policies of the Western world, such as social insurance, trade policies, and independent central banks. Their influence extends to issues such as policies to deal with the coronavirus pandemic, and in particular in informing policies concerning the coronavirus quarantine measures. However, in modern democracies it is the view of the general public that matters more, because without public support policies cannot be fostered and implemented.

There is a strong divergence between the views and ways of thinking of experts and the general public when it comes to economic affairs, and the difference is systematic (Leiser and Krill, 2017). In fact, this phenomenon is so old that more than a century ago Newcomb (1893) made a few arguments that could easily be made today: that the pubic seems to support detrimental policies for their self-interest, such as restrictions on immigration and trade and opposition to labor-saving technology. There are good reasons why this divide exists, and both economists and psychologists have conducted research in recent years in order to shed light to this phenomenon.

First of all, economics exhibits idiosyncratic characteristics as a science (Arthur, 2000). The economic way of thinking is often counter-intuitive and takes years of training (Colander, 2005). Despite this fact (and unlike many natural sciences) laypersons are prone to fostering personal simplistic theories of the economy without much technical training (Leiser et al., 2010; Dixon et al., 2014; Leiser and Krill, 2017). Moreover, economic issues might elicit emotional, rather than analytic responses (e.g. views on immigration or trade policy), and ideology often exacerbates disagreements (Stantcheva, 2020). The complexity of economic problems (especially in macroeconomics) raises the issue of whether knowledge of economic issues has the same meaning as expert knowledge in the physical or even biomedical sciences (Javdani and Chang, 2019). However, Caplan (2002) shows that economic expertise is meaningful, in the sense that training - rather than other factors (e.g.

demographics) - accounts for differing beliefs between experts and laypeople.

Several cognitive biases have been proposed and empirically verified as responsible for the gap of economic understanding between laypeople and professional economists. Caplan (2011) categorizes four types of systematic biases in laypeople's economic reasoning: anti-market bias, anti-foreign bias, 'make-work' bias (the idea that work itself is valuable and not the fruit of it) and 'pessimistic bias' (the systematic tendency to see the economic future as grimmer than the past). These four biases are also reflected in the results and analysis presented by Blendon et al. (1997). The GBG-heuristic ('Good-Begets-Good', Leiser and Aroch, 2009) explains why laypeople believe in false causal links between pairs of economic variables when they are both perceived as 'good' (tax cuts and employment: Bartels, 2005) or 'bad' (inflation and unemployment: Dräger et al., 2016; inflation and interest rate increases: Andre et al., 2019). Metaphors are also powerful tools that laypeople often rely on in order to transform the incomprehensible into accessible (Oberlechner et al., 2004), but over-reliance on them may lead to significant judgment errors (Krugman, 2010). Finally, one cannot ignore the ability of the human mind to perceive teleological forces behind random events or behind the outcomes of complex dynamic systems (Kelemen and Rosset, 2009; Kelemen et al., 2013). Although conspiracy-based explanations of economic events have not been studied as extensively as the other biases, there is evidence that people have a systematic tendency to ignore general equilibrium effects (Dal Bó et al., 2018), and that they are more likely to attribute economic crises to individual errors than institutional failings (Gangl et al., 2012; Aprea and Sappa, 2014).

It is very important for economists to further study the determinants of this discrepancy and how to reduce it. In the absence of a well-functioning communication between experts and laypeople, populists are likely to fill in the vacuum by pandering to the public's prior beliefs. This, in turn, fosters a vicious cycle of reduced credibility for the economists and an increasing distrust from the public. Secondly, economists may run the risk of thinking in paternalistic terms. Zingales (2020) forcefully argues that economists might fall into the temptation of not making the implications of policies clear to citizens, from fear that if they fully understand those policies they might oppose them.[1]

We contribute to this literature by examining the robustness of the discrepancy in views between laypeople and economists in an incentivised environment with direct feedback. We find that the

[1]For example, the true views of the American public are likely to be opposed to social insurance, but this domain is very difficult to fully understand, and conscious obfuscation likely prevents direct opposition.

discrepancy is pretty sizable and persistent. In addition, unlike previous studies, we examine experimentally how it affects people's choice of representatives: it can be used to manipulate public support by a populist charlatan.

3 Overview of Experimental Design and Predictions

The basic design is a simple three-stage laboratory experiment with the following structure. Participants are seated in computer terminals and do not interact with each other. They are exposed to an economics questionnaire with eleven multiple-choice questions, as presented in our online Appendix.[2] First, in Stage 1 they are asked to answer these eleven questions without any feedback. After answering each question, they observe the proposed answer by two 'advisers'. They are informed that one of them is a high-accuracy adviser (whom we call the 'Expert' in the paper but not during the experiment), who answers on average 85 (or 70) percent of questions correctly, and that the other one, the low-accuracy adviser (whom we call the 'Charlatan' in the paper but not in the experiment), only answers 55 percent of the answers correctly.

In Stages 2 and 3, participants have monetary incentives to detect who the high accuracy adviser is. In particular, in Stage 2 they make a choice of adviser once and for all, and the answers of their chosen adviser will apply to all the questions. Finally, in Stage 3, they have the opportunity (in every round) to update their choice of adviser to apply to the next round, after receiving feedback on the correct answer in the current round. For each correct answer in stages 1, 2 and 3 participants receive 4, 16 and 8 points, respectively. Each point counts for 0.05 British Pounds (approximately 0.06 US dollars).

In order to keep the incentives appropriate in Stage 3 of the experiments,[3] there is an independent draw for each question that determines whether the high-accuracy adviser gives the correct answer or not. The probability of a correct answer is fixed within each experiment to 85% (or 70%). When it comes to the low-accuracy adviser, the answers are fixed ex ante. In particular, for each question the low-accuracy adviser gives the modal answer chosen in a pilot with laypersons run in Greece (see Appendix A.1 for details). These modal answers turned out to be correct for six out of the eleven questions, and the average percentage of subjects that chose the modal answer in each question was 63%. This is a critical design aspect: the less accurate adviser chooses by design the

[2] As we shall see, in essence participants answer each of the eleven questions three times.

[3] Since in Stage 3 there is sequential feedback question-by-question on the correct answer, if the fraction of correct answers of the Expert was fixed to 85%, then after a long sequence of correct answers by the Expert, subjects would have an incentive to avoid using the Expert as their representative, since a wrong answer is due.

question most likely to be chosen by a person drawn randomly from a population of laypersons.

For each question, participants had to pick one out of four answers.[4] One of the answers is correct and the other three are wrong. Since these are not mere mathematical exercises, in order to determine the correct answer, we looked for expert consensus using a set of academic economists from university departments in Europe and the USA. We selected departments of economics based on research performance and randomly chose members who were invited to participate. If they agreed, they simply answered a subset of our questionnaire without any feedback. In order to be validated, a correct answer required 70 percent consensus among these academic economists. The full validation process is described in Appendix A.

We have two main experiments. In the 'low-information' Experiment 1, our participants are informed about the structure of the experiment and the overall accuracy of the two advisers. However, we do not explain to them the process that we used to choose the answers of the low-accuracy adviser. This benchmark treatment corresponds to the natural setting in which the public simply observes prescriptions about the economy (in the media) without knowing how they were formed. In our 'high-information' Experiment 2 participants have all relevant information about the process with which the low-accuracy adviser chose their answers, including the average popularity of modal answers in the Greek pilot (63%). The key question is whether this comprehensive information suffices to make participants realize that they are more likely to have a common answer with the low-accuracy adviser rather than with the high-accuracy one.

We are principally interested in participants' choices in stages 2 and 3. Considering Stage 2 behavior, we experimentally examine how the advisers' suggestions alone drive laypersons' choice of representatives/influencers. Thus, we examine how the opinions of laypeople about economic matters affect whom they choose to influence or represent them. Stage 3 examines the robustness of such potentially biased choice to feedback. How much feedback on the correct answer is needed for participants to select the high-accuracy adviser (if possible at all)?

To address such questions, we build an explicit framework of how participants may view the experimental interaction and choose to behave in our experiments. This framework guided our experimental design, especially with regards to the information that participants would need in the high-information experiment. It also enables us to predict behavior in stages 2 and 3 using simple models of Bayesian learning. We construct and test two such models, which correspond to different levels of information from the participants' side. The first one examines the case where participants

[4]Only one question diverged from this pattern, having five options instead of four.

do not know anything about the method through which the Charlatan selects her answers. We refer to this model as the 'low information' model or simply the 'low-info' model. The second one analyses the case where participants are given full information on how the Charlatan selects her answers. We refer to it as the 'high information' model or simply the 'high-info' model. The two models are presented below.

3.1 The Low Information Model

First, we lay down the set of assumptions. Let us assume a questionnaire with four possible answers for each question, and fix answer 1 as the correct one, without loss of generality. Consider three agents who provide answers to the questionnaire: the participant (p), the low-accuracy adviser (l), and the high-accuracy adviser (h). Let us make the simplifying assumption that the probability of each individual answering correctly each question is fixed and independent across questions.[5] We make the straightforward assumption that the participant has subjective probability $p_p \geq 1/4$ of giving a correct answer in every round, the low-accuracy adviser has analogous probability $p_l > p_p$, and the high-accuracy adviser has respective probability $p_h > p_l$. We also assume that wrong answers are equally likely (this applies to all three agents), and that the participant merely wishes to maximize experimental earnings by choosing the high-accuracy adviser as much as possible.

If a participant has this simple model of the world, which adviser should she choose at stages 2 and 3 of the experiment? We are interested in answering this question through the lens of the above model, keeping in mind that the participant does not know who the high-accuracy adviser is. What she knows is how many common answers with each adviser she has (in Stage 2) and which questions each adviser answered correctly (in Stage 3). Thus, from the participant's perspective, there are two states of the world and she needs to form posteriors for each one of them. The two relevant states of the world are as follows.

- $\mathbf{S_1}$: Adviser 1 ($A1$) is the high-accuracy adviser, adviser 2 ($A2$) is the low-accuracy adviser.
- $\mathbf{S_2}$: Adviser 1 ($A1$) is the low-accuracy adviser, adviser 2 ($A2$) is the high-accuracy adviser.

We are ultimately interested in the participant's posteriors $P(S_1/h)$, where h is her information set, namely the set of all of her answers and all the advisers' suggestions for Stage 2, along with

[5] In reality, the probability of a correct answer may differ across questions. Moreover, note that the low-accuracy adviser's answers are not random, but the participant does not know this. Finally, both correct and wrong answers could be correlated, since the answers to a block of questions may depend on which theory about the world is correct.

feedback on the correct answer for a subset of questions in Stage 3. But for now, let us focus on Stage 2 alone.

3.1.1 Bayesian Updating in Stage 2

Consider the case of a single question for the moment. Denote the participant's answer as α_p, $A1$'s as α_1 and $A2$'s as α_2. Given that there are 4 possible answers, this gives $4^3 = 64$ different possible answer configurations from the three agents. Formally, these configurations are denoted as $\boldsymbol{\alpha} = \{\alpha_p, \alpha_1, \alpha_2\}$. From the perspective of the participant, who does not know the correct answer, these 64 configurations form 5 distinct *events* on the basis of the coincidence of the answers across the agents. These events are as follows.

- *Event 1 (E1):* All three agents give the same answer to the question.
- *Event 2 (E2):* The participant gives a common answer with $A1$ and a non-common with $A2$.
- *Event 3 (E3):* The participant gives a common answer with $A2$ and a non-common with $A1$.
- *Event 4 (E4):* The two advisers give the same answer, but the participant's answer is different.
- *Event 5 (E5):* All three agents select different answers.

For each of the above events the participant calculates the probability of this event conditional on the state being S_1 or S_2. The calculations of these probabilities for all events and states are provided in Appendix B, where we show that the formula for Bayesian updating of the probability of event S_1 is:

$$P(S_1|h) = \frac{\pi_0}{\pi_0 + (1-\pi_0)\cdot(OR)^K} \tag{1}$$

where OR (which is less than 1) is defined as:

$$OR \equiv \frac{p_p \cdot p_l \cdot (1-p_h) + \frac{(1-p_p)(1-p_l)p_h}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}}{p_p \cdot p_h \cdot (1-p_l) + \frac{(1-p_p)(1-p_h)p_l}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}} \tag{2}$$

In addition, π_0 is the prior and $K \equiv$ [number of times $E2$ has been observed]-[number of times $E3$ has been observed]. Using the natural prior $\pi_0 = 1/2$, we can easily see that if $K > 0$ the posterior is greater than 1/2. This implies that a participant who believes the low-info model would use the simple heuristic in Stage 2 of picking the adviser with whom she shares the most common answers. Let us call this the 'Standard Heuristic' from now on. The key question is whether participants behave according to the predictions of this model or according to the high-info model, which will be presented shortly. As we shall see, under the latter model, participants may choose in Stage 2 the adviser with the least common answers with themselves. However, note that in Experiment 1 the information that subjects have corresponds naturally to the low-info model, whereas in Experiment 2 it corresponds to the high-info model.

3.1.2 Bayesian Updating in Stage 3

Now, let us consider the implications of the model for Bayesian updating in Stage 3. Recall that in this stage the participant receives feedback for each question and decides whether to switch to the other adviser for answers or not. The feedback consists of the correct answer and the answer choice of the currently selected adviser. Thus, if the participant remembers the suggestions of both advisers (reiterated in both Stage 1 and Stage 2), she can infer who answered it correctly and who did not.

Therefore, the information set of the participant gets updated as the stage progresses. Specifically, at an intermediate point of Stage 3, the participant evaluates her information on the basis of the correctness of the two advisers' responses for the questions for which she has received feedback. On the other hand, for the remainder questions, for which no feedback is yet available in Stage 3, she continues to evaluate her information on the sole basis of the number of common answers with the two advisers. We will formally derive the implications of this distinction soon.

With regards to the questions with Stage-3 feedback, there are four distinct events to consider for each question. Notation-wise, we distinguish these events from the events of Stage 2 by using $\hat{E}$ instead of E:

- *Event 1 ($\hat{E}1$):* Only $A1$ gave the correct answer to the question.
- *Event 2 ($\hat{E}2$):* Only $A2$ gave the correct answer to the question.
- *Event 3 ($\hat{E}3$):* They were both correct.
- *Event 4 ($\hat{E}4$):* They were both wrong.

Let us consider the case of a single question with feedback. For each one of the above events, the conditional probabilities $P(\hat{E}|S_1)$ and $P(\hat{E}|S_2)$ can be explicitly derived. For example, $P(\hat{E}1|S_1) =$

$p_h(1-p_l)$. The full table of these probabilities is provided in Appendix B. There we also show that Bayesian updating for any part of Stage 3, after observing a history h that contains both Stage-2 and Stage-3 events, can be described by the formula:

$$P(S_1/h) = \frac{\pi_0}{\pi_0 + (1-\pi_0)\cdot(OR)^{K_{-\hat{N}}}\cdot \hat{OR}^{K_{\hat{N}}}} \qquad \text{with } \hat{OR} \equiv \frac{p_l\cdot(1-p_h)}{p_h\cdot(1-p_l)} \tag{3}$$

where $K_{\hat{N}}$ is the number of times that $A1$ gave a correct answer and $A2$ gave a wrong answer in Stage 3, minus the number of times that $A2$ gave a correct answer and $A1$ gave a wrong answer in Stage 3. $K_{-\hat{N}}$ denotes the number of times that $A1$ alone had a common answer with the participant minus the number of times that $A2$ alone had a common answer with the participant in the questions that have not yet been included in Stage 3.

