Choosing between Causal Interpretations:
An Experimental Study

Sandro Ambuehl Ⓡ *Heidi C. Thysen*

Choosing between Causal Interpretations: An Experimental Study

Abstract

Good decision-making requires understanding the causal impact of our actions. Often, we only have access to correlational data that could stem from multiple causal mechanisms with divergent implications for choice. Our experiments comprehensively characterize choice when subjects face conflicting causal interpretations of such data. Behavior primarily reflects three types: following interpretations that make attractive promises, choosing cautiously, and assessing the fit of interpretations to the data. We characterize properties of interpretations that obscure bad fit to subjects. Preferences for more complex models are more common than those reflecting Occam's razor. Implications extend to the Causal Narratives and Model Persuasion literatures.

JEL-Codes: C910, D010, D830.

Sandro Ambuehl
Department of Economics & UBS Center for
Economics in Society
University of Zurich / Switzerland
sandro.ambuehl@econ.uzh.ch

Heidi C. Thysen
Department of Economics
Norwegian School of Economics
Bergen / Norway
heidi.thysen@nhh.no

1 Introduction

To make good decisions, understanding the *causal* impact of our actions on desired outcomes is essential. Yet, we frequently encounter only correlational data, which could stem from multiple causal mechanisms that may diverge greatly in their implications for optimal choice. Experts, the media, and politicians present causal interpretations of the facts that align with the data to various degrees. This multiplicity forces us to decide which interpretations, if any, inform our actions.

For example, parents might ponder whether to spend more on their child's college education, enabling attendance at a more prestigious but costlier university. If prestigious universities indeed add more value, a positive correlation between university prestige and graduate success justifies the higher expenditure. However, this correlation could also result if graduate outcomes depend solely on IQ as long as more prestigious institutions admit more capable students. In this case, extra spending on tuition would not produce the desired outcomes. Insights into the actual causal mechanism may arise from empirical correlations on which the two explanations disagree, such as that between university rank and earnings conditional on IQ. While some parents might successfully use this information, others might choose to view the world through rose-tinted glasses and follow the interpretation that justifies high hopes for their children's future. Yet others might prefer interpretations that do not require high tuition payments whose causal impact they do not understand.

In this paper, we experimentally study decision-making in the face of conflicting causal interpretations of correlational data. We focus on three interrelated questions. First, to what extent are individuals able and willing to discard interpretations that do not fit the data they observe? Second, in what way, if any, does decision-making depend on the hopes a causal interpretation raises or on the actions it urges? Third, do individuals have preferences over the structure of interpretations? Do they, for example, favor simpler over more complex interpretations, as Occam's razor suggests, and how do they conceptualize simplicity? Overall, we study how individuals use information to make choices when they do not know the structure of the world, in contrast to the vast literature on belief updating (Benjamin, 2019) that considers the case in which the structure of the world is known.

Our results provide empirical foundations for the emergent literature on mental models whose applications span fields as diverse as behavioral economics (Spiegler, 2016), macroeconomics (Molavi, 2019), finance (Molavi et al., 2021; Shiller, 2017), strategic management (Felin and Zenger, 2017; Camuffo et al., 2023), institutional economics (Denzau and North, 1994), and contract theory (Schumacher and Thysen, 2022). They are especially relevant for the literatures on *Narrative Competition* (Eliaz and Spiegler, 2020; Eliaz et al., 2022; Levy et al., 2022; Angrisani et al., 2023) and *Model Persuasion* (Schwartzstein and Sunderam, 2021, 2022; Ichihashi and Meng, 2021; Jain, 2023; Aina, 2024), whose divergent behavioral assumptions we test in a single experiment. Models of Narrative Competition explain the pervasive multiplicity of conflicting causal narratives in the public discourse, with fascinating implications: multiplicity is an equilibrium that guarantees the survival of false nar-

ratives;[1] it is possible to predict which narratives will persist; they permit comparative statics on the relative prevalence of conflicting narratives; and, when embedded in dynamic political economy frameworks, cycles of populism emerge (Levy et al., 2022). These phenomena hinge on the assumption that individuals adopt the interpretation that promises the most appealing outcome—or, more generally, choose based on any attribute that depends on the narratives' relative support in the population. We examine which, if any, of these attributes determine choice. The Model Persuasion literature characterizes how principals can persuade agents to undertake specific actions by suggesting interpretations of data that agents might not independently conceive. It rests on the premise that individuals choose the interpretation that most closely fits the observable data. Limits in fit-checking and preferences over the structure of interpretations alter the set of persuasive interpretations.