3.2 The High Information Model

Recall that the 'most popular answer' in the Greek pilot questionnaire was on average selected by about 63% of these laypersons, and this answer was used as the low-accuracy adviser's recommendation in our experiments. In Experiment 1, participants know nothing about this fact, but only know the accuracy of the two advisers. In Experiment 2 participants have full information about the manner in which the two advisers choose their answers. Accordingly, the low-info model applied to Experiment 2 could be considered as a prediction of the behavior of a participant with bounded rationality, who fails to realize that the correlation of their answers with the Charlatan's answers is higher that the correlation with the Expert's answers. We will now provide a model which describes the reasoning of a more sophisticated participant, one who takes into account the information on how the low-accuracy adviser's recommendations were selected and adjusts her posteriors accordingly. Since the participant knows how the Charlatan selects answers, we term this model the 'high-info' model.

Note that although the high-info model is the natural rational benchmark to consider, its predictions run contrary to a powerful psychological tendency to maintain a positive view of oneself (Akerlof and Dickens, 1982) and a preference for self-consistency (Falk and Zimmermann, 2018). In particular, believing that the Expert is the one providing the most similar answers to them allows participants to maintain a view of themselves as being smart and making correct decisions. In a similar vein, a participant with motivated reasoning (see Bénabou and Tirole (2002) and Zimmermann (2020)) will want to believe that they are more similar to the Expert than to the

Charlatan. We will comment in later chapters on how such behavioral models can illuminate aspects of our participants' behavior.

Once more, assume four possible answers for each question and that answer 1 is the correct one. Let $\pi > 1/4$ be the fixed probability, same for each question, that the low-accuracy adviser gives the same answer as the participant. This captures the manner in which the low-accuracy adviser's answers were selected. Here we implicitly treat our participants as members of the same population of laypersons as the participants to the Greek pilot. On the other hand (and as in the low-info model), the high-accuracy adviser and the participant have probability p_h and p_p respectively of selecting a correct answer, with $p_h > p_p \geq 1/4$.

3.2.1 Bayesian Updating in Stage 2

Calculating the posteriors in Stage 2 is very similar to the low-info model. As in section 3.1, there are two states, S_1 and S_2, and the same five events described in page 11. The configurations of possible answers $\boldsymbol{\alpha} = \{\alpha_p, \alpha_1, \alpha_2\}$ fall into these five events and the probability of each event can be calculated once the correlation between the answers of p and l is taken into consideration.

For example, the probability of event $E1$ in state S_1 is $P(E1|S_1) = p_p \cdot p_h \cdot \pi + (1-p_p) \cdot (1-p_h) \cdot \pi/3$. This is derived as follows. $E1$ is the event where the three agents give all the same answer, so that the possible configurations in this event are $\{1,1,1\}$, $\{2,2,2\}$, $\{3,3,3\}$, and $\{4,4,4\}$. Configuration $\{1,1,1\}$ happens when all three give (the correct) answer 1, which happens with probability $p_p \cdot p_h \cdot \pi$. Note that $p_p \cdot \pi$ is the probability that p answers correctly and l gives the same answer. Each one of the other configurations corresponds to a wrong answer and this has probability $3 \cdot \frac{(1-p_p)\cdot\pi}{3} \cdot \frac{(1-p_h)}{3}$. Summing up the two numbers gives the expression above for $P(E1|S_1)$. The probabilities of all other events in all states are computed similarly and they are provided in Appendix B.

With the above calculations at hand, Bayesian updating in Stage 2 is similar as in the low-info model. Given any history h of events, $P(S_1|h)$ is computed by equation (4) below with the corresponding odds ratio given in (5):

$$P(S_1|h) = \frac{\pi_0}{\pi_0 + (1-\pi_0) \cdot (OR_{II})^K} \tag{4}$$

$$OR_{II} \equiv \frac{p_p \cdot \pi \cdot (1-p_h) + (1-p_p) \cdot \pi \cdot p_h + 2/3[(1-p_p) \cdot \pi \cdot (1-p_h)]}{p_p \cdot (1-\pi) \cdot p_h + [(1-p_p) \cdot (1-\pi) \cdot (1-p_h)]/3} \tag{5}$$

Note that the low-info model predicts that a participant with prior 1/2 should *always* pick the adviser with the most answers common with her (following the Standard Heuristic). However, the high-info model is richer, in that it allows p to choose the adviser with the least common answers. When could this happen? Let us assume that $K > 0$, so that $E2$ has been observed more times than $E3$. Accordingly, $A1$ is the adviser with the most common answers with the participant. In order for the participant to prefer $A2$ (the adviser with the least common answers) it has to be the case that the posterior for S_1 is less than half. Assuming the natural prior $\pi_0 = 1/2$, we have that the relevant inequality is $P(S_1/h) = \frac{1}{1+(OR_{II})^K} < 1/2$. This holds whenever $OR_{II} > 1$. By substituting in the value of OR_{II}, we can find the range of values of π for which this inequality holds. Doing so gives us the inequality:

$$\pi > p_p p_h + (1 - p_p)(1 - p_h)/3 \tag{6}$$

For instance, let us set $p_h = 0.85$ (as in most of our high-information experiments) and $\pi = 0.63$. This value of π is the average popularity of the modal answer in the Greek pilot (the low-accuracy adviser's answer) which participants know in the high-information treatments. Let us also assume that $p_p = 0.47$, meaning that the participant believes that she is as accurate as the average participant in our experiments. In this case, the left-hand side of inequality (6) is equal to 0.63 and the right-hand size is 0.426. As a result, we see that for parameters corresponding with our experimental conditions, this model predicts that rational and fully-informed participants should choose the adviser with the least amount of common answers. The condition in (6) is very intuitive, since π is the expected fraction of answers common with the low-accuracy adviser and $p_p p_h + (1 - p_p)(1 - p_h)/3$ is the expected share of answers that are common with the high-accuracy adviser. Whenever the former is greater, the participant rationally infers that the adviser with the most common answers with herself is most likely the low-accuracy adviser, hence choosing the other adviser.

Notice that in our experiments the only parameter that differs across subjects is p_p, the perceived accuracy of the participant. Participants with different confidence in their economic knowledge could choose differently. Define $T \equiv p_p p_h + (1 - p_p)(1 - p_h)/3$. For values of p_p between 0.05 to 0.6, T is less than 0.63 (meaning that the participant should choose the adviser with the least common answers), while for values of p_p greater than 0.6, T is greater than 0.63, so the participant should follow the Standard Heuristic. In other words, overconfident participants could still choose the low-accuracy adviser, despite using the correct model of the world to analyze the problem. This is interesting, because the model predicts that overconfident people are easier targets for populists. We delve more deeply into this issue in our data analysis.

3.2.2 Bayesian Updating in Stage 3

Stage-3 Bayesian updating in the high-info model is slightly different than in the low-info model. This is because the probability of the low-accuracy adviser getting a correct answer in this model depends on whether the participant got the correct answer or not, and so there are now eight possible cases with regards to the feedback that the participant can receive in a given period of Stage 3. Before we present these cases, notice that the probability of the low-accuracy adviser being right, conditional on the participant being wrong, is $(1-\pi)/3$. This is because l picks a different answer from p with probability $1-\pi$ and, conditional on p being wrong, l picks the correct answer from the remaining three with probability $1/3$. Also, the probability of l being wrong, conditional on p being wrong, is $1-(1-\pi)/3 = (2+\pi)/3$. The eight possible events of Stage 3 in the high-info model are as follows.

- *Event 1 ($\tilde{E}1$):* Everyone (participant, $A1$ and $A2$) give the correct answer.
- *Event 2 ($\tilde{E}2$):* Only the participant is correct.
- *Event 3 ($\tilde{E}3$):* The participant and $A1$ are correct, $A2$ is not.
- *Event 4 ($\tilde{E}4$):* Only $A1$ is correct.
- *Event 5 ($\tilde{E}5$):* The participant and $A2$ are correct, $A1$ is not.
- *Event 6 ($\tilde{E}6$):* $A1$ and $A2$ are correct, the participant is not.
- *Event 7 ($\tilde{E}7$):* Only $A2$ is correct.
- *Event 8 ($\tilde{E}8$):* They are all wrong.

The probability $P(\tilde{E}|S)$ of each event in each state is provided in Appendix B. Moreover, events $\tilde{E}1$, $\tilde{E}2$, $\tilde{E}6$, and $\tilde{E}8$ have the same probability in both states, and so their odds ratio is equal to one. Thus, the only informative events are $\tilde{E}3$, $\tilde{E}4$, $\tilde{E}5$, and $\tilde{E}7$, which are paired in terms of odds ratios. $\tilde{E}5$ has the inverse odds ratio of $\tilde{E}3$ and $\tilde{E}7$ has the inverse odds ratio of $\tilde{E}4$. The odds ratios for $\tilde{E}3$ and $\tilde{E}4$ are derived in Appendix B and are given below.

$$OR_{III} \equiv \frac{P(\tilde{E}3|S_2)}{P(\tilde{E}3|S_1)} = \frac{\pi(1-p_h)}{p_h(1-\pi)} \tag{7}$$

$$OR_{IV} \equiv \frac{P(\tilde{E}4|S_2)}{P(\tilde{E}4|S_1)} = \frac{(1-\pi)(1-p_h)}{p_h(2+\pi)} \tag{8}$$

Intuitively, OR_{III} applies to the case where only one adviser and the participant are correct, while OR_{IV} applies to the case where only one adviser is correct. Let us now define $k^1 \equiv$ [number of times $\tilde{E}3$ has occurred in Stage 3] $-$ [number of times $\tilde{E}5$ has occurred in Stage 3], and $k^2 \equiv$ [number of times $\tilde{E}4$ has occurred in Stage 3] $-$ [number of times $\tilde{E}7$ has occurred in Stage 3]. Again, $\hat{N}$ denotes the set of rounds for which feedback has been provided in Stage 3 and $-\hat{N}$ denotes the set of rounds for which feedback has not been provided in Stage 3. Therefore, $k^i_{\hat{N}}$ denotes the measure $k^i, i \in \{1, 2\}$ applied to the rounds with Stage-3 feedback, and $K_{-\hat{N}}$ denotes the measure K (as defined for the low-info model) applied to the rounds without Stage-3 feedback. Overall, the final updating formula for Stage 3 of the high-info model is:[6]

$$P(S_1/h) = \frac{\pi_0}{\pi_0 + (1-\pi_0)\cdot (OR_2)^{K_{-\hat{N}}} \cdot (OR_{III})^{k^1_{\hat{N}}} \cdot (OR_{IV})^{k^2_{\hat{N}}}}. \tag{9}$$

4 The Experimental Sessions

We run a series of experiments, with the key manipulation pertaining to the information provided. We also conducted some additional checks, such as varying the accuracy of the high-accuracy adviser from 85 to 70 percent, and examining the role of financial incentives. The timeline of all the experimental sessions is the same (Figure 1) and let us describe it in detail. In Stage 1

[6] The specification of the model that we describe here assumes that participants do not update their beliefs about how accurate they are in answering questions (as denoted by the parameter p_p) after they receive feedback in Stage 3. The high-info model could be extended to allow for this updating. We refrained from doing this in order to keep the benchmark model as simple as possible. Moreover, the extension would be immaterial for all our main results. As we show in section 5, even in the extreme case where participants have exact (i.e. correct) beliefs about their accuracy in answering questions, the gap between participants' actual choices in selecting experts and the choices predicted by the high-info model remains (see figure 3 on page 25). Because this extension would give very similar predictions for the last questions of Stage 3 to the high-info model where p_p is calibrated according to the participants' actual accuracy (as is the case for figure 25), it would not improve prediction accuracy, while sacrificing parsimony.

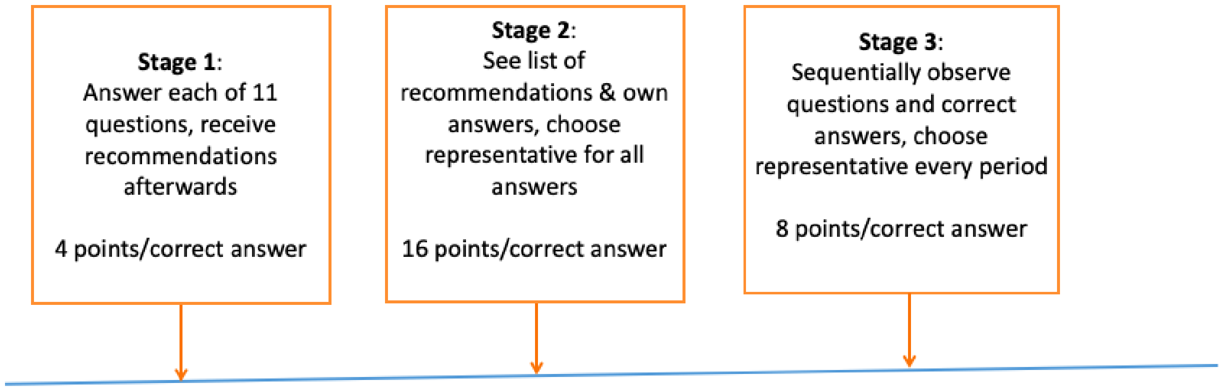
participants answered eleven questions sequentially. Each correct answer carried a prize of 4 points (there was no punishment for wrong answers at any stage). Participants were informed of the existence of two advisers labeled 'Adviser X' and 'Adviser M'. In each round, after choosing their own answer, participants were informed of the answers that the two advisers suggested for the particular question (although they could not use this knowledge any more for this stage). After they answered all eleven questions Stage 1 ended.

In Stage 2, participants were first presented with an overview of the answers provided by themselves and the two advisers for each question of Stage 1. They then had to choose a representative among the two advisers, whose answers would be used for all questions and apply to the participant's payoffs. Each correct answer at this stage carried a prize of 16 points. In other words, in Stage 2, the selected adviser answered the same eleven questions from Stage 1 and the participant earned points based on the performance of the adviser. Hence, participants were financially incentivized to pick the adviser whom they thought had the highest accuracy among the two.

Finally, in Stage 3 participants had the chance of using the advisers in order to answer the same set of questions one last time. However, in this stage the correct answers were revealed sequentially after each question. In particular, each participant was informed about the correct answer for the immediately previous question and the chosen adviser's answer for that question. After this feedback on each question, participants chose the adviser who would answer the next question. Participants could switch advisers as many times as they wished. Each correct answer in this stage was worth 8 points. The selected adviser from Stage 2 was used to answer the first question of Stage 3, after which the participant could choose between her current or the other adviser.

After the main part of each experimental session, participants answered an additional questionnaire. We collected additional information on several dimensions, namely cognitive reasoning (CRT), strategic sophistication (using the 'undercutting game' from Georganas et al., 2015), psychological indicators (measures of confidence and egocentricity), political attitudes (especially towards redistribution) and demographics. Our three-stage design maximizes the information that can be elicited from the given set of questions. First we get the the participants' personal answers, then elicit the choice of representative without feedback, and finally we are able to examine the role of feedback.

Figure 1: *The timeline of the experiments.*



4.1 Experiment 1

Our first experiment emulates a natural setting in a strong sense. In real life, the public observes the opinions and suggestions of influencers regarding important matters, without knowing the process underlying these views and suggestions. This means that if there are populists pandering to the public to gain its support, this is not overt. Accordingly, in our first experiment participants were only informed of the overall accuracy of the two advisers (they did not know which one of X or M was the high-accuracy adviser, of course). No information regarding the exact manner in which answers were chosen for the Charlatan was provided. We argue that this information environment captures realistically the chances that the Charlatan has of gaining public support in the current state of affairs in the public sphere. In this setting, we expect that our simple low-info model makes reasonable predictions for the behavior of participants, since it is natural that they consider that this model describes the actual interaction.