We answer our questions using laboratory experiments because they provide certainty about the data-generating process and facilitate identification by letting us freely mold the decision environment. Their stylized nature allows us to abstract from confounding factors such as prior beliefs or attachment to political groups.[2]

In our experiment, subjects choose an action at a cost. The action affects a monetary outcome via a stochastic causal model (the data generating process; DGP), which, in our case, is a system of linear regression equations. Next to the action and the outcome, each DGP includes two additional variables we call covariates. While the action is always exogenous, the outcome and the covariates can take various roles. They might be endogenous or exogenous, and they might be causes or symptoms of other variables, or mediators between them. Subjects do *not* know the DGP, but they observe the (large-sample) correlational information it generates. They also observe a menu of causal *interpretations* that will inform their action choice. At the core of each interpretation lies a potentially misspecified model that postulates how the four variables are causally related. When fit to the data generated by the DGP, each such model implies a distinct optimal action (the *recommendation*), as well as a payoff that the subject can expect if the model is correctly specified and the recommended action is taken (the *promise*), both of which subjects observe. They choose an action by selecting one of the available interpretations. Its action recommendation is then executed and determines the subjects' payout according to the DGP. To aid their choice, subjects also observe each interpretation's model specification in the form of a directed acyclic graph (DAG) along with the interpretations' recommendation and

---

[1] For example, consider the question of whether mask-wearing causally reduces COVID-19 transmission. Focus on two causal models. Model 1 accurately describes reality and states that increased masking reduces COVID-19 transmission. Model 2 surmises that masking has no effect on disease transmission. If most individuals adopt the first model, they will wear masks, and case counts will be low. According to Model 1, this situation can be maintained only by continued masking, which has small hassle costs. Model 2 is more attractive because it predicts that ending mask-wearing will eliminate hassle costs without affecting case counts. The literature assumes that due to its greater attractiveness, individuals will flock to Model 2. Contrarily, if Model 2 is more popular, individuals will not wear masks, and case counts will be high. According to Model 2, masking cannot change this situation. Model 1 makes the more attractive prediction that case counts can be lowered at a small hassle cost. Individuals will thus flock to Model 1. Overall, the more popular one model, the more attractive the other. Accordingly, no model, including the true one, can survive alone; multiplicity is an equilibrium.

[2] For the same reason, research in experimental game theory (Camerer, 2003) and experimental decision theory (e.g. Benjamin, 2019) also tends to focus on stylized settings.

promise (see Figure 1). Subjects can access unconditional and conditional correlational data produced by the DGP (see Figure 2). In principle, this data suffices to rule out incorrect interpretations, for instance if a predicted correlation between two variables is absent in the data.

We model subjects as applying criteria from three classes: (1) Subjects use *data-based criteria* to discard models that poorly align with the data. These criteria differ in the model implications and correlational information they process. For example, some subjects might draw inferences only from unconditional but not from conditional correlations. (2) *Structure-based criteria* capture preferences over model structures that do not reference the data, such as a preference for simpler over more complex models. (3) *Advice-based criteria* operate on an interpretation's promise or recommendation; they reference neither the model structure nor the data. Subjects who like to view the world through rose-tinted glasses, for instance, might follow high-promise interpretations whereas others might opt for interpretations that recommend the lowest action. These three classes of criteria correspond to our three research questions. The model accommodates simultaneous application of criteria from different classes.

Our primary objective is to quantify how often subjects use each decision criterion. We employ two complementary identification strategies. Experiment 1 involves a sequence of menus, each characterized by a DGP and two or three interpretations from which subjects select one. Any combination of decision criteria, which we call a *type*, implies a distinct *fingerprint* of choices across the menus. Averaging fingerprints across types, weighted by the probability of each type, yields a probability distribution over choices across the menu sequence. We estimate the empirical type distribution by fitting a finite mixture model that minimizes the distance between the model-implied and observed choice distributions. Constructing a menu sequence that pinpoints all elements of the type distribution is challenging because changing any single aspect of a decision problem generally has multiple effects. For example, the structure of a model is inherently linked to it's correlational implications. Therefore, changing the misspecified model will likely affect its consistency with both data- and structure-based criteria. We use several theoretical insights to construct a menu sequence. We then prove formally that it identifies all type probabilities. Experiment 2 identifies the popularity of the decision criteria by hiding selected elements of the decision problem. For example, we infer the fraction of subjects who effectively utilize correlational data by comparing the proportion of participants who choose the correct interpretation when that data is available to the corresponding frequency when it is unavailable. The two identification strategies complement each other because they depend on different identification assumptions. Close congruence in the estimates derived from the two strategies suggests that neither identifying assumption is severely mistaken.

We conduct Experiments 1 and 2 using university student samples. To assess the subject pool dependency of our results, Experiment 3 recruits a U.S. general population sample that completes a simplified and abbreviated version of Experiment 1.

Empirically, which interpretations guide subjects' actions? Broadly, our data feature three predominant groups of subjects. The first seeks to objectively determine the correct interpretations the way a scientist would, with varying levels of success. The second opts for the interpretation that offers the most optimistic outlook, consistent with a preference to view the world through rose-tinted glasses. The third minimizes spending on the action, plausibly due to a reluctance to invest in actions with consequences they may not understand clearly. Among subjects who choose by model structure, a preference for complexity is twice as common as a preference for simplicity. The prevalence of these groups varies across samples. Laboratory participants much more frequently exclude incorrect interpretations compared to individuals from the U.S. general population, who more often favor the most optimistic interpretations. This high-level summary of our results, however, hides much nuance.

In response to our first research question (discarding ill-fitting interpretations), we document pronounced dispersion in the use of data-based criteria. Close to 40% of student participants consistently choose the correct interpretation. This number that drops to 5% in the U.S. general population sample both in the presence and in the absence of a treatment that causes subjects to view data charts thirteen percentage points more often. Another substantial minority of subjects (15% among students and 20% among the U.S. general population) successfully utilize unconditional correlations but do not draw correct inferences from conditional correlations. The largest fractions in both samples (45% of students and 75% of the U.S. general population) fail to use any correlational information effectively. Reassuringly, we estimate similar distributions of criteria usage through both identification strategies in our student samples.

The popularity of high-promise (15% of student subjects and over 40% of U.S. general population subjects) and low-action interpretations (30% of student subjects and 40% of U.S. general population subjects) underscores the significance of our second question (advice-based criteria). Many individuals exclusively rely on these criteria, not merely as a fallback when other criteria do not identify an interpretation. Ex ante, two additional advice-based criteria appear plausible. First, subjects choosing according to Gilboa and Schmeidler (1989)'s maximin criterion should consistently opt for the low-promise interpretation in our setting. Second, individuals subject to the illusion of control (Langer, 1975) may prefer actively doing something over doing nothing, and hence might choose high-action interpretations. Yet, both these criteria receive minimal support across all experiments.

In answer to our third question (preferences over the structure of interpretations), we find that nearly 40% of student subjects employ structure-based criteria. Ten percent of subjects prefer simpler models over more complex ones, whereas 23% show a reverse preference. Though puzzling from the perspective of Occam's razor, this result is intuitive once one realizes that simplification often involves restricting models. Subjects who are not sufficiently confident about the validity of such restrictions will thus prefer the unrestricted, more complex specifications. (The general population version of the experiment is not equipped to identify structure-based criteria.)

To test whether our findings capture stable behavioral tendencies, we turn to out-of-sample prediction using a leave-one-out approach. Compared to a random benchmark, our model reduces the distance between predicted and observed choice distributions by a reassuring 70 percent. Given the impracticability of a large number of types, we then turn to the question of whether a much simpler model could perform as well as our comprehensive model. Indeed, we find that a model consisting of only three types—High Promise (no other criteria), Low Action (no other criteria), and Correct Interpretation—achieves over 95% of the out-of-sample predictive power of the comprehensive model. Predictive power deteriorates, however, as we include even fewer types. The most predictive two-type model achieves just over 80% of the full model's out-of-sample predictive power, while the most predictive single-type model reaches not even 50% of that benchmark. Behavioral heterogeneity is a hallmark of choice under conflicting causal interpretations.

We examine whether individual characteristics can predict this heterogeneity. Reassuringly, we find that holding a graduate degree, being a STEM major, and having greater background knowledge about statistical causal inference all increase the chance of selecting the correct interpretation. Student subjects with a stronger belief in pseudoscience (Torres et al., 2020) select the correct interpretation less often, though we do not observe this correlation in the U.S. general population. We find no evidence that political preference predicts correct choices. While this result contrasts with the common view that disagreements with one's own political views stem from others' objective inferential errors, that view itself may be mistaken (*naïve realism*, Griffin and Ross, 1991).

Our study contributes to the rapidly growing body of research on mental models with a particular focus on the Narrative Equilibria and Model Persuasion literatures cited above.[3] The preference for high-promise interpretations is consistent with the core assumption of the Narrative Equilibria literature. Although fewer than half of the individuals in each sample display this preference, the substantial minorities that do may still be pivotal in the political economy settings the Narrative Equilibrium literature seeks to inform. Our results on data- and structure-based criteria also raise exciting new questions for the Model Persuasion literature: how do persuasion opportunities change with heterogeneous recipients who may not detect certain inconsistencies with the data, or refuse to consider interpretations that are too simple or too complex?

A small number of recent experiments study questions related to ours. None of them shares this paper's *comprehensive characterization* of decisions under conflicting causal interpretations of correlational data. Along with Barron and Fries (2023), the paper most closely related to ours, Kendall and Charles (2022), shows that externally supplied mental models that can be used to interpret raw data significantly influence individuals' choices. In a setting that qualitatively differs from ours (it displays neither recommendations nor promises to subjects, among other differences) that paper finds that conflicting interpretations cause subjects to choose intermediate actions. In contrast to that

---

[3]Izzo et al. (2023) and Horz and Kocak (2022) also study narrative equilibria albeit in different frameworks than those we address here.

paper, we include a detailed investigation of fit-checking, preferences for simpler or more complex explanation, explore different samples, and perform out-of-sample prediction. Further, Frechette et al. (2023) ask what mental models subjects spontaneously form when presented with raw data, focusing on predictions rather than causal interventions. Alysandratos et al. (2020) study choice between interpretations in hypothetical real-world settings with strong priors. Existing work also abstracts from limits to subjects' fit-checking abilities and from preferences over model structures. A more applied literature complements our inquiry into foundational human tendencies by studying mental models in specific policy domains (Andre et al., 2022, 2023a,b).

A substantial body of work in cognitive psychology studies how individuals learn causal structures and the characteristics of persuasive explanations (see Waldmann, 2017, for a review). That work differs from ours in two key aspects. First, it often allows subjects to interact with the system they are learning about. Our study, by contrast, is motivated by scenarios like policy choice where experimentation is not feasible. Hence, it requires individuals to choose based on purely observational data. Second, the psychology literature concentrates either on individual's ability to pinpoint the correct causal model when explicitly asked, or on motivated reasoning without reference to causal structures (see, for example, Kunda, 1990; Epley and Gilovich, 2016, for reviews). Unlike the present paper, it does not feature an integrated, comprehensive investigation of decisions related to causal mechanisms.