Let us recall that our Stage 1 has 11 experimental choices, Stage 2 has only one choice, and Stage 3 has 11 choices but only 10 of them have payment-related consequences.[7] Our Experiment 1 took place at the University of York (EXEC Lab) in June 2019. 69 participants took part in six sessions. The experiments were designed using z-tree (Fischbacher, 2007), and each session lasted about 40 minutes, with average payoffs equal to £11.5 or approximately $14. In half of the sessions the accuracy of the Expert was 70% and in the other half this accuracy was 85%. Although our preferable specification was the one where the Expert had 85% accuracy (to maximize the treatment

[7] Recall that participants choose an adviser after observing the eleventh (i.e. the last) question of Stage 3. However, since there are no more questions, this choice does not impact their payment. Nevertheless, their behavior is in line with those of the preceding rounds, hence we do not observe an end-game effect.

effect), we were skeptical about whether it would lead to uniform behavior in Stage 3. Therefore, we also used Expert accuracy equal to 70% in some sessions in order to make sure that participants do not choose exclusively the Expert in Stage 3. This concern did not turn out to be relevant, as we will show in our results. The exact instructions for all experiments can be found in the online Appendix.

4.2 Experiment 2

The second experiment took place at the University of Southampton (Southampton Social Sciences Experimental Laboratory) and the University of York (EXEC Lab) during the Fall of 2019 and the Winter of 2020. In this experiment, participants were provided with complete information regarding the manner in which the two advisers chose their answers to the questionnaire. In particular, we provided several examples to illustrate the concept of modal answers, and there was also a quiz to test for participants' understanding. In addition, we elicited their beliefs regarding how many answers they believed they answered correctly in Stage 1 and also their subjective probability that they chose the high-accuracy adviser in Stage 2. In all other aspects, this experiment was identical to Experiment 1 (except the incentives in the Southampton sessions, as explained below). This environment allowed us to test whether complete information on how advisers select answers affects the popularity of an influencer who gives similar answers to themselves. Apart from testing a rigorous Bayesian model, this experiment informs us about the strength of populism and fake news in a controlled environment. In particular, how can we de-bias receivers of news who have strong home-made theories about the economy? Does providing strong evidence about the existence of influencers/politicians who follow populist strategies change who the public chooses as a representative?

Moreover, in Experiment 2 we wanted to examine the role on financial incentives. In particular, Caplan (2011) summarizes very strongly the point that people may have motivated reasoning regarding economic problems. Models of motivated reasoning argue that people attribute real value to holding certain personal beliefs (Bénabou and Tirole, 2002) and that as long as maintaining them is not too costly, they are unlikely to try and correct these wrong beliefs. Varying the level of financial incentives in our Experiment 2 may provide us with insights on the importance of such motivations, and in particular regarding the trade-off between holding cherished beliefs and financial incentives. Moreover, there is a long-standing debate about whether financial incentives make a difference in the behavior of experimental participants, with many scholars arguing that participants are intrinsically motivated and therefore extrinsic financial incentives often make a small difference (Camerer and Hogarth, 1999). Our Southampton laboratory sessions of Experiment 2 did not provide financial incentives to participants,[8] while in the York sessions incentives were identical as in Experiment 1 (20 points translated to £1).

Finally, in Experiment 2 we conducted an additional check, examining whether the order of the questions matters. This is of particular concern in Stage 3, where path dependence may matter significantly in how participants form their beliefs. Accordingly, in the York experiments we examined two different sequences (both randomly generated), and we shall also comment on the importance of the order. Table 1 below summarizes the two experiments and the checks carried out in each one of them.

Table 1: *Characteristics of different sessions in the two experiments.*

	Experiment 1	**Experiment 2**
Treatment	Low-Info	High-Info
Expert Accuracy (baseline)	85%	85%
Location	York (incentivised)	York (incentivised) Southampton (non-incentivised)
Additional Checks	70% accuracy of Expert	No incentives (Ch. 1); Alternative Question Order (Ch. 2)
Sessions	6 (3 for add. check)	10 (4 for Ch. 1; 2 for Ch. 2)
No of Participants	69 (31 for add. check)	203 (54 for Ch. 2) (70 for Ch. 1)

[8] In the Southampton experiments, participants gathered points per correct answer as in all other experiments, but these points did not translate into cash.

5 Results

5.1 Descriptives

We start with some simple descriptive statistics. On aggregate, over both experiments, we recruited 272 participants across both institutions (University of York and University of Southampton). The participants were registered students, mainly undergraduate, although there was substantial participation from postgraduate students and some life-long learning students. Table 2 below gives a precise break-down of the participants from the two experiments across gender, age and study area.

Table 2: *Demographics of participants across the two experiments.*

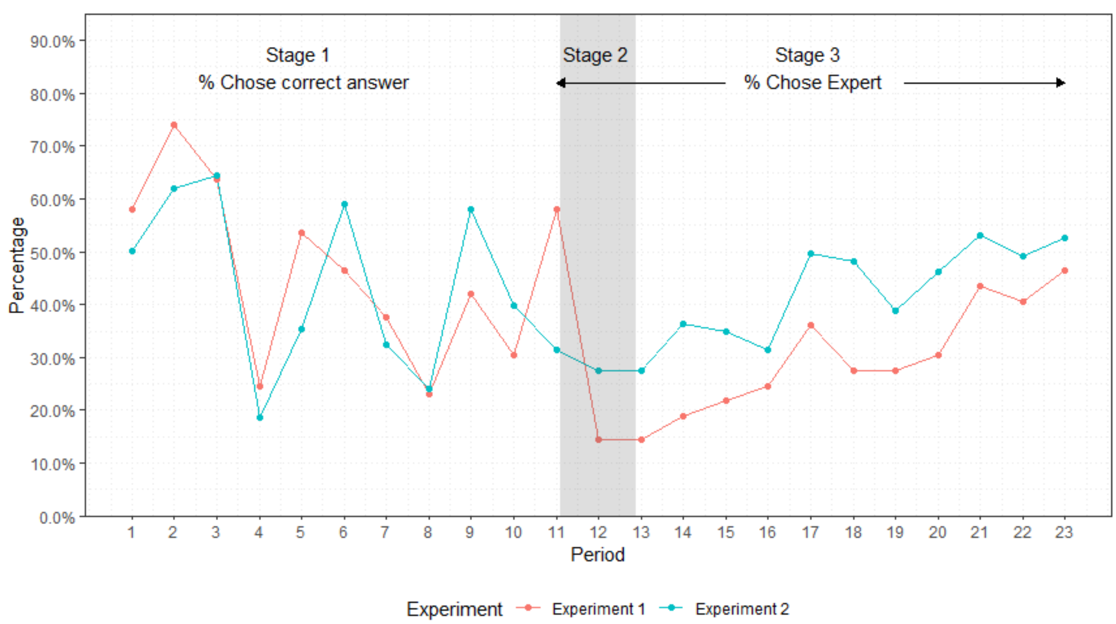
	Experiment 1	**Experiment 2**
Participants By Subject Area		
Economics	12	89
Finance and Accounting	7	2
Other Business	3	6
Other Social Sciences and Humanities	20	64
Positive Sciences	27	42
Participants By Gender		
Female	39	102
Male	27	98
Other	3	3
Participants By Age Group		
18-21	28	139
22-29	32	41
30 and over	9	23

We observe that there is adequate variance regarding the course of study in the participant pool. Although more than a third of the participants (101 in total) came from economics background, a substantial fraction (84 in total) came from social sciences or humanities, and another sizable fraction (69 in total) came from positive sciences. Hence, there is considerable representation of students from non-economic backgrounds. Similarly, we see that gender representation is balanced, while age representation is predominantly between 18 and 21 (167 out of 272), indicating a mainly undergraduate level of study. Overall, the participant pool is quite balanced and with some variation for both experiments across most dimensions.

5.2 Choices

We now look at the frequency of correct answers across stages. This information is summarized in Figure 2 below, where round 13 is by construction the same as round 12 (participants in the first round of Stage 3 inherit their choice from their last choice of Stage 2). In the first stage participants answer less than half of the questions correctly in the absence of feedback. This reflects the fact that the questions are not trivial, the answers are often counter-intuitive, and this seems to hold for various participant groups. On average, participants get 47% of the answers right in Experiment 1 and 43% in Experiment 2.

Figure 2: *Accuracy of Participants' Choices.*



Notes: *The figure shows the percentage of participants that chose the correct answer in each round of Stage 1 and the analogous percentage of those that chose the Expert at each round of stages 2 and 3.*

Recall that participants choose an adviser in Stage 2 without feedback regarding the correctness of the answers. Given that most participants give answers that are similar to the Charlatan's in Stage 1, uninformed choice goes overwhelmingly in favor of the Charlatan. Only 15% of the participants choose the Expert in Experiment 1 and about 28% choose her in Experiment 2 (Wilcoxon test for the difference yields no significance, p-value=0.07).

In Stage 3 we observe some learning, but it is very slow. Even after receiving feedback on 11 questions, close to half of the participants choose the Charlatan in both experiments. In particular, in Experiment 1, only 46% of participants choose the Expert for the last question, and in Experiment 2 this percentage rises slightly to 52%.

5.3 Stage-3 Updating

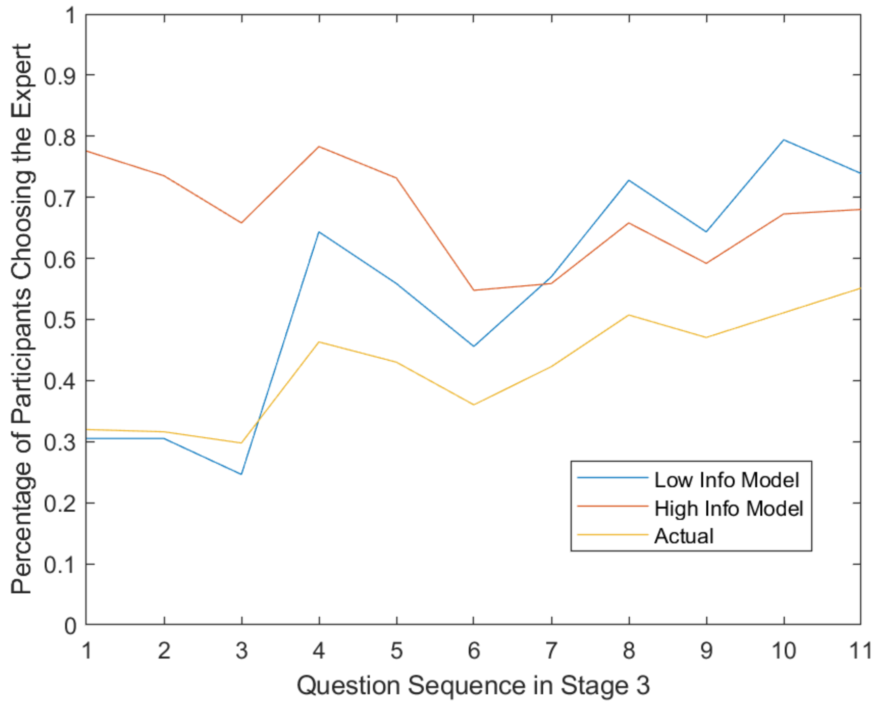
In this section we will have a closer look at updating behavior in Stage 3. Assuming a prior of 50%, our Bayesian models yield exact updating behavior for each participant, given the answer revelations she observes. The other parameters that need to be calibrated are π, i.e. the correlation between the answers of the average participant and the Charlatan, and p_p, i.e. the accuracy of the participant in answering the questions. One can set p_p equal to the actual one or the believed one (according to our belief elicitation). It turns out it does not make a difference, as both specifications for p_p give very similar results, so in this section the results presented use the actual precision.

The parameter π is conceptually more complicated, as it is the probability that the participant gives an identical answer to the Charlatan. It is related to how close the participant believes she is to the median member of the population of laypersons. Since participants are explicitly informed in Experiment 2 about the average popularity of the modal answers from the Greek pilot, we set π equal to this popularity, i.e. 0.63.[9] Given these parameters, we may calculate the optimal updating behavior for each participant individually and can then infer which adviser they should be choosing. We provide a hawk-eye view of the performance of the two models in Figure 3, presenting the aggregate predictions of each of the two models and juxtapose it with actual behavior, pooling together all sessions, including additional checks. 'Aggregate predictions' pertain to the fraction of all players in each round of Stage 3 that choose the Expert. Importantly, if the models fail to explain aggregate behavior then we should not expect them to be very useful for illuminating individual behavior, either.

Participants chose the Expert substantially less often than theory would suggest. Even after 11 occasions of feedback, a bit more than half the participants chose the Expert, while theory suggests this should be done in 70% to 75% of the cases. Excluding aberrant sessions, where the Charlatan actually got in total (weakly) more questions right than the Expert, the difference is even starker. The low-info model predicts that everyone should choose the Expert in the last two questions, while less than 60% actually do so. The particular comparisons pertaining to our additional checks can

[9] But note that it is not clear that the participants' perceived probability is actually equal to this; the Kahneman quote from the introduction actually suggests that few people understand the extent of their own lack of knowledge. Put differently, a participant may not wish to believe that she belongs to the population of laypersons.

Figure 3: *Participants choosing the Expert in Stage 3: theory vs actual.*



Notes: *The graph presents the average adviser choice predicted by each of the two models at the aggregate level, compared to the actual fraction of participants that choose the Expert in each respective round of Stage 3.*

be found in Appendix C.

Overall, the most sophisticated of the two models, the high-info model, has a fit of 50.9% across all participants and periods, meaning that it predicts about half of the participants' choices correctly, while the low-info model does better, at 56.08%, but is still far from perfect. Clearly, some elements of participants' behavior cannot be captured by the Bayesian models, so in the next section we build and estimate an exploratory behavioral model to explain actual choices in Stage 3.

5.3.1 Alternative Behavioral Model for Stage 3

One possible behavioral model in the spirit of fictitious play would have the participants simply count the correct answers each adviser gives in Stage 3 and choose the one with the most correct ones. On top of this count, the participant can give additional weight, ω_{init}, to the initial choice they made in Stage 2. We allow this weight to be different for different participants, reflecting the strength of their beliefs (or stubbornness). We also allow for a common extra weight ω_{cur} on the current observation the participant is making. That is, the Expert's attractiveness for participant i after t questions is simply:

$$E_t^i = N_{correct}^{exp} + \omega_{init}^i \chi_{exp} + \omega_{cur} \chi_t^{exp}$$

where χ_{exp} is an indicator function, equal to one if the participant chose the Expert in Stage 2, and χ_t^{exp} is an indicator function equal to one if the Expert is giving the correct answer to the current question. The Charlatan's attractiveness is equivalently:

$$C_t^i = N_{correct}^{ch} + \omega_{init}^i \chi_{ch} + \omega_{cur} \chi_t^{ch}$$

Participant i, after receiving feedback for question t, selects the Expert if $E_t^i > C_t^i$, the Charlatan if $E_t^i < C_t^i$, and randomizes otherwise. Notice that this model has a close correspondence to a simple reinforcement learning (RL) model (Erev and Roth, 1998). In such a model, players have a propensity p for each option or strategy they can play. In our setup the advisers are the options, and after each correct answer is revealed, the player would update the propensity of choosing an adviser by one unit if the adviser got it right, zero otherwise. Propensities are then converted to probabilities, usually setting the probability equal to the relative propensity. In our case, choice is simpler: participants choose the adviser with the highest propensity with probability 1.[10]The strength of the initial propensities is the free parameter in the basic reinforcement learning model, as in ours. We are also allowing for an extra weight on the current observation, which is related to versions of RL that allow for *recency* or *forgetting*, although our results do not depend on this modeling choice.

We estimate ω_{init}^i and ω_{cur} using an exhaustive grid search up to the second decimal. The estimation yields $\omega_{cur} = 0.2$ and a relatively wide, declining distribution of ω_{init}^i from 0 to 1. This means that about half the participants have low inertia, and switch relatively fast to the adviser with the best record, while the other half have high inertia, with a small percentage even sticking to their initial choice until the end, despite all evidence that the chosen adviser might be the Charlatan.