The remainder of this paper proceeds as follows. Section 2 outlines the choice setting and defines the choice criteria we study. Section 3 explains our identification strategies along with details concerning the experimental design. Section 4 showcases our main empirical results. Section 5 complements our laboratory findings with results from the U.S. general population sample. Finally, Section 6 concludes.

2 Setting and choice criteria

2.1 Choice problem

In each round of each of our experiments, subjects choose between two or three real-valued action levels A at the quadratic cost $c(A) = \frac{c}{2}A^2$. They know that the action stochastically maps into an outcome Y, and that their payoff will be $\pi(A) = Y(A) - c(A)$, but they do not know the data-generating process (DGP) that determines whether and how the action causally affects the outcome.

In Experiments 1 and 2, DGPs involve four variables, the action A, the outcome Y, and two covariates X and Z.[4] These variables are related through a recursive system of linear Gaussian equations (colloquially, a system of linear regression equations) in which A is exogenous, meaning

---

[4]This is the minimum number of variables that allows us to answer our research questions. In the simplified Experiment 3, DGPs only involve three variables, at the cost of providing coarser insight into decision-making: the action, the outcome, and a single covariate.

Table 1: Example of a recursive equation system and its DAG representation

Recursive equation system	DAG representation
$A = \beta_A + \varepsilon_A$	
$X = \beta_X + \varepsilon_X$	
$Y = \beta_Y + \beta_{AY}A + \beta_{XY}X + \varepsilon_Y$	$A \qquad Z$
$Z = \beta_Z + \beta_{YZ}Y + \varepsilon_Z,$	$X \longrightarrow Y$

Notes: The left-hand side displays an example of a recursive system of linear equations and the right-hand side displays its graphical representation. Parameters β are chosen to fit the data (in interpretations), or are given exogenously (in the DGP). Random variables ε are independent, mean-zero Gaussian errors with variances chosen to fit the data (in interpretations) or given exogenously (in the DGP). In the graphical representation of this system, no arrows point to A and X, as they are exogenous. Arrows from A and X into Y indicate that both A and X appear in the equation that describes Y, and the arrow from Y to Z indicates that Y is the only regressor in the equation that describes Z.

that is not influenced by any other variable. Other variables may be endogenous or exogenous. Throughout, our DGPs include no isolated subsets of variables and feature generic (non-knife-edge) parameters.[5] The left-hand side of Table 1 provides an example.

Subjects choose an action by picking one of two or three exogenously provided interpretations. An *interpretation* consists of a causal model, a recommendation concerning what action to take, and a promise. A *causal model* specifies a recursive system of linear regression equations that posits the causal relations among the variables. The inclusion of a variable as a regressor in an equation signifies a direct causal influence of that variable on the corresponding endogenous variable. The *recommendation* derives from an OLS fit of the model to the population moments implied by the DGP (informally: a large sample of data generated by the DGP). It is the action level that maximizes the expected payoff based on the action's inferred causal effect on the outcome. The *promise* is the corresponding expected payoff. In each round, one interpretation's model is correctly specified. All other models are misspecified.

We use directed acyclic graphs (DAG) to convey model specifications to subjects.[6] A DAG $G = (N, E)$ consists of a finite set of nodes, N, and a set of directed links $E \subset N \times N$. Nodes represent variables, so that $N = \{A, X, Y, Z\}$ in our setting.[7] Edges represent direct causation. If the DAG $G = (N, E)$ represents a given recursive system of linear Gaussian equations, and I is a regressor in the equation that corresponds to the endogenous variable J, then $(I, J) \in E$, which we also write as $I \rightarrow J$. Table 1 shows an example of the DAG representation of a recursive system of linear Gaussian equations.

---

[5]A subset of variables is *isolated* if it is neither influenced nor influences variables outside that subset. Formally, a property of a parameter vector is *generic* in our setting if it is violated only on a subset of the parameter space with Lebesgue measure zero.

[6]Every recursive linear Gaussian system can be represented as a DAG. This is a bijection up to a re-parametrization.

[7]We use the same symbols to refer to a node and the random variable it represents.

Subjects can check whether an interpretation is based on a misspecified model by comparing it to the DGPs' population moments (informally: the correlations in the data generated by the DGP). The key observational prediction of any causal models is the set of (conditional) independence relationships it implies. Because of our focus on linear Gaussian systems, (conditional) independence is equivalent to the absence of (conditional) correlation. Assessing whether a model is misspecified thus entails identifying the implied independence relationships and checking whether the corresponding empirical correlations are indeed zero. In the case of generic parameter values, one can also conclude that a model is misspecified if it predicts non-zero correlations that are zero in the data.

We frame the experiment to subjects by explaining that a *mechanism* (DGP) governs the relation between an *action* (A, depicted as a hand emoji), the number of *circles* (covariate X, depicted as a circle emoji), the number of *squares* (covariate Z, depicted as a square emoji), and a *bonus* (outcome Y, depicted as a money emoji). Subjects know that the action is costly, and that their study payment equals the bonus minus the cost of the action for one randomly selected round. Subjects select an action by deciding which one of two or three *advisors* (interpretations) to follow; that advisor's recommended action will be implemented. Because the scaling of the action is arbitrary, recommendations concern spending on the action ($c(A)$) rather than the level of the action itself (A).

Figure 1 shows the experiment's interface. The *data dashboard* on top of the screen lets subjects access *data charts* that may help them identify the correct interpretation. Subjects can observe the correlations between any pair of variables, both unconditional and conditional on holding any third variable fixed, in charts of the type shown in Figure 2.[8] Section 3.3 explains the remaining elements of the interface.

2.2 Archetypical causal models

A handful of archetypical causal models provide the key insights required for excluding misspecified models in our setting:[9]

Observation 1. *Consider a DAG $G = (N, E)$, with $N = \{I, J, K\}$.*

 (i) If $I \to J$, then generically $\mathrm{cov}(I, J) \neq 0$, and we say that I directly influences J.

 (ii) (a) If $G : I \to K \to J$ or $G : I \leftarrow K \to J$, then generically $\mathrm{cov}(I, J) \neq 0$, and we say that I indirectly influences J or a common cause *influences both I and J, respectively.*

 (b) If $G : I \to K \leftarrow J$, then $\mathrm{cov}(I, J) = 0$.

---

[8] Charts conveying conditional correlations show two data series, one for the conditioning variable held fixed at its above-median average and one holding it fixed at its below-median average. We do not show correlations that condition on pairs of variables; the experiment is designed such that such information is never needed to identify correct interpretations.

[9] These archetypical causal models employ three or fewer nodes. Appendix A.1 provides the natural formal extension of Observation 1 to the case of four-node DAGs.

Figure 1: Decision screen
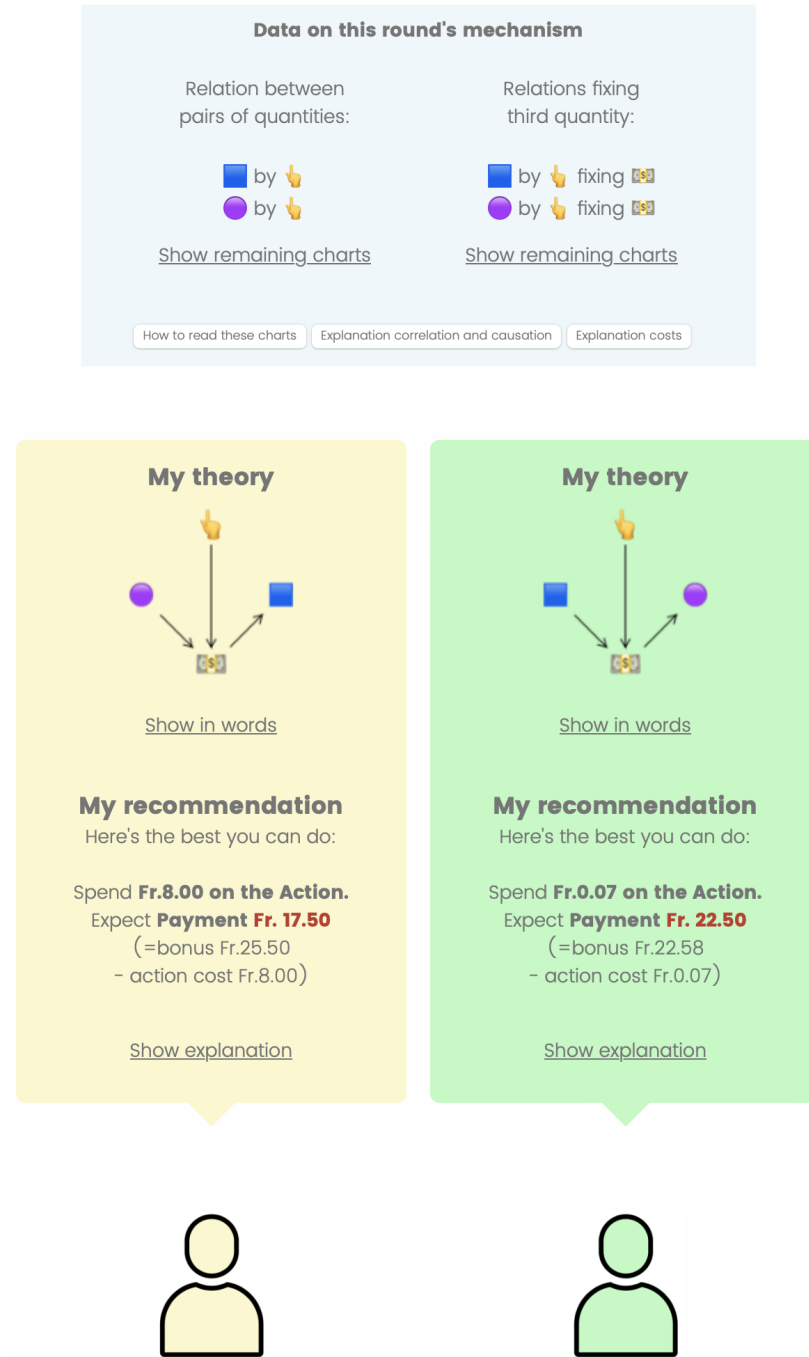
Notes: All links in the interface (underlined, buttons, and relations between quantities) are clickable. The links '■ by 👆' and '■ by 👆 fixing 💵' each open a pop-up (in the same window) that show the charts in Panels A and B of Figure 2, respectively. Other links in the data dashboard show similar charts. The links 'show remaining charts' expand the data dashboard with a list of links of the same format. For any pair of variables, and for any pair conditioning on any third variable, a data chart displaying the corresponding relation is available in every round, except when noted otherwise. For the advisor on the left-hand side, the link 'show explanation' displays the following text: 'A higher action leads to a higher bonus, when we hold the number of ● fixed. Raising the action by one raises the bonus by Fr. 0.80. Therefore you should spend Fr. 8.00 on the action. The influence on the number of ■ should not matter for your decision.' The link 'show in words' displays the following text: '👆 can directly affect 💵, ● can directly affect 💵, 💵 can directly affect ■. There are no other direct effects.' The corresponding links for all other advisors feature the same format.