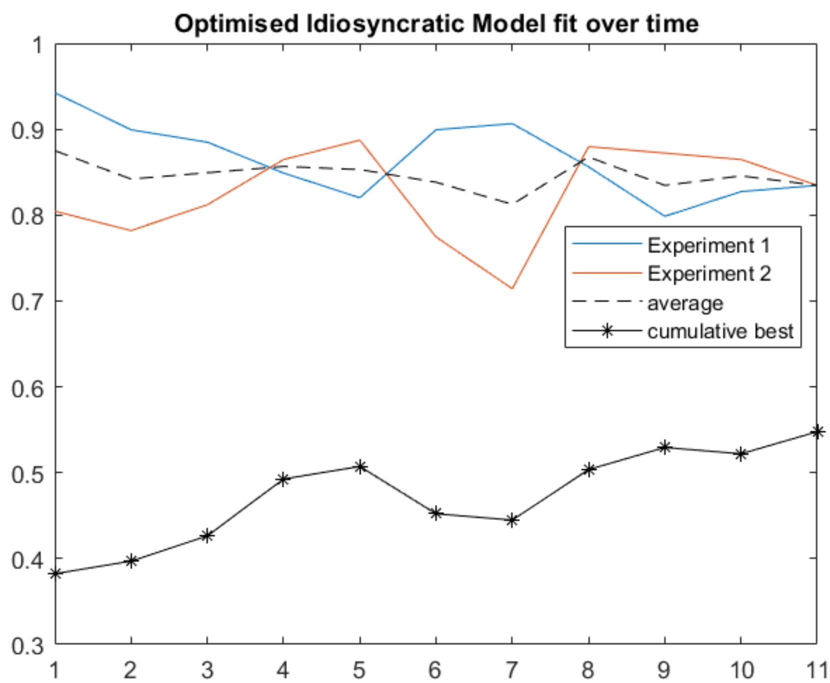
The fit of the model is 0.844, meaning it predicts correctly more than 8 out of 10 participants' decisions across all periods. On the other hand, forcing ω_{init} to be common for all subjects yields an estimated value of 0.6. The impact on the model's fit is not high, as it falls to 0.7332. However

[10]This is similar to versions of RL that allow for cut-off parameters below which the probability of choosing an option becomes zero.

there is a qualitative difference between the two specifications: with a common ω_{init} the model fit falls with time, while the idiosyncratic ω_{init}^i yields a consistent fit across all periods/questions, indicating individual differences in learning speed: some individuals switch fast while others slowly.

Figure 4 shows the model fit over time, for the two different experiments. As a benchmark, the figure also presents the fit of a very simple counting model, where one chooses the adviser with the highest historical number of correct answers, without putting any weight on the initial choice. This benchmark model starts with a low fit and improves with time, but its fit is always substantially below that of our idiosyncratic behavioral model. This shows that the increased complexity of our exploratory behavioral model comes with a substantial benefit.

Figure 4: *Exploratory behavioral model fit over time in Stage 3 of the two experiments.*



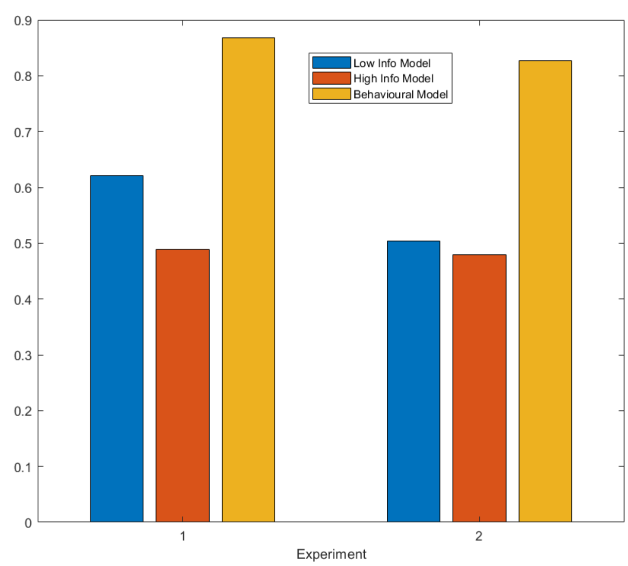
Notes: *The graph presents the fraction of total participant choices that agree with the model predictions in each round of Stage 3 in each of the two experiments. The model is the simple counting behavioral model with heterogeneous initial weights on the adviser chosen in Stage 2. The 'cumulative best' line represents the fit of the simplest possible counting model, where participants would only choose the adviser with the best record, without putting any weight on the initial choice. This line should be compared with the 'average' line, to give us an overall comparison between the simplest model and our idiosyncratic behavioral model.*

5.4 Model Comparison for Stage 3

In this section we compare the fit of the different models. Examining the mere percentage of choices explained by the models (Figure 5), we see that the simple behavioral model fits the data consistently best, explaining 84.7% in Experiment 1 and 84.9% in Experiment 2. The low-info model is somewhat better than the high-info model in both experiments (62.13% and 50.44% against 48.86% and 47.92% respectively).

A more sophisticated approach is to calculate a noisy model, where participants make mistakes with probability ϵ and play what is predicted by each model with probability $1-\epsilon$. We estimate the model parameters for this specification using a maximum likelihood procedure. The ranking of the models is similar to the one using just the fractions of explained choice as a measure of goodness of fit, but we can now compare the fit formally using likelihood ratio tests and the Bayesian Information Criterion. Fitting the behavioral model with idiosyncratic inertia weights brings a jump in the fit, but the Bayesian Information Criterion punishes the high number of parameters in this case. Table 3 reveals that the behavioral model with homogeneous ω is the best according to this criterion. Interestingly, a likelihood ratio test between the two nested behavioral models chooses the unrestricted model as best.

Figure 5: *Comparing the fit of the three models in the two Experiments.*



Notes: *The bars show the fraction of total individual choices, aggregated across all rounds of Stage 3, that agree with the respective model's prediction for the particular participant and round.*

Table 3: *Estimation and goodness of fit of the various models, using all data.*

Model	Low-info	High-info	Beh. CW	Beh. IW
estimated error rate	0.45	0.52	0.28	0.15
LL	-2052	-2074	-1738	-1284
free parameters	1	1	1	273
BIC	4111.4	4155.1	3484.6	4745.4

Notes: *Low-info refers to the Bayesian low-info model, High-info refers to the Bayesian high-info model, Beh. CW refers to the behavioral model with parameter* ω *common across subject, and Beh. IW refers to the behavioral model with idiosyncratic* ω *for each subject.*

Table 4: *Relationship between Stage-1 actual and perceived performance and choice at Stage 2.*

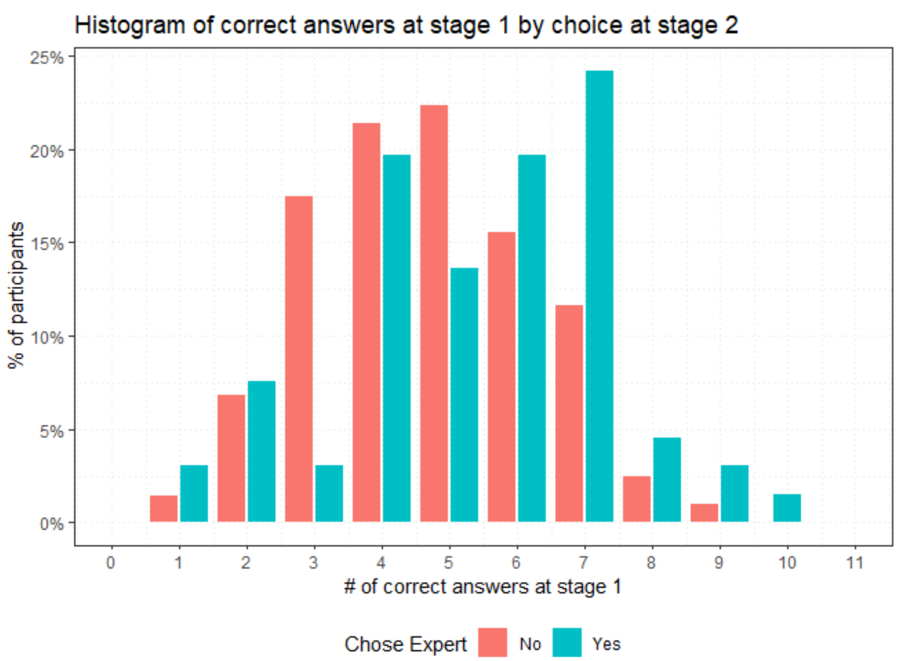
Choice at Stage 2	N	Average number of correct answers	Average perceived correct answers	Average belief in adviser
Expert	66	5.41	5.99	59.7%
Charlatan	206	4.67	5.77	65.6 %
Wincoxon Mann-Whitney p-value		0.000247	0.5456	0.02993

5.5 Examining the Determinants of Behavior at Stage 2

So far we have gotten a view of the general picture: in terms of aggregate performance, the two Bayesian models do not predict behavior well in stages 2 and 3. However, since in Experiment 2 we are eliciting participants' beliefs on the number of their own correct answers, we can have a more careful look into how their perceived performance at the questionnaire affects their choice of adviser at Stage 2. In general, we are interested in having a closer look at the possible heuristics that participants use in Stage 2. Participants' elicited confidence in their choosing the correct adviser at Stage 2 will also be useful in assessing this. We shall use the convention of calling participants who chose the Expert and the Charlatan at Stage 2 'Expert-choosers' and 'Charlatan-choosers', respectively.

Figure 4 provides an overview (for Experiment 2 only) of the relationship between choice at Stage 2, the number of actual and expected correct answers at Stage 1, and participants' subjective probability of having chosen the correct adviser in Stage 2. What we see is that Expert-choosers have a higher number of correct answers in Stage 1, but a similar number of perceived correct answers. Charlatan-choosers were significantly more confident in their choice at Stage 2 relative to Expert-choosers.[11] It is therefore a notable fact that Expert-choosers tended to be less confident, both regarding the number of questions that they answered correctly, and about their choice at Stage 2. A possible interpretation for these results is that confidence across tasks is correlated. Expert-choosers feel less strongly about their capacity in answering the questionnaire and this translates to tentativeness regarding their other choices. Figure 6 shows a detailed picture of the Stage-1 performance of Expert and Charlatan choosers. The distribution of correct answers of Expert-choosers is clearly to the right of the analogous distribution for Charlatan-choosers. This difference is statistically significant with the p-value of the Kolmogorov-Smirnov test being 0.01298. On the other hand, if we were to look into the two distributions of perceived answers they would look identical (Figure 9 in Appendix C).

Figure 6: *Distributions of Stage-1 performance for Expert-choosers and Charlatan-choosers.*



The Pearson correlation between actual performance and beliefs about this performance at Stage 1 was 0.45 and highly significant. The correlation between the number of correct answers

[11] This fact seems surprising but can be explained: if you feel you got many answers right, then you also believe that your choice according to the Standard Heuristic must be correct. If you feel you have many answers wrong, however, choosing the person with the least common answers to yourself feels much less reassuring. It seems that there is something inherently unnatural in choosing the adviser we do not agree with.

and belief in having chosen the correct adviser at Stage 2 was 0.029 for Charlatan-choosers and 0.1 for Expert-choosers (both not significant). However, the number of answers that participants believe they answered correctly is significantly related to their confidence in the adviser ($\rho = 0.4$ and 0.28 for Charlatan-choosers and Expert-choosers respectively, p-values < 0.001 and < 0.027). As an overconfidence measure, we can look at the number of questions that participants answered correctly in Stage 1 minus the number they *believe* they answered correctly (we do not have data on this from Experiment 1). Median overconfidence across Experiment 2 is one, meaning that the median participant believed that they gave one more correct answer than they actually did. Overconfidence is significantly higher for Charlatan-choosers.

It is now time to delve deeper into the drivers of behavior in the high-info Experiment 2. As we saw from the high-info model, assuming that subjects believe that $p_h = 0.85$ and $\pi = 0.63$ (as is common knowledge from the instructions), they should be using the Standard Heuristic if and only if they believe that they have at least seven correct answers. Let us define as 'below-threshold participants' those who believe that they have six or less correct answers, and define the remaining participants as 'above-threshold participants'. Table 5 presents the breakdown of participants who had more common answers either with the Expert or with the Charlatan, and the choice of these participants (participants with a tie were disregarded from this exposition).

The table illustrates that the bulk of participants had more common answers with the Charlatan, as predicted. Overall, 89 out of the 134 participants below the threshold (66.4%) adhered to the Standard Heuristic. On the other hand, 56/69 participants above the threshold (81.2%) adhered to it. The difference is considerable and indicates that at least some participants do take into account their accuracy (as they perceive it) when choosing whether to stick to the Standard Heuristic. On the other hand, the fraction of those who use the Standard Heuristic among below-threshold participants is clearly not zero, as the model predicts. This fact clearly indicates that despite the common knowledge of variables that should make a rational participant drop the intuitive Standard Heuristic, very few people in fact drop it. This indicates the existence of some fundamental deficiency in analyzing this environment in the appropriate strategic manner.

Finally, in the post-experimental survey we gathered a large amount of complementary information about participants and we shall now briefly comment on those. Appendix C.2 contains a series of tables presenting correlation and regression analysis between participants' performance (getting many questions right, either directly or via adviser choice) and these complementary variables. Although the tables also contain the analysis broken down by experiment, we will only comment

Table 5: *Juxtaposing Stage-2 choice to Stage-1 agreement with advisers: Experiment 2.*

	Below-Threshold Participants: 134	**Above-Threshold Participants: 69**
More common answers with Expert	22 out of 134	6 out of 69
Choosing Expert	13 out of 22	3 out of 6
More common answers with Charlatan	99 out of 134	52 out of 69
Choosing Expert	23 out of 99	10 out of 52

on the pooled data, since we have no theory about the possible interaction of these observable variables with the main experimental manipulation (the information condition). As can be seen, high-level courses in technical disciplines and economics are associated with high performance at Stage 1, and the same holds for being male and having scored highly in the classical Cognitive Reflection Test. Interestingly, courses in sociology seem to be associated with good performance in Stages 2 and 3. This may be tentatively interpreted as an effect of sociology in ameliorating the problem of 'knowing one's ignorance' in economic matters. No other variable seems associated with performance at Stage 3. Psychological variables (except the CRT score) appear to have very weak and non-significant effects, while right-wing political orientation seems to have a weak but significant effect on choosing the Expert.

6 Conclusions

We ran a series of experiments with financial incentives to investigate whether the discrepancy of opinions between laypersons and economists is persistent, and the implications for the appeal of populist influencers on the general public. To address these questions we develop a novel questionnaire on economic policy, comprised of 11 questions. As is frequently the case in economics, most of the questions had counter-intuitive answers, while the popular answer is, following Mencken (2012), neat, plausible and *wrong*.[12]

[12] We did not choose only questions with counter-intuitive answers, to avoid the possibility that our participants may recognize the pattern and consciously choose counter-intuitive answers because of this. From their answers in Stage 1 we can safely conclude that this did not happen at any substantial rate.

The first general result is that our experimental participants overwhelmingly chose the Charlatan, who offered popular but frequently wrong advice, over the more accurate Expert. In some sessions less than 10% of participants chose the Expert, while their own responses to the questions were correct much more often, which means that ignorance at the individual question level is amplified when choosing an adviser. This failure to choose the right adviser resulted in significant foregone experimental profits.

Our second general result is that providing information regarding how the Charlatan selects her answers hardly helps. In our high-info treatment (Experiment 2) we explained the Charlatan's *modus operandi* carefully, just short of telling the participants who the Charlatan actually is. Yet, this did not significantly reduce the frequency that they choose the Charlatan compared to the simple experiment where participants lack any information about the Charlatan's thinking (Experiment 1). This is a strong indication of *confirmation bias*. Most participants followed the simple Standard Heuristic, and this strategic choice was not particularly responsive to their beliefs about how ignorant they are of the questionnaire answers. An alternative interpretation that we cannot rule out is that participants might understand they are wrong, but derive direct utility from believing that the adviser with the most common answers with themselves is the high-accuracy adviser (*motivated reasoning*). This may also explain part of the results of Stage 3.

Providing feedback on the correct answers and allowing for learning does surprisingly little in reducing Charlatan-choosers. On the contrary, almost half of the participants stuck with the Charlatan in spite of the strong evidence against this choice. Even after 11 occasions where the correct answer to each question was revealed, only about 55% of participants chose the Expert. This is way below the benchmarks set by the canonical model of learning, i.e. the Bayesian, which predicts between 69% to 75%, depending on the model details. In cases where the Expert was actually more frequently correct than the Charlatan, the theory predicts that the Expert would be chosen 90% to 100% of the time, while participants only chose her in about 60% of the cases.

Note that in real life such perfect and direct feedback is very rare in public policy matters. For instance, even among the academic community, the effect of a minimum wage reform takes decades to be measured properly and in a way acceptable to most economists. In more complicated issues one needs to be even more pessimistic. How many years will it take for the economic effects of Brexit to be cleanly demonstrated and measured? This means that feedback in our experiment was, if anything, too swift and strong. And still often ignored.

What accounts for our results, especially at Stage 3? It seems that the two Bayesian models significantly overestimate the speed with which subjects abandon one adviser for another once they observe them providing wrong answers. Indeed, both the low-info and the high-info model predict that it would take 3 to 5 questions with feedback for the average participant to identify the Expert. This clearly did not happen in our experiments. On the other hand, the exploratory behavioral model, which explains our data much better, suggests that the Charlatan gets a head-start of about 6 questions on average simply by being selected over the Expert at Stage 2 (this estimate is approximately accurate for the majority of participants). This explains why feedback was a weak instrument in improving participants' choices and why, in fact, many ended up choosing the Charlatan until the very end.

Our results have several implications. First, the problem of economic education that Newcomb (1893) lamented on more than a century ago does not seem to have been addressed. When it comes to counter-intuitive topics, it is very hard to provide adequate feedback about the accuracy of economic advice in order to convince the untrained public. Social media need to be used competently by the experts, at the same time as regulation regarding "fake news" is being contemplated. Our evidence suggests that information campaigns to help the public identify charlatans in a decentralized way can hardly help. Most of our participants are unwilling or unable to process this information properly. In Kahneman's words, they are unable to accept their ignorance. Or, to put it differently, they do not seem to come to terms with the fact that, in technocratic issues, what sounds right to them is probably incorrect.

The experimental design in this study can be fruitfully applied to other disciplines that are witnessing a rise in anti-scientific populist movements. Medicine is a good example, given the prevalence of antivaxxers and coronavirus-deniers. Several other disciplines are also experiencing a rising distrust in experts, from biology to even physics. Our study suggests that a common mechanism possibly lies behind these phenomena; more research is needed.

Acknowledgments

Maniadis and Boukouras are supported by BA/Leverhulme small research grant SG162362. We are grateful to John Hey, Paul J. Healy, Roberto Weber, David Cooper, Constantine Sedikides, Santiago Sanchez-Pages as well as seminar participants at the University of Southampton, City University, King's College London, Leicester University, University of Maastricht, University of Durham, ESA Global Meeting in Los Angeles and CRETE 2019 in Tinos, Greece.

A Pilots and validation studies

Since many of our questions involve relatively complex policy scenarios, the correct answer cannot be determined in a purely logical or mathematical sense. For this reason, we employed expert consensus to determine what we consider as a correct answer. To achieve this, we asked a number of academics to answer a set of questions. From those, we chose a priori to include in the experiment the ones that achieve an agreement of at least 70% on any one option.

Our validation process had two steps. First, we chose departments of Economics from which to randomly select economists. For the United Kingdom we used the Research Excellence Framework 2014 (REF 2014), the official research assessment procedure in the United Kingdom. We included all the institutions listed assessed in "Economics and Econometrics", except for City, Southampton and Leicester universities because of potential conflicts of interest with the affiliation of the authors. We chose REF 2014 as that was the most recent assessment available at the time and we ended up including 25 institutions. For Germany, we used the 2013 Handelsblat ranking, accessed on 7 October 2018. We chose the 2013 ranking as it was the closest to the REF 2014 and we wanted to maintain comparability. We included the top 10 departments of Economics from that list. For the rest of Europe we used the Tilburg Ranking for the period 2008-2013, using the journal impact factor to rank departments. From this list we included the top 10 institutions that were not already on our list.

For the second stage of our validation we visited the websites of the Departments of Economics of the institutions we had included in our list and randomly selected 8 members of staff from each Department. This resulted in 360 academics in total. We randomly split the questionnaire into two parts of 8 questions each to increase the response rate. Each academic in our list received an email with a link to one of the two subsets of the questionnaire. In total, we received 43 responses, which amounts to a response rate of 11.94%.

When designing our list of institutions, we had also included the global top 30 Departments of Economics according to REPEC (assessed in October 2019), excluding Columbia Department of Economics and Finance at the Graduate Business School because we did not want to include more than one departments from each institution. Similarly to the European procedure, we randomly chose 8 members of staff from each Department of Economics in our list. In May 2020, we wished to expand our sample of academics after receiving feedback that the number of answers was not large enough. We thus sent an email to each person in our REPEC list containing a link to one of the two subsets of the questionnaire. We received 24 additional responses in total with a response rate of 10%. Overall, we received 66 responses, 31 for the first subset of the questionnaire and 35 for the second subset. Our combined response rate was 11%. Table 6 presents the self-described main field of study of the economists who answered the questionnaire.

Table 6: *Distribution of economists by field*

	Field					
	Microeconomic Theory	Macroeconomic Theory	Econometric Theory	Applied Microeconomics	Applied Macroeconomics	**Total**
Subset A	10	4	0	14	3	**31**
Subset B	10	6	4	15	0	**35**
Total	20	10	4	29	3	**66**

Table 7 presents the responses of the distribution of answers for the 11 questions that we ended up including in the experiments. We must note that when we ran the experiments, we had validated the correct answers according to expert consensus with economists from European-based institutions only. At the time of question choice, option D in question 11 had received 75% of the answers. Hence, it was above the threshold for inclusion. In the end, when we included economists from American based institutions, option D ended up receiving 65.7% of the answers from our overall validation sample.

A.1 Greek pilot

On top of eliciting expert consensus for the included questions, we needed to calibrate the most popular answers among laypeople that the Charlatan would like to emulate. For this purpose, in June 2018 we recruited 73 participants to answer from the participants pool of 'paignia.net', a Greek on-line lab tailored to experimental economics. The participants were volunteers from the general Greek population. For full disclosure purposes, Sotiris Georganas, one of the co-authors of the study, is the founder of paignia.net. We asked the participants to answer all 16 questions that we were considering for inclusion in our experiments at the time. Table 8 presents the distribution of answers for the 11 questions that we ended up including in the experiments (those that passed the 70% threshold in the European part of the validation study). The bold number indicates the 'modal answer' that the Charlatan gave by construction in all of our experiments. The average age

Table 7: *Distribution of answers from the validation exercise with academic economists.*

	Option					
Question	A	B	C	D	E	% of agreement
1	0	**23**	7	1		74.19
2	0	3	**24**	4		77.42
3	1	**30**	0	0		96.77
4	0	0	**29**	2		93.55
5	1	2	**25**	3		80.65
6	3	0	0	**26**	2	83.87
7	2	1	**28**	4		80.00
8	3	2	4	**26**		74.29
9	1	**29**	1	4		82.86
10	**28**	6	0	1		80.00
11	10	1	1	**23**		65.71

of the participants, excluding six who did not want to disclose this information, was 27.8 years.

Table 8: *Distribution of answers in the Greek pilot.*

	Option					
Question	A	B	C	D	E	% of agreement
1	**7**	16	**48**	2		65.75
2	9	**34**	18	12		46.58
3	0	14	**53**	6		72.60
4	11	3	**53**	6		72.60
5	3	16	**45**	9		61.64
6	10	2	2	**59**	0	80.82
7	13	11	**26**	23		35.62
8	9	8	**36**	20		49.32
9	**53**	6	9	5		72.60
10	**47**	14	5	7		64.38
11	3	13	5	**52**		71.23

B Theoretical derivations

B.1 Low Information Model: Stage 2

Table 9: *Low-info model, Stage 2: Configurations of answers belonging to each event.*

Events	Configurations	Total
$E1$	{1,1,1}, {2,2,2}, {3,3,3}, {4,4,4}.	4
$E2$	{1,1,2}, {1,1,3}, {1,1,4}, {2,2,1}, {3,3,1}, {4,4,1}, {2,2,3}, {2,2,4}, {3,3,2}, {3,3,4}, {4,4,2}, {4,4,3}.	12
$E3$	{1,2,1}, {1,3,1}, {1,4,1}, {2,1,2}, {3,1,3}, {4,1,4}, {2,3,2}, {2,4,2}, {3,2,3}, {3,4,3}, {4,2,4}, {4,3,4}.	12
$E4$	{2,1,1}, {3,1,1}, {4,1,1}, {1,2,2}, {1,3,3}, {1,4,4}, {3,2,2}, {4,2,2}, {2,3,3}, {4,3,3}, {2,4,4}, {3,4,4}.	12
$E5$	{1,2,3}, {1,2,4}, {1,3,2}, {1,3,4}, {1,4,2}, {1,4,3}, {2,1,3}, {2,1,4}, {3,1,2}, {3,1,4}, {4,1,2}, {4,1,3}, {2,3,1}, {2,4,1}, {3,2,1}, {3,4,1}, {4,2,1}, {4,3,1}, {2,3,4}, {2,4,3}, {3,2,4}, {3,4,2}, {4,2,3}, {4,3,2}.	24

Notes: *The last column of the table gives the total number of configurations belonging to that event. Recall that a configuration $\{\alpha_p, \alpha_1, \alpha_2\}$ denotes the answers provided by the participant, adviser A1 and adviser A2 respectively. Each question has four possible answers. The full list of events is provided in page 11.*

Table 10: *Low-info model, Stage 2: Probability of each event for every state.*

Events	$\mathbf{P(E\|S_1)}$
$E1$	$p_p \cdot p_h \cdot p_l + \frac{(1-p_p)(1-p_h)(1-p_l)}{9}$
$E2$	$p_p \cdot p_h \cdot (1-p_l) + \frac{(1-p_p)(1-p_h)p_l}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}$
$E3$	$p_p \cdot p_l \cdot (1-p_h) + \frac{(1-p_p)(1-p_l)p_h}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}$
$E4$	$p_h \cdot p_l \cdot (1-p_p) + \frac{(1-p_h)(1-p_l)p_p}{3} + \frac{2\cdot(1-p_p)(1-p_h)(1-p_l)}{9}$
$E5$	$\frac{2}{3} \cdot \left[p_h \cdot (1-p_l) \cdot (1-p_p) + p_p(1-p_h)(1-p_l) + p_l(1-p_h)(1-p_p) + \frac{(1-p_p)(1-p_h)(1-p_l)}{3}\right]$
Events	$\mathbf{P(E\|S_2)}$
$E1$	$p_p \cdot p_l \cdot p_h + \frac{(1-p_p)(1-p_l)(1-p_h)}{9}$
$E2$	$p_p \cdot p_l \cdot (1-p_h) + \frac{(1-p_p)(1-p_l)p_h}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}$
$E3$	$p_p \cdot p_h \cdot (1-p_l) + \frac{(1-p_p)(1-p_h)p_l}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}$
$E4$	$p_l \cdot p_h \cdot (1-p_p) + \frac{(1-p_l)(1-p_h)p_p}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}$
$E5$	$\frac{2}{3} \cdot \left[p_l \cdot (1-p_h) \cdot (1-p_p) + p_p(1-p_l)(1-p_h) + p_h(1-p_l)(1-p_p) + \frac{(1-p_p)(1-p_l)(1-p_h)}{3}\right]$

To illustrate the calculation of the entries of Table 10, let us suppose that the state of the world is S_1 and consider the probability $P(E1|S_1)$. With probability $p_p \cdot p_l \cdot p_h$ all three agents select answer 1, the correct answer, while the probability of all of them selecting answer 2, which is wrong, is equal to $\frac{(1-p_p)}{3} \cdot \frac{(1-p_l)}{3} \cdot \frac{(1-p_h)}{3}$. The latter probability also applies for answers 3 and 4. Thus:

$$P(E1|S_1) = p_p \cdot p_l \cdot p_h + 3 \cdot \frac{(1-p_p)}{3} \cdot \frac{(1-p_l)}{3} \cdot \frac{(1-p_h)}{3} = p_p \cdot p_l \cdot p_h + \frac{(1-p_p)(1-p_l)(1-p_h)}{9}$$

One can easily show that the same probability applies for $P(E1|S_2)$. Using Table 10 we can obtain the odds ratio for any event. Clearly, for events $E1, E4$, and $E5$, $P(E|S_1) = P(E|S_2)$ and so $OR(E) = 1$. Intuitively, events where the two advisers agree have zero diagnostic value for determining who the Expert is. Only events $E2$ and $E3$ generate different probabilities under the two states. Moreover, $\frac{P(E2|S2)}{P(E2|S1)} = \frac{P(E3|S1)}{P(E3|S2)}$. The expression for $OR(E2)$ in equation (11) below is derived by dividing $P(E2|S_2)$ by $P(E2|S_1)$. It is easy to verify that OR(E3) is the inverse of OR(E2).

B.2 Deriving the Formula for Bayesian Updating

To calculate the posterior $P(E|S_1)$ after the participant observes a *single* event $E \in \{E1, E2, E3, E4, E5\}$, let π_0 denote the prior of the participant for state S_1. Then:

$$P(S_1|E) = \frac{\pi_0 \; p(E|S_1)}{\pi_0 P(E|S_1) + (1-\pi_0)P(E|S_2)} = \frac{\pi_0}{\pi_0 + (1-\pi_0)\frac{P(E|S_2)}{P(E|S_1)}} = \frac{\pi_0}{\pi_0 + (1-\pi_0)OR(E)} \quad (10)$$

where $OR(E) = \frac{P(E|S2)}{P(E|S1)}$ is the *'Odds Ratio'* of event E. Under events $E1, E4$, and $E5$, $OR(E) = 1$ and so $P(S_1|E) = \pi_0$. By replacing the Odds Ratio for events $E2$ and $E3$ in equation (10) one obtains the corresponding posteriors. Under $E2$ the Odds Ratio is given below, while for $E3$ we have that $OR(E3) = [OR(E2)]^{-1}$.

$$OR(E2) = \frac{p_p \cdot p_l \cdot (1-p_h) + \frac{(1-p_p)(1-p_l)p_h}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}}{p_p \cdot p_h \cdot (1-p_l) + \frac{(1-p_p)(1-p_h)p_l}{3} + \frac{2\cdot(1-p_p)(1-p_l)(1-p_h)}{9}} \quad (11)$$

A *history* h is a sequence $\{E_1, E_2, ...\}$ of events. Since the correct answers to the questions are assumed to be uncorrelated, the events are independent of each other and the posterior can be calculated recursively in the usual way.

$$P(S_1|h) = \frac{P(S_1|h-1)P(E|S_1)}{P(S_1|h-1)P(E|S_1) + P(S_2|h-1)P(E|S_2)}$$

However, the calculations are simplified by the observation that for any two independent events E_1 and E_2, the posterior of the joint events is equal to:

$$P(S_1|E_1, E_2) = \frac{\pi_0}{\pi_0 + (1-\pi_0)OR(E_1)OR(E_2)}$$

and so

$$P(S_1|h) = \frac{\pi_0}{\pi_0 + (1-\pi_0)\prod_{E\in h} OR(E)}$$

Taking into account that $OR(E) = 1$ for events $E1, E4$, and $E5$ and $OR(E3) = [OR(E2)]^{-1}$, the expression simplifies further to:

$$P(S_1|h) = \frac{\pi_0}{\pi_0 + (1-\pi_0)\cdot(OR)^K} \tag{12}$$

where OR is simply $OR(E2)$, and $K \equiv$ [number of times $E2$ has been observed]-[number of times $E3$ has been observed].