9

Figure 2: Example data charts shown to subjects

A. Unconditional correlation

B. Conditional correlation
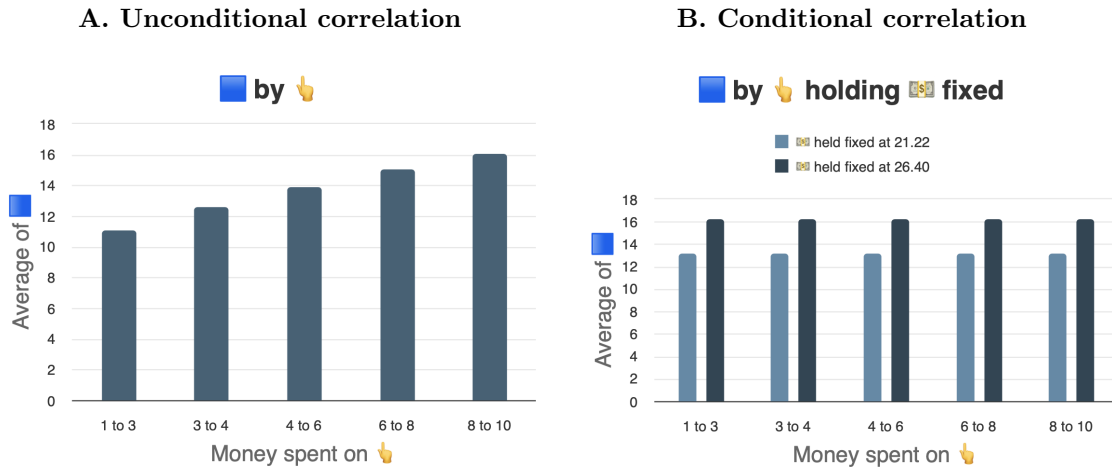
Notes: Subjects can retrieve each chart by clicking on the corresponding link in their 'data dashboard.' Panel A shows an example of a chart displaying an unconditional correlation. Panel B shows an example of a chart displaying a conditional correlation.

(iii) (a) *If $G: I \rightarrow K \rightarrow J$ or $G: I \leftarrow K \rightarrow J$, then $\text{cov}(I, J|K) = 0$, and we say that K blocks the path between I and J.*

(b) *If $G: I \rightarrow K \leftarrow J$, then generically $\text{cov}(I, J|K) \neq 0$.*

In the example of Figure 1, observation (i) is insufficient to rule out any interpretation. Each of the remaining observations can identify the correct interpretation.

To understand observations (ii)(a) and (iii)(a), if three variables are related in a *chain*, $G: I \rightarrow K \rightarrow J$, then I indirectly causes J, so the two variables are unconditionally correlated. But holding K fixed means that changes in I can no longer translate into changes in J. Hence, conditioning on K causes the correlation between I and J to vanish. In Figure 1, for example, the left-hand side advisor's model contains the chain 👆 → 💵 → 🟦. Consistent with this model, the data in Figure 2 show a nonzero unconditional correlation between 👆 and 🟦 that disappears upon conditioning on 💵.

If three variables are related in the form $G: I \rightarrow K \leftarrow J$ as in (ii)(b) and (iii)(b) (a *v-collider*), then the (unconditionally) uncorrelated variables I and J become correlated once we condition on K (the *collider node*). The intuition becomes most apparent through an example. Suppose intelligence and parental wealth are uncorrelated in the general public. Therefore, learning the magnitude of an individual's inheritance does not change our beliefs about their cognitive ability. Now, consider students at a highly reputable private university that offers two admission pathways: intelligence and parental donations. In this context, information that a student has wealthy parents suggests they may be less bright than their peers—money may be the reason they got in, in which case they did

not need to be smart. In this example, university admission is the collider node (K) that generates a conditional correlation between the two otherwise unrelated variables intelligence (I) and parental wealth (J). The right-hand side advisor's model in Figure 1, for instance, contains the v-collider 👆 → 💷 ← ▮, which is inconsistent with the data in Figure 2.

Some of these archetypical causal structures and their correlational implications are easier to understand than others. It is this variation that we use to characterize the limits to subjects' fit-checking abilities.

2.3 Decision criteria and types

Our main objective is to characterize the way in which subjects choose actions when facing conflicting interpretations of the data they observe. To achieve this, we propose a list of decision criteria and estimate how often subjects use each of them. We then collect evidence suggesting that our list contains all empirically relevant ways of decision-making in our setting.

We consider three classes of decision criteria, of which subjects may use multiple. They correspond to the three elements of our decision problems: the data, the model structure, and the *advice* (an interpretation's recommendation and promise).[10]

Data-based criteria rule out misspecified models based on correlational information in the data. Employing the principles in Observation 1 and the assumption that principle (iii) is more challenging than principle (ii), which, in turn, is more challenging than principle (i),[11] we define three criteria that correspond to the most challenging principle in use.