B.3 Low Information Model: Stage 3

Table 11: *Low-info model, Stage 3: Probability of event $\hat{E}$ in state S.*

Events	$\mathbf{P(E\|S_1)}$	$\mathbf{P(E\|S_2)}$
$\hat{E}1$	$p_h\cdot(1-p_l)$	$p_l\cdot(1-p_h)$
$\hat{E}2$	$(1-p_h)\cdot p_l$	$(1-p_l)\cdot p_h$
$\hat{E}3$	$p_h\cdot p_l$	$p_l\cdot p_h$
$\hat{E}4$	$(1-p_h)\cdot(1-p_l)$	$(1-p_l)\cdot(1-p_h)$

Notes: *The full description of events in Stage 3 is provided in page 12.*

Given the probabilities expressed in Table 11, the odds ratio $\hat{OR}(\hat{E}) \equiv \frac{P(\hat{E}|S_2)}{P(\hat{E}|S_1)}$ can be calculated in the usual manner. Similarly to Stage 2, the odds ratio for events $\hat{E}3$ and $\hat{E}4$ is equal to one, while $\hat{OR}(\hat{E}2) = \left[\hat{OR}(\hat{E}1)\right]^{-1}$. Dividing $P(\hat{E}1|S_2)$ with $P(\hat{E}1|S_1)$ gives $\hat{OR}$ used in equation (13) below. The posterior of S_1 after feedback on a single question and with no other information is given by replacing OR with $\hat{OR}$ in equation (10), but with π_0 interpreted as the prior at the beginning of Stage 3.

Now, a history $\hat{h}$ in Stage 3 is a collection of events $\{\hat{E}_1, \hat{E}_2, ..., \hat{E}_k,\}$ (for questions with feedback) and a collection $\{E_1, E_2, ..., E_{11-k},\}$ (for questions without feedback). The posterior of S_1 after a history $\hat{h}$ can be derived by modifying equation (12) accordingly. Let us define as $\hat{N}$ the set of questions for which feedback has been provided in Stage 3. Let $K_{\hat{N}}$ measure the number of times that $A1$ gave a correct answer and $A2$ gave a wrong answer in $\hat{N}$, minus the number of times that $A2$ gave a correct answer and $A1$ gave a wrong answer in $\hat{N}$. Let $K_{-\hat{N}}$ measure the number of times that $A1$ alone had a common answer with the participant minus the number of times that $A2$ alone had a common answer with the participant in the set of questions that do not belong to $\hat{N}$. $K_{-\hat{N}}$ is defined in the same way as measure K in Stage 2, but for only a subset of questions. Given the above definitions, the formula for Bayesian updating for Stage 3 can be expressed as:

$$P(S_1/h) = \frac{\pi_0}{\pi_0 + (1-\pi_0)\cdot(OR)^{K_{-\hat{N}}}\cdot \hat{OR}^{K_{\hat{N}}}} \qquad \text{with} \;\; \hat{OR} = \frac{p_l\cdot(1-p_h)}{p_h\cdot(1-p_l)} \tag{13}$$

Note that the above equation implies that for questions with feedback in Stage 3, the events of Stage 2 (whether or not the participant had common answers with the advisers) are not relevant for the calculation of the posterior. Only the events of Stage 3 (whether the advisers were correct or not) matter for these questions. Let us explain why this is the case. Consider two histories, h and $\hat{h}$, which correspond to Stage 2 and Stage 3 respectively, and which concern the same questions, say questions 1 to 3.

For instance, assume h specifies that in questions 1 to 2 the participant agreed only with $A1$ and then in question 3 she only agreed with $A2$. On the other hand, let $\hat{h}$ specify that in questions 1 to 3, $A2$ was always right (so that $A1$ was always wrong). Under the information of only Stage 2, the participant infers that she has more common answers with $A1$ than $A2$ and so S_1 is more likely than S_2. However, once the feedback of Stage 3 is made available in $\hat{h}$, this conclusion is overturned since $A1$ was wrong in all of them. Intuitively, the information on who is correct on each question from Stage 3 generates a finer partition of the participant's information set, so that the coarser partition from Stage 2 is obsolete for these questions. Continuing our example, for questions beyond question 3, namely 4 to 11, the participant has no feedback from Stage 3 and so she can utilize only the coarse partition of Stage 2. Since the correct answers across questions are independent, this gives this convenient formula.

B.4 High Information Model: Stage 2

The list of events and the configurations of answers per event are the same as in the low-info model and are presented in Table 9. The table below gives the probabilities of these events in each state.

These are computed by estimating the probability of each configuration as in the example provided in the main text.

Table 12: *High-info model, Stage 2: Probability of each event in every state.*

Events	$\mathbf{P(E\|S_1)}$
$E1$	$p_p \cdot \pi \cdot p_h + (1-p_p) \cdot \pi \cdot (1-p_h)/3$
$E2$	$p_p \cdot (1-\pi) \cdot p_h + (1-p_p) \cdot (1-\pi) \cdot (1-p_h)/3$
$E3$	$p_p \cdot \pi \cdot (1-p_h) + (1-p_p) \cdot \pi \cdot p_h + (2/3) \cdot (1-p_p) \cdot \pi \cdot (1-p_h)$
$E4$	$(1/3) \cdot (1-p_p) \cdot (1-\pi) \cdot p_h + (2/9) \cdot (1-p_p) \cdot (1-\pi) \cdot (1-p_h) + (1/3) \cdot p_p \cdot (1-\pi) \cdot (1-p_h)$
$E5$	$(2/3) \cdot p_p \cdot (1-\pi) \cdot (1-p_h) + (4/9) \cdot (1-p_p) \cdot (1-\pi) \cdot (1-p_h) + (2/3) \cdot (1-p_p) \cdot (1-\pi) \cdot p_h$
Events	$\mathbf{P(E\|S_2)}$
$E1$	$p_p \cdot \pi \cdot p_h + (1-p_p) \cdot \pi \cdot (1-p_h)/3$
$E2$	$p_p \cdot \pi \cdot (1-p_h) + (1-p_p) \cdot \pi \cdot p_h + (2/3) \cdot (1-p_p) \cdot \pi \cdot (1-p_h)$
$E3$	$p_p \cdot (1-\pi) \cdot p_h + (1-p_p) \cdot (1-\pi) \cdot (1-p_h)/3$
$E4$	$(1/3) \cdot (1-p_p) \cdot (1-\pi) \cdot p_h + (2/9) \cdot (1-p_p) \cdot (1-\pi) \cdot (1-p_h) + (1/3) \cdot p_p \cdot (1-\pi) \cdot (1-p_h)$
$E5$	$(2/3) \cdot p_p \cdot (1-\pi) \cdot (1-p_h) + (4/9) \cdot (1-p_p) \cdot (1-\pi) \cdot (1-p_h) + (2/3) \cdot (1-p_p) \cdot (1-\pi) \cdot p_h$

Table 12 allows us to calculate the odds ratios for each event. Since $P(E|S_1) = P(E|S_2)$ for events $E1, E4$, and $E5$, these events have odds ratios equal to one. The odds ratio of equation (5) in page 14 is derived by dividing $P(E2|S_2)$ by $P(E2|S_1)$. It is straightforward to verify that the odds ratio for $E3$ is the inverse of $E2$.

B.5 High Information Model: Stage 3

The full description of the possible events in questions with feedback for the high-info model is given in page 16. The following table gives the probabilities of each event for each state.

Table 13: *High-info model, Stage 3: Probability of event $\tilde{E}$ in state S.*

Events	$\mathbf{P(E\|S_1)}$	$\mathbf{P(E\|S_2)}$
$\tilde{E}1$	$p_p p_h \pi$	$p_p \pi p_h$
$\tilde{E}2$	$p_p(1-p_h)(1-\pi)$	$p_p(1-\pi)(1-p_h)$
$\tilde{E}3$	$p_p p_h(1-\pi)$	$p_p \pi(1-p_h)$
$\tilde{E}4$	$(1-p_p)p_h\frac{(2+\pi)}{3}$	$(1-p_p)\frac{(1-\pi)}{3}(1-p_h)$
$\tilde{E}5$	$p_p(1-p_h)\pi$	$p_p(1-\pi)p_h$
$\tilde{E}6$	$(1-p_p)p_h\frac{(1-\pi)}{3}$	$(1-p_p)\frac{(1-\pi)}{3}p_h$
$\tilde{E}7$	$(1-p_p)(1-p_h)\frac{(1-\pi)}{3}$	$(1-p_p)\frac{(2+\pi)}{3}p_h$
$\tilde{E}8$	$(1-p_p)(1-p_h)\frac{(2+\pi)}{3}$	$(1-p_p)\frac{(2+\pi)}{3}(1-p_h)$

From the above table we observe that the odds ratios are equal to one for events $\tilde{E}1$, $\tilde{E}2$, $\tilde{E}6$, and $\tilde{E}8$. Dividing $P(\tilde{E}|S_2)$ by $P(\tilde{E}|S_1)$ for events $\tilde{E}3$ and $\tilde{E}4$ one obtains the odds ratios in equations (7) and (8) respectively in page 17. Finally, one can easily verify that the odds ratio for $\tilde{E}5$ is the inverse of the analogous ratio for $\tilde{E}3$ and the odds ratio for $\tilde{E}7$ is the inverse of the analogous ratio for $\tilde{E}4$.

C Additional Results

C.1 Stage-3 results broken down by experimental condition

Our experiments had a total of 5 conditions, including the additional checks. Conditions 1 and 2 are low-info conditions corresponding to Experiment 1. Conditions 3, 4 and 5 are high-info conditions, corresponding to Experiment 2. In Condition 1 we allowed the Expert to be correct 70% of the time, against 85% in all other four conditions. Condition 3 was high-info and it was the only condition without financial incentives. Conditions 4 and 5 differed in the order of the questions, in that in Condition 4 the order was harder than in Condition 5 for the participants to distinguish between the two advisers. Table 14 below summarizes the characteristics of each condition.

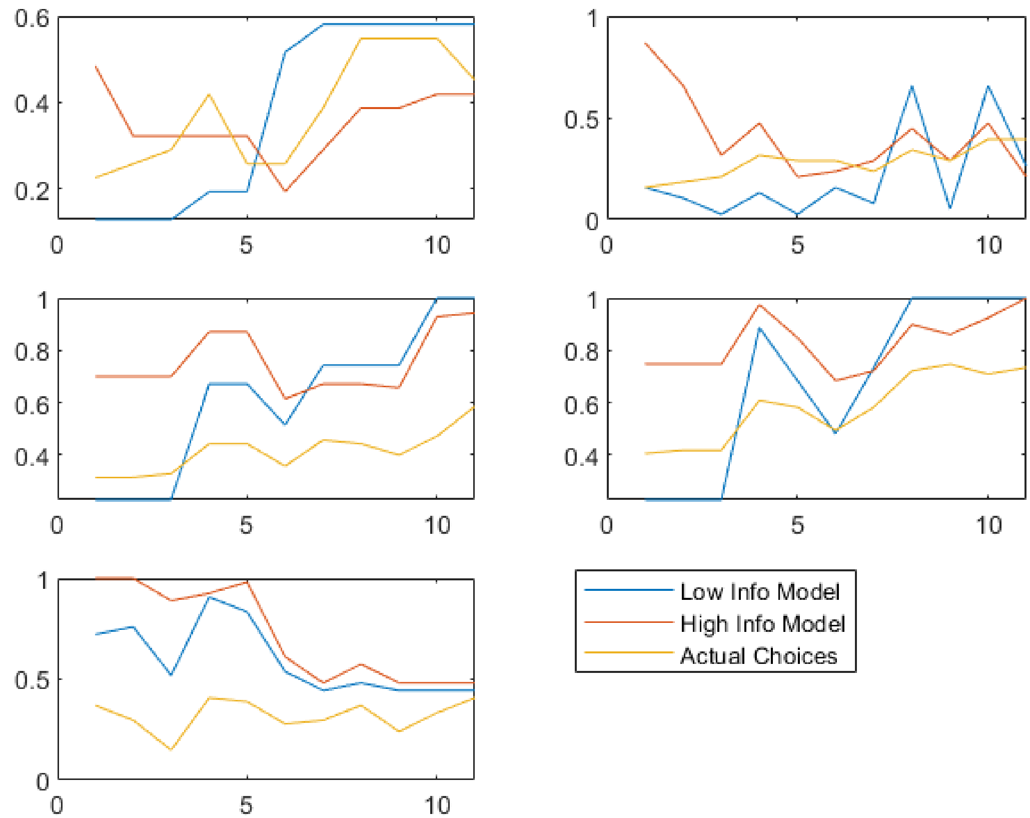
Table 14: *Summary of the five experimental conditions.*

Condition	Description
1	low-info, 70% expert accuracy, incentivized
2	low-info, 85% expert accuracy, incentivized
3	high-info, 85% expert accuracy, not incentivized
4	high-info, 85% expert accuracy, incentivized
5	high-info, 85% expert accuracy, incentivized, reversed order

Figure 7 presents the aggregate predictions of the two models broken down by condition and juxtapose it with actual behavior. Since this is a learning environment, more emphasis is placed on convergence in the last few rounds. As we can see, in all high-info conditions, the fraction of participants choosing the Expert is much lower in the last few rounds relative to the two models' predictions. In Condition 1, the low-info model does not seem to capture the data well, and the predictions of the high-info model are closer to aggregate behavior. However, in Condition 2 the low-info model seems to follow the data relatively closely.

In the remaining three conditions it is the high-info model that captures the environment theoretically, but in all of these conditions it consistently over-predicts the tendency of participants to choose the Expert. In Condition 3 and Condition 4, the high-info model more closely resembles actual aggregate behavior in late rounds than the the low-info model. The stronger tendency of participants to choose the Expert in late rounds of Condition 4 relative to Condition 3 indicates that incentives may have played a role in Condition 4. Finally, in Condition 5 the high-info model predicts that a high fraction of participants chooses the Expert at the beginning of Stage 3, which is not what happens in the data. However, both models seem to converge to actual behavior in

Figure 7: *Aggregate performance of the Bayesian models in Stage 3 broken down by experimental condition (conditions 1-5).*

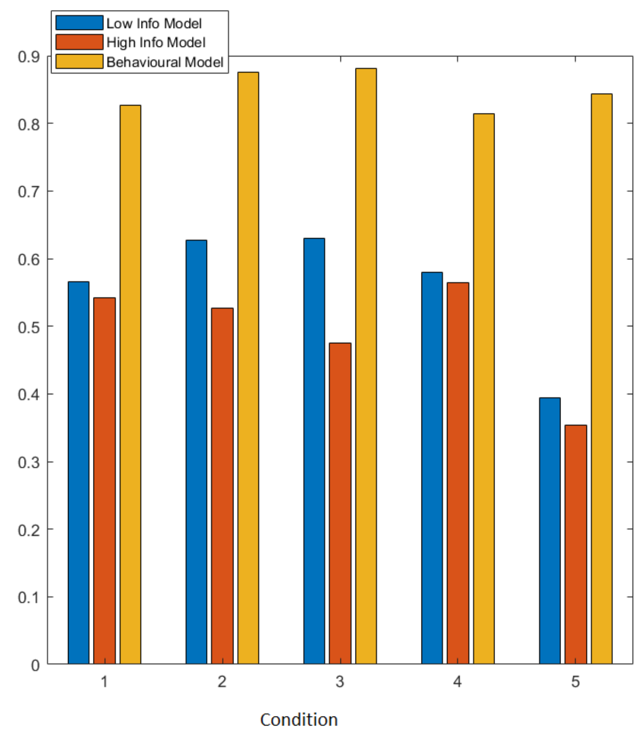


Notes: *The horizontal axis in each subplot represents Stage-3 rounds. The top left panel depicts condition 1, top right condition, 2 and so on. The horizontal axis shows the relevant round number of Stage 3 and the vertical axis the fraction of participants who chose the 'Expert' according the two models and in the data.*

late rounds, in terms of their prediction of the fraction of the participants that choose the Expert.

Figure 8 illustrates the fit of the three main models (the two Bayesian ones and the idiosyncratic behavioral model) for each of the five conditions. Remarkably, the high-info model fails to outperform the low-info model in any of the five conditions. This is particularly surprising, since this model captures closely the experimental environment of the high-info conditions 1-3. Finally, in Condition 5 we observe a considerable drop in fit for both models.

Figure 8: *Comparison of the fit of the three models broken down by experimental condition.*



Notes: *The bars show the fraction of total individual choices, aggregated across all rounds of Stage 3, that agree with the respective model's prediction for the particular condition, participant and round.*

C.2 Further analysis on complementary observable variables

Variable	**Explanation**
Correct i	Number of correct answers at Stage $i = 1, 2, 3$
Chose Expert	Participant chose the expert at Stage 2
Economics (Sociology)	Participant took Economics (Sociology) courses
Science/Maths	Participant took science/maths courses
Politics/Law	Participant took politics/law courses
CRT sum	Sum of correct answers in the classic CRT questionnaire
Self Enhancement	Sum of answers in the self enhancement questionnaire
Self-esteem	Rosenberg's self-esteem 10 items scale (sum)
Narcissism	Narcissism question
Econ (Hard) Science	Answer to the question if participant believes Economics is a (hard) science
Sex	1 is female, 2 is male, 3 is other
Age	Participant's age
Qualifications	The highest qualification the participant has obtained 0 is none or other, 1 is A-Levels or equivalent, 2 is Bachelor's degree, 3 is Master's degree, 4 is Doctorate
Knowledgeable	Answer to the question: "When it comes to matters of public policy, such as the minimum wage, taxes, or public investments, how knowledgeable do you consider yourself?" 1 is lowest, 7 is highest
Political Spectrum	Seven point scale, 1 is left, 7 is right
Level k	Depth of reasoning implied by answer in the undercutting game
Confidence # correct	Number of questions at Stage 1 the participants believe they have answered correctly
Confidence in Adviser	Belief the participant has chosen the best adviser at Stage 2
Questions order	Order of questions used

Table 15: *Simple correlations, all experiments pooled.*

	Correct 1	Correct 2	Correct 3	Chose Expert
A Levels				
Economics	0.15*	0.10	−0.01	0.06
Sociology	−0.18 **	0.13*	0.06	0.13*
Science/Maths	0.20**	0.01	−0.09	−0.02
Politics/Law	0.01	0.00	0.08	0.01
University courses				
Economics	0.15*	0.08	−0.02	0.06
Sociology	−0.10	0.11	0.09	0.13*
Science/Maths	0.13*	−0.09	−0.05	−0.10
Politics/Law	0.01	0.03	0.09	0.02
Psychological measures				
CRT sum	0.28***	0.04	0.00	0.03
Self Enhancement	−0.06	−0.07	0.02	−0.09
Narcissism	0.10	−0.09	−0.09	−0.06
Self-esteem	−0.02	−0.07	−0.04	−0.07
Other measures				
Econ Hard Science	−0.12	0.01	0.07	0.03
Econ Science	−0.03	0.00	0.05	0.03
Sex	0.25***	−0.05	−0.07	−0.07
Age	−0.19 **	0.01	0.12*	0.04
Qualifications	−0.07	0.06	0.11	0.10
Knowledgeable	0.10	−0.01	0.03	−0.01
Political Spectrum	−0.04	0.08	0.01	0.12*
Level k	−0.10	−0.06	0.05	−0.02
Confidence # correct	0.45***	0.02	−0.04	−0.04
Confidence in Adviser	0.24***	−0.14	−0.06	−0.15 *
Questions Order	−0.13 *	−0.04	−0.03	0.02

Table 16: *Simple correlations, Experiment 1.*

	Correct 1	Correct 2	Correct 3	Chose Expert
A Levels				
Economics	0.11	−0.07	−0.11	−0.07
Sociology	−0.13	0.35**	0.16	0.37**
Science/Maths	0.14	0.12	−0.02	0.11
Politics/Law	0.08	0.15	0.14	0.16
University courses				
Economics	0.12	−0.08	0.15	−0.08
Sociology	0.02	0.28*	0.14	0.30*
Science/Maths	0.19	−0.13	−0.18	−0.14
Politics/Law	−0.04	−0.08	0.13	−0.08
Psychological measures				
CRT sum	0.38**	0.16	−0.11	0.17
Self Enhancement	−0.20	−0.25 *	−0.32 **	−0.26 *
Narcissism	0.05	−0.02	−0.07	−0.01
Self-esteem	0.01	−0.09	−0.22	−0.10
Other measures				
Econ Hard Science	−0.17	−0.03	0.28*	−0.01
Econ Science	−0.08	−0.16	0.01	−0.15
Sex	0.23	−0.04	−0.20	−0.04
Age	−0.18	0.13	0.18	0.13
Qualifications	−0.25 *	0.13	0.17	0.13
Knowledgeable	0.09	−0.34 **	−0.07	−0.33 **
Political Spectrum	−0.02	0.30*	0.31**	0.30*
Level k	−0.42 ***	−0.16	0.13	−0.17
Questions Order	0.10	−0.08	−0.09	−0.09

Table 17: *Simple correlations, Experiment 2.*

	Correct 1	Correct 2	Correct 3	Chose Expert
A Levels				
Economics	0.18**	0.12	0.00	0.07
Sociology	−0.20 **	0.08	0.04	0.07
Science/Maths	0.22**	−0.03	−0.12	−0.07
Politics/Law	0.01	−0.04	0.06	−0.04
University courses				
Economics	0.18*	0.11	−0.08	0.08
Sociology	−0.18 **	0.10	0.10	0.12
Science/Maths	0.09	−0.06	0.00	−0.06
Politics/Law	0.04	0.05	0.08	0.04
Psychological measures				
CRT sum	0.25***	0.02	0.03	0.00
Self Enhancement	−0.02	−0.03	0.10	−0.05
Narcissism	0.10	−0.11	−0.10	−0.07
Self-esteem	−0.01	−0.09	−0.01	−0.09
Other measures				
Econ Hard Science	−0.10	0.03	0.01	0.04
Econ Science	−0.01	0.05	0.08	0.09
Sex	0.26***	−0.06	−0.04	−0.08
Age	−0.24 ***	0.03	0.16*	0.07
Qualifications	−0.03	0.08	0.14*	0.14*
Knowledgeable	0.13	0.05	0.03	0.05
Political Spectrum	−0.04	0.04	−0.07	0.08
Level k	0.01	−0.03	0.02	0.02
Confidence # correct	0.45***	0.02	−0.04	−0.04
Confidence in Adviser	0.24***	−0.14	−0.06	−0.15 *
Questions Order	−0.14 *	−0.12	−0.07	−0.05

Table 18: *Regressing experimental performance on 'A-level courses'.*

A Levels	Correct 1			Correct 2			Correct 3			Chose Expert		
	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	4.62^{***}	5.65^{***}	6.84^{***}	6.71^{***}	6.17^{***}	6.84^{***}	6.84^{***}	6.75^{***}	6.81^{***}	0.23^{*}	0.06	0.27^{*}
	(0.42)	(0.73)	(0.59)	(0.47)	(0.49)	(0.59)	(0.46)	(0.73)	(0.59)	(0.12)	(0.14)	(0.15)
Economics	0.30	0.18	0.47^{*}	0.31	−0.27	0.47^{*}	0.05	−0.38	0.15	0.07	−0.07	0.11
	(0.21)	(0.39)	(0.24)	(0.20)	(0.28)	(0.24)	(0.20)	(0.34)	(0.25)	(0.05)	(0.09)	(0.06)
Sociology	-0.71^{**}	−0.57	0.52	0.70^{*}	1.06	0.52	0.00	0.08	−0.17	0.18^{*}	0.38	0.09
	(0.33)	(0.66)	(0.51)	(0.40)	(0.69)	(0.51)	(0.36)	(0.55)	(0.47)	(0.11)	(0.23)	(0.12)
Science/Maths	0.41^{*}	0.10	−0.12	0.04	0.24	−0.12	−0.13	0.30	−0.31	−0.02	0.04	−0.07
	(0.23)	(0.51)	(0.26)	(0.21)	(0.33)	(0.26)	(0.21)	(0.41)	(0.26)	(0.06)	(0.09)	(0.07)
Politics/Law	0.25	0.34	−0.22	−0.11	0.21	−0.22	0.31	0.59	0.20	−0.03	0.09	−0.07
	(0.23)	(0.46)	(0.30)	(0.25)	(0.53)	(0.30)	(0.26)	(0.60)	(0.32)	(0.07)	(0.17)	(0.08)
Sex: Male	0.72^{***}	0.59	−0.26	−0.16	0.07	−0.26	−0.16	−0.48	−0.11	−0.05	0.03	−0.08
	(0.22)	(0.48)	(0.24)	(0.20)	(0.38)	(0.24)	(0.22)	(0.44)	(0.25)	(0.06)	(0.10)	(0.07)
Sex: Other	−0.31	0.56	0.19	−0.15	−0.47	0.19	−0.01	−0.48	0.59	−0.09	−0.13	−0.01
	(0.76)	(1.36)	(1.40)	(0.68)	(0.33)	(1.40)	(0.70)	(1.00)	(0.98)	(0.17)	(0.10)	(0.36)
Age	−0.02	−0.04	0.00	−0.00	0.00	0.00	0.02	0.01	0.03	0.00	0.00	0.00
	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.00)	(0.00)	(0.01)
R^2	0.14	0.10	0.03	0.03	0.12	0.03	0.02	0.07	0.04	0.03	0.16	0.03
Adj. R^2	0.12	−0.00	−0.00	0.00	0.02	−0.00	−0.00	−0.03	0.00	0.00	0.06	−0.01
Num. obs.	272	69	203	272	69	203	272	69	203	272	69	203

Results from linear regressions with robust standard errors. In columns (1) to (3) the dependent variable is the number of correct answers a participant had at Stage 1. In columns (4) to (6) the dependent variable is the number of correct answers of the adviser the participant chose at Stage 2. In columns (7) to (9) the dependent variable is the number of correct answers the participant attained at Stage 3. In columns (10) to (12) the dependent variable is 1 if the participant chose the Expert at Stage 2 and 0 if not. In columns (1), (4), (7) and (10) data from both experiments have been used. In columns (2), (5), (8) and (11) only data from the low information experiment have been used. In columns (3), (6), (9) and (12) only data from the high information experiment have been used. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 19: *Regressing experimental performance on 'university courses'.*

University courses	Correct 1			Correct 2			Correct 3			Chose Expert		
	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	4.93^{***}	5.38^{***}	4.67^{***}	6.85^{***}	6.62^{***}	6.76^{***}	6.91^{***}	6.79^{***}	6.96^{***}	0.25^{**}	0.18	0.23^{*}
	(0.40)	(0.80)	(0.43)	(0.43)	(0.50)	(0.53)	(0.46)	(0.85)	(0.55)	(0.11)	(0.13)	(0.13)
Economics	0.41^{**}	0.37	0.49^{**}	0.19	−0.33	0.34	−0.08	0.40	−0.25	0.05	−0.08	0.09
	(0.21)	(0.38)	(0.24)	(0.20)	(0.29)	(0.24)	(0.21)	(0.39)	(0.25)	(0.05)	(0.08)	(0.06)
Sociology	−0.06	0.70	−0.44	0.56^{*}	1.28^{***}	0.55	0.11	−0.04	0.23	0.18^{**}	0.45^{***}	0.14
	(0.33)	(0.57)	(0.37)	(0.34)	(0.46)	(0.48)	(0.33)	(0.45)	(0.45)	(0.09)	(0.15)	(0.12)
Science/Maths	0.46^{*}	0.55	0.38	−0.32	−0.33	−0.25	−0.14	−0.27	0.02	-0.10^{*}	−0.12	−0.07
	(0.24)	(0.46)	(0.29)	(0.21)	(0.29)	(0.26)	(0.22)	(0.37)	(0.29)	(0.06)	(0.08)	(0.07)
Politics/Law	0.16	−0.66	0.32	0.01	-0.94^{*}	0.13	0.30	0.30	0.28	−0.03	-0.31^{**}	−0.00
	(0.25)	(0.58)	(0.29)	(0.29)	(0.47)	(0.33)	(0.27)	(0.64)	(0.31)	(0.07)	(0.15)	(0.08)
Sex: Male	0.83^{***}	0.72	0.92^{***}	−0.05	0.42	−0.15	−0.14	−0.36	−0.10	−0.02	0.13	−0.06
	(0.21)	(0.43)	(0.25)	(0.20)	(0.31)	(0.24)	(0.21)	(0.38)	(0.25)	(0.05)	(0.09)	(0.06)
Sex: Other	−0.07	0.96	-1.09^{***}	−0.18	−0.29	0.26	0.08	−0.18	0.73	−0.11	−0.06	−0.01
	(0.78)	(1.17)	(0.37)	(0.68)	(0.32)	(1.31)	(0.71)	(1.01)	(1.00)	(0.17)	(0.10)	(0.33)
Age	-0.04^{**}	−0.04	-0.03^{*}	−0.00	−0.01	0.00	0.02	0.01	0.02	−0.00	−0.00	0.00
	(0.02)	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)	(0.00)	(0.00)	(0.01)
R^2	0.13	0.13	0.17	0.03	0.17	0.03	0.02	0.06	0.04	0.04	0.23	0.03
Adj. R^2	0.11	0.03	0.15	0.00	0.07	−0.00	−0.00	−0.05	0.00	0.01	0.15	−0.00
Num. obs.	272	69	203	272	69	203	272	69	203	272	69	203

Results from linear regressions with robust standard errors. In columns (1) to (3) the dependent variable is the number of correct answers a participant had at Stage 1. In columns (4) to (6) the dependent variable is the number of correct answers of the adviser the participant chose at Stage 2. In columns (7) to (9) the dependent variable is the number of correct answers the participant attained at Stage 3. In columns (10) to (12) the dependent variable is 1 if the participant chose the Expert at Stage 2 and 0 if not. In columns (1), (4), (7) and (10) data from both experiments have been used. In columns (2), (5), (8) and (11) only data from the low information experiment have been used. In columns (3), (6), (9) and (12) only data from the high information experiment have been used. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 20: *Regressing experimental performance on 'psychological measures'.*