Definition 1.

(i) *Direct Links: Subjects rule out an interpretation if it posits a direct link between two nodes I and J but I and J are not correlated in the data.*

(ii) *Unconditional Correlations: Subjects rule out an interpretation if any of its implied unconditional correlations (or absence thereof) are inconsistent with the data.*

(iii) *Conditional Correlations: Subjects rule out an interpretation if any of its implied conditional or unconditional correlations (or absence thereof) are inconsistent with the data.*

Data-based criteria differ not only in the type of empirical correlations they employ but also in the implications of the causal structures they recognize. Both the Direct Links and Unconditional Correlations criterion, for instance, require processing information on unconditional correlations, but

---

[10]The variance of a subject's payoff does not depend on the option they select from any given menu. This is because the variance of the payoff is determined entirely by the DGP, which is assumed to be homoscedastic.

[11]While indirect causation may appear just as simple to understand as direct causation, the former requires a step of inference. In many decisions outside of the domain of causal inference, individuals often fail to draw similarly simple inferences, as conveyed by the adage, 'What you see is all there is' (Kahneman, 2011).

the latter criterion draws inferences from a larger number of them. Experiment 2 will introduce a similar distinction and consider two versions of the Conditional Correlations criterion that depend on the involvement of v-colliders.

Structure-based criteria capture preferences over model structures, especially concerning their simplicity (*cf.* Occam's razor). Our setting permits three specific interpretations of simplicity, where we refer to exogenous variables as roots.

Definition 2. *Consider two models G and G'.*

   (i) *G has greater* Root Simplicity *than G' if G has* fewer roots *than G'.*

   (ii) *G has greater* Link Simplicity *than G' if G has* fewer links *than G'.*

   (iii) *G has greater* Subset Simplicity *than G' if the links in G are a* subset *of the links in G'.*

Subjects may prefer more complex over simpler models; we call the corresponding criteria *Root Complexity*, *Link Complexity*, and *Superset Complexity*. A potential rationale for a complexity preference is the fact that a model whose links are a subset of those of another model is a restricted version of that model. In large samples, an analyst unsure about the validity of that restriction will prefer the unrestricted version.[12] Interacting a preference for or against simpler models with the three notions of simplicity yields six structure-based criteria.

Subjects' preferences may also depend on the structure of the choice set. The *Median Action* criterion encodes a preference for the model with the median recommended action. Vacuous in the two-option case, this criterion only applies to three-option menus.

Advice-based criteria reference neither data nor models. Instead, they select an interpretation based on its recommendation or promise. In a large-data context, only these criteria can satisfy the core premise of the Narrative Equilibrium literature, which posits that individuals are drawn to causal interpretations based on attributes that depend on the interpretations' prevalence in the population.[13] The reason is that the relative popularity of various interpretations does not affect the structure of interpretations or the match between predicted and observed conditional independence relationships. Therefore, data- and structure-based criteria cannot generate the required dependence.

There are four natural candidates for advice-based criteria.

Definition 3. *Let C denote the set of interpretations under consideration.*

   (i) *The* High Promise *criterion selects the interpretation with the highest promise in C.*

   (ii) *The* Low Promise *criterion selects the interpretation with the lowest promise in C.*

---

[12]Montiel Olea et al. (2022) provides related intuition.

[13]Based on action recommendations, the Median Action criterion can also satisfy the core premise.

(iii) The Low Action *criterion selects the interpretation that implies the lowest optimal action in C.*

(iv) The High Action *criterion selects the interpretation that implies the highest optimal action in C.*

The Narrative Equilibrium literature (Eliaz and Spiegler, 2020; Eliaz et al., 2022; Levy et al., 2022) operationalizes its core premise through the assumption that individuals' decisions are governed solely by the High Promise criterion.

The Low Promise criterion, in our setting, is equivalent to Gilboa and Schmeidler (1989)'s *maximin* criterion,[14] which captures much of the spirit of the empirical literature of choice under ambiguity (Trautmann and Van De Kuilen, 2015).

Because the equivalence between the *maximin* and Low Promise criteria may be unintuitive, cautious subjects may alternatively follow the heuristic to refrain from investing in an action whose returns they do not understand. Such subjects will implement the Low Action criterion. By contrast, a subject who prefers to believe in her ability to actively enhance her circumstances, for instance, due to the illusion of control (see, e.g., Stefan and David, 2013; Klusowski et al., 2021), will follow the High Action criterion.

Types We consider individuals who potentially use multiple criteria, up to one from each class, applied in the order data-based, structure-based, and advice-based.[15] We assume that individuals uniformly randomize across all interpretations that remain after applying the decision criteria.

Definition 4. *A* Type *is a triple that consists of a (possibly empty) data-based criterion, a (possibly empty) structure-based criterion, and a (possibly empty) advice-based criterion, applied in this order. Indeterminacies are resolved through uniform randomization.*

While the four data-based, eight structure-based, and five advice-based criteria (including the empty criterion in each class) combine to form 160 different types, only 111 of them are behaviorally distinct, for two reasons. First, a subject who applies the Conditional Correlations criterion consistently selects the correct interpretation and hence never reveals the structure- or advice-based criteria she would have used otherwise. Second, in our setting, all roots of a given model are uncorrelated. Hence, any two models that differ in the number of roots also differ in the unconditional correlations they imply. Once a subject has discarded interpretations based on the Unconditional Correlations criterion, all remaining models have the same number of roots, which renders the Root Simplicity and Complexity criteria vacuous.[16]

---

[14]This insight is generally true in the case of two-option menus, as we formally show in Appendix A.2. We ensure that it also holds in our three-option menus.

[15]We place the application of advice-based criteria last because each advice-based criterion determines a unique choice. Placing them first would be equivalent to placing them last and imposing the restriction that advice-based criteria cannot be combined with other criteria. We place structure-based criteria after data-based criteria because, in the tradition of Occam's razor, the former are typically used to distinguish between interpretations that cannot be separated based on the data alone.

[16]The first reason renders 40 types behaviorally equivalent. The second reason implies that for each combination of one of the 5 advice-based criteria (including the empty criterion) with the Unconditional Correlations criterion, we

3 Identification

To estimate the frequency with which subjects use each of the decision criteria, we employ two complementary identification strategies. Experiment 1 follows the usual revealed preference logic. Each subject makes a choice from each menu in a sequence of independent menus, which are structured so that the overall choice distribution reveals the distribution of criteria used. We estimate type frequencies using a finite mixture approach.

In Experiment 2, we reveal only selected elements of the decision problem to subjects. Comparing these choices with those in the case where the selected elements are hidden informs us about the fraction of subjects who use the corresponding information. For example, we estimate the proportion of subjects that effectively utilize correlational information based on how often subjects choose correct interpretation when they can access only correlational data (but not recommendations and promises) in comparison to a random choice benchmark.

The two identification approaches complement each other, given their reliance on different identifying assumptions. The finite mixture approach in Experiment 1 assumes that our set of criteria does not exclude any empirically relevant decision strategies. Experiment 2 assumes that withholding elements of the decision problem does not trigger the use of decision criteria that subjects would not otherwise have employed. If the two identification approaches yield similar estimates, we have an indication that neither approach relies on substantially inaccurate assumptions and, in particular, that our analysis considers all empirically relevant types.

Design principles Identifying the distribution of criteria in each of the three classes requires variation in, respectively, observed and implied correlations, model structures, as well as recommendations and promises. The principal design challenge consists of the fact that we generally cannot modify any single property without affecting others. For example, altering the model structure changes the implied correlations, while adjusting the DGP parameters shifts the recommendations and promises of each interpretation in a menu. A second challenge consists of the fact that no two interpretations may recommend the same action, for otherwise, subjects would not have an instrumental reason to distinguish between the available interpretations.

We construct menus to identify types based on four key insights: First, if two interpretations in a menu differ only in that the two covariates are interchanged (e.g., menu 7 in Figure 2), structure-based criteria are vacuous. This insight helps identify advice- and data-based criteria independent of structure-based confounds. Second, the relevant correlational implications of the DGP are not sensitive to its (non-knife-edge) parametrization,[17] and neither are those of any interpretation. Accordingly, once we fix a DGP and a set of causal models that determine the interpretations, we can freely select the

---

cannot distinguish between the structure-based criteria More Roots, Fewer Roots, and not using a structure-based criterion at all, which renders 3 types behaviorally equivalent in each case.

[17]Data-based criteria only concern (conditional) independence relationships. They do not depend on the magnitudes or signs of correlations.

parameters to identify advice-based criteria. Third, we can partition the set of models we consider into 15 equivalence classes. Any two models in the same equivalence class imply the same optimal action and promise for any given DGP. This fact helps us to satisfy the restriction that no two interpretations in a menu recommend the same action—all models in the menu need to stem from different equivalence classes. It also enables us to alter a menu's correlational and structural properties while maintaining constant action recommendations and promises by substituting one misspecified interpretation with another from the same equivalence class. Appendix A.3 details the construction of these equivalence classes. Fourth, for any pair of interpretations, we can choose the action distribution in the DGP to independently vary whether the interpretation that recommends the higher action makes the higher or the lower promise of the two interpretations.[18] As we show in the same Appendix section, the former case applies if and only if the mean of the action distribution exceeds some threshold. This result implies that in three-option menus, the DAG associated with the median action recommendation can never yield the highest promise.

Based on these principles, we heuristically construct candidate sequences of menus for Experiment 1. We aim for sequences that are short enough to be manageable for subjects. We choose DGP parameters that imply few negative correlations, which subjects might find more demanding to process than positive correlations, especially if chained in sequence, and we ensure that recommendations differ perceptibly.[19] To increase the statistical precision of our estimates, we further aim to create large distances between any pair of types.[20] Subsequently, we formally check whether a given candidate sequence identifies the full vector of type probabilities, as explained later (subsection 3.1). Table 2 displays the resulting sequence of menus we use in Experiment 1.

3.1 Experiment 1: Varying choice sets

We use a finite mixture model to infer criterion frequencies from the choice distributions we observe.[21] The model relies on the fact that we can predict each type's distribution of choices across the sequence of menus. Figure 3 graphically displays the predicted distributions for a selected subset of types. Summing up these distributions across all types, weighted by type frequencies, yields a prediction about the choice distribution we should observe in the overall sample. We