Psychological measures	Correct 1			Correct 2			Correct 3			Chose Expert		
	Overall	Low info	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	4.37^{***}	4.36^{***}	4.24^{***}	7.36^{***}	6.46^{***}	7.58^{***}	7.08^{***}	7.97^{***}	6.70^{***}	0.35^{**}	0.11	0.40^{**}
	(0.52)	(0.82)	(0.59)	(0.53)	(0.75)	(0.67)	(0.57)	(1.15)	(0.70)	(0.15)	(0.19)	(0.18)
CRT sum	0.35^{***}	0.52^{**}	0.27^{**}	0.07	0.15	0.07	0.07	−0.07	0.13	0.02	0.06	0.01
	(0.10)	(0.20)	(0.11)	(0.10)	(0.23)	(0.11)	(0.10)	(0.22)	(0.11)	(0.03)	(0.06)	(0.03)
Self Enhancement	−0.02	-0.20^{*}	0.03	−0.05	-0.14^{**}	−0.03	−0.01	-0.15^{*}	0.03	−0.01	-0.05^{**}	−0.01
	(0.06)	(0.11)	(0.07)	(0.04)	(0.07)	(0.05)	(0.05)	(0.08)	(0.06)	(0.01)	(0.02)	(0.02)
Narcissism	0.04	−0.02	0.07	−0.09	−0.05	−0.09	−0.07	−0.05	−0.07	−0.01	−0.01	−0.01
	(0.06)	(0.10)	(0.08)	(0.06)	(0.09)	(0.08)	(0.06)	(0.09)	(0.08)	(0.02)	(0.03)	(0.02)
Self-esteem	0.00	0.02	0.01	−0.01	0.00	−0.02	−0.01	−0.02	−0.00	−0.00	−0.00	−0.00
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.00)	(0.00)	(0.00)
Sex: Male	0.81^{***}	0.85^{**}	0.90^{***}	−0.17	0.11	−0.28	−0.19	−0.28	−0.17	−0.06	0.02	−0.09
	(0.22)	(0.39)	(0.25)	(0.20)	(0.38)	(0.24)	(0.21)	(0.40)	(0.26)	(0.06)	(0.10)	(0.07)
Sex: Other	0.02	0.67	−1.01	−0.26	-0.61^{**}	0.32	0.02	−0.37	0.72	−0.11	-0.19^{**}	0.01
	(0.73)	(0.78)	(0.77)	(0.79)	(0.29)	(1.46)	(0.79)	(0.96)	(1.05)	(0.20)	(0.09)	(0.36)
Age	-0.03^{*}	0.00	-0.03^{**}	0.00	0.02	0.00	0.02	0.01	0.03	0.00	0.01	0.00
	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)	(0.00)	(0.00)	(0.01)
R^2	0.15	0.23	0.17	0.02	0.06	0.03	0.02	0.10	0.04	0.02	0.09	0.02
Adj. R^2	0.13	0.14	0.14	−0.01	−0.05	−0.01	−0.00	−0.00	0.00	−0.01	−0.01	−0.01
Num. obs.	272	69	203	272	69	203	272	69	203	272	69	203

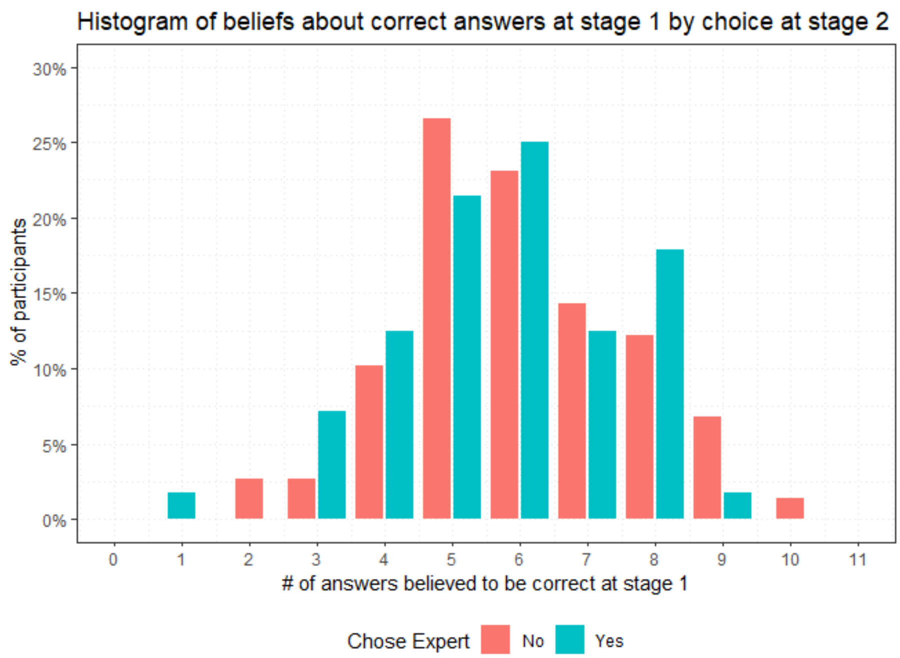
Results from linear regressions with robust standard errors. In columns (1) to (3) the dependent variable is the number of correct answers a participant had at Stage 1. In columns (4) to (6) the dependent variable is the number of correct answers of the adviser the participant chose at Stage 2. In columns (7) to (9) the dependent variable is the number of correct answers the participant attained at Stage 3. In columns (10) to (12) the dependent variable is 1 if the participant chose the Expert at Stage 2 and 0 if not. In columns (1), (4), (7) and (10) data from both experiments have been used. In columns (2), (5), (8) and (11) only data from the low information experiment have been used. In columns (3), (6), (9) and (12) only data from the high information experiment have been used. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 21: *Regressing experimental performance on 'other measures'.*

Other measures	Correct 1			Correct 2			Correct 3			Chose Expert		
	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2	Overall	Experiment 1	Experiment 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	5.47^{***}	7.08^{***}	4.94^{***}	6.51^{***}	6.62^{***}	6.45^{***}	6.37^{***}	5.34^{***}	6.65^{***}	0.07	0.16	0.02
	(0.54)	(1.00)	(0.61)	(0.61)	(0.68)	(0.74)	(0.59)	(0.86)	(0.72)	(0.16)	(0.19)	(0.19)
Qualifications	0.08	-0.41^{*}	0.09	0.07	-0.06	0.15	0.13	-0.05	0.22	0.03	-0.01	0.06
	(0.13)	(0.24)	(0.15)	(0.11)	(0.15)	(0.15)	(0.12)	(0.20)	(0.16)	(0.03)	(0.05)	(0.04)
Knowledgeable	0.08	0.01	0.11	-0.01	-0.31^{***}	0.07	0.06	-0.01	0.07	-0.00	-0.08^{***}	0.01
	(0.07)	(0.10)	(0.08)	(0.07)	(0.10)	(0.09)	(0.07)	(0.12)	(0.09)	(0.02)	(0.03)	(0.02)
Level k	-0.08	-0.47^{***}	0.06	-0.06	-0.19^{**}	-0.04	0.03	0.07	-0.00	-0.01	-0.06^{**}	0.01
	(0.08)	(0.12)	(0.09)	(0.07)	(0.09)	(0.09)	(0.07)	(0.11)	(0.09)	(0.02)	(0.03)	(0.02)
Political Spectrum	-0.07	0.01	-0.06	0.11	0.28^{***}	0.05	0.02	0.31^{**}	-0.06	0.04^{**}	0.08^{***}	0.04
	(0.08)	(0.13)	(0.09)	(0.08)	(0.09)	(0.10)	(0.09)	(0.13)	(0.11)	(0.02)	(0.03)	(0.03)
Sex: Male	0.88^{***}	0.39	0.96^{***}	-0.11	0.31	-0.22	-0.17	-0.25	-0.12	-0.03	0.08	-0.07
	(0.21)	(0.36)	(0.25)	(0.20)	(0.29)	(0.24)	(0.21)	(0.38)	(0.25)	(0.05)	(0.07)	(0.06)
Sex: Other	-0.17	-0.22	-1.26^{*}	-0.11	-0.44	0.48	0.18	0.06	0.90	-0.06	-0.14	0.07
	(0.77)	(0.76)	(0.68)	(0.76)	(0.35)	(1.36)	(0.74)	(0.90)	(1.23)	(0.19)	(0.11)	(0.33)
Age	-0.04^{***}	0.00	-0.05^{***}	0.00	0.02	-0.00	0.02	0.02	0.02	0.00	0.01	-0.00
	(0.01)	(0.03)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.00)	(0.01)	(0.01)
R^2	0.12	0.28	0.15	0.01	0.26	0.02	0.02	0.10	0.04	0.03	0.26	0.03
Adj. R^2	0.10	0.19	0.12	-0.01	0.17	-0.02	-0.00	-0.00	0.00	0.00	0.18	-0.00
Num. obs.	272	69	203	272	69	203	272	69	203	272	69	203

Results from linear regressions with robust standard errors. In columns (1) to (3) the dependent variable is the number of correct answers a participant had at Stage 1. In columns (4) to (6) the dependent variable is the number of correct answers of the adviser the participant chose at Stage 2. In columns (7) to (9) the dependent variable is the number of correct answers the participant attained at Stage 3. In columns (10) to (12) the dependent variable is 1 if the participant chose the Expert at Stage 2 and 0 if not. In columns (1), (4), (7) and (10) data from both experiments have been used. In columns (2), (5), (8) and (11) only data from the low information experiment have been used. In columns (3), (6), (9) and (12) only data from the high information experiment have been used. $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Figure 9: *Distributions of Stage-1 perceived performance for Expert-choosers and Charlatan-choosers.*



References

Akerlof, G. A. and W. T. Dickens (1982): "The Economic Consequences of Cognitive Dissonance," *The American Economic Review*, 72, 307–319.

Andre, P., C. Pizzinelli, C. Roth, and J. Wohlfart (2019): "Subjective Models of the Macroeconomy: Evidence from Experts and a Representative Sample," *CESifo Working Paper*.

Aprea, C. and V. Sappa (2014): "Variations of Young Germans' Informal Conceptions of Financial and Economic Crises Phenomena," *Journal of Social Science Education*, 13, 57–67.

Arthur, W. B. (2000): "Cognition: The Black Box of Economics," *in The Complexity Vision and the Teaching of Economics. D. Colander (Ed.), Northampton, MA: Edward Elgar Publishing.*

Bartels, L. M. (2005): "Homer Gets a Tax Cut: Inequality and Public Policy in the American Mind," *Perspectives on Politics*, 3, 15–31.

Bénabou, R. and J. Tirole (2002): "Self-Confidence and Personal Motivation," *The Quarterly Journal of Economics*, 117, 871–915.

Blendon, R. J., J. M. Benson, M. Brodie, R. Morin, D. E. Altman, D. Gitterman, M. Brossard, and M. James (1997): "Bridging the gap between the public's and economists' views of the economy," *Journal of Economic Perspectives*, 11, 105–118.

Camerer, C. F. and R. M. Hogarth (1999): "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 19, 7–42.

Caplan, B. (2002): "Systematically Biased Beliefs about Economics: Robust Evidence of Judgemental Anomalies from the Survey of Americans and Economists on the Economy," *The Economic Journal*, 112, 433–458.

——— (2011): *The Myth of the Rational Voter: Why Democracies Choose Bad Policies-New Edition*, Princeton University Press.

Chakraborty, A., P. Ghosh, and J. , Roy (2020): "Expert Captured Democracies," *American Economic Review*, 110, 1713–51.

Colander, D. (2005): "The Making of an Economist Redux," *Journal of Economic Perspectives*, 19, 175–198.

Dal Bó, E., P. Dal Bó, and E. Eyster (2018): "The Demand for Bad Policy when Voters Underappreciate Equilibrium Effects," *The Review of Economic Studies*, 85, 964–998.

Dixon, R., W. Griffiths, and G. Lim (2014): "Lay People's Models of the Economy: A Study Based on Surveys of Consumer Sentiments," *Journal of Economic Psychology*, 44, 13–20.

Dräger, L., M. J. Lamla, and D. Pfajfar (2016): "Are Survey Expectations Theory-Consistent? The Role of Central Bank Communication and News," *European Economic Review*, 85, 84–111.

Erev, I. and A. E. Roth (1998): "Predicting how People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review*, 848–881.

Falk, A. and F. Zimmermann (2018): "Information Processing and Commitment," *Economic Journal*, 128, 1983–2002.

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics*, 10, 171–178.

Frederick, S. (2005): "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 19, 25–42.

Gangl, K., B. Kastlunger, E. Kirchler, and M. Voracek (2012): "Confidence in the Economy in Times of Crisis: Social Representations of Experts and Laypeople," *The Journal of Socio-Economics*, 41, 603–614.

Gentzkow, M. and J. M. Shapiro (2006): "Media Bias and Reputation," *Journal of Political Economy*, 114, 280–316.

Georganas, S., P. J. Healy, and R. A. Weber (2015): "On the Persistence of Strategic Sophistication," *Journal of Economic Theory*, 159, 369–400.

Javdani, M. and H.-J. Chang (2019): "Who Said or What Said? Estimating Ideological Bias in Views Among Economists," *SSRN No 3356309 Working Paper.*

Jerrim, J., P. Parker, and D. Shure (2019): "Bullshitters. Who Are They and What Do We Know about Their Lives?" *IZA Discussion Paper*.

Kahneman, D. (2011): *Thinking, Fast and Slow*, Macmillan.

Kelemen, D. and E. Rosset (2009): "The Human Function Compunction: Teleological Explanation in Adults," *Cognition*, 111, 138–143.

Kelemen, D., J. Rottman, and R. Seston (2013): "Professional Physical Scientists Display Tenacious Teleological Tendencies: Purpose-based Reasoning as a Cognitive Default." *Journal of Experimental Psychology: General*, 142, 1074.

Krugman, P. (2010): "Block Those Metaphors," *The New York Times*, 12.

Leiser, D. and R. Aroch (2009): "Lay Understanding of Macroeconomic Causation: The Good-Begets-Good Heuristic," *Applied Psychology*, 58, 370–384.

Leiser, D., S. Bourgeois-Gironde, and R. Benita (2010): "Human Foibles or Systemic Failure—Lay Perceptions of the 2008–2009 Financial Crisis," *The Journal of Socio-Economics*, 39, 132–141.

Leiser, D. and Z. Krill (2017): "How Laypeople Understand the Economy," *Economic Psychology*, 139–154.

Mencken, H. L. (2012): *Mencken Chrestomathy*, Vintage.

Moore, D. A. and P. J. Healy (2008): "The trouble with overconfidence," *Psychological Review*, 115, 502–517.

Newcomb, S. (1893): "The Problem of Economic Education," *The Quarterly Journal of Economics*, 7, 375–399.

Nichols, T. (2017): *The Death of Expertise: The Campaign Against Established Knowledge and Why It Matters*, Oxford University Press.

Oberlechner, T., T. Slunecko, and N. Kronberger (2004): "Surfing the Money Tides: Understanding the Foreign Exchange Market Through Metaphors," *British Journal of Social Psychology*, 43, 133–156.

PENNYCOOK, G., J. A. CHEYNE, N. BARR, D. J. KOEHLER, AND J. A. FUGELSANG (2015): "On the reception and detection of pseudo-profound bullshit," *Judgment and Decision Making*, 10, 549–563.

PENNYCOOK, G., Z. EPSTEIN, M. MOSLEH, A. A. ARECHAR, D. ECKLES, AND D. G. RAND (2019): "Understanding and Reducing the Spread of Misinformation Online," .

PENNYCOOK, G., J. MCPHETRES, Y. ZHANG, J. G. LU, AND D. G. RAND (2020): "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy Nudge Intervention," .

RONAYNE, D. AND D. SGROI (2018): "Ignoring Good Advice," *Competitive Advantage in the Global Economy (CAGE), Working Paper No. 359*.

SCHOTTER, A. (2003): "Decision Making with Naive Advice," *American Economic Review*, 93, 196–201.

STANTCHEVA, S. (2020): "Understanding Economic Policies: What do People Know and How Can they Learn?" Tech. rep., Harvard University Working Paper.

TVERSKY, A. AND D. KAHNEMAN (1980): "Causal Schemas in Judgments under Uncertainty," *Progress in Social Psychology*, 1, 49–72.

VOSOUGHI, S., D. ROY, AND S. ARAL (2018): "The Spread of True and False News Online," *Science*, 359, 1146–1151.

WILLS, G. (2002): *A necessary evil: A history of American distrust of government*, Simon and Schuster.

ZIMMERMANN, F. (2020): "The Dynamics of Motivated Beliefs," *The American Economic Review*, 110, 337–361.

ZINGALES, L. (2020): "The Political Limits of Economics," *Paper Presented at the 2020 American Economic Association Meetings*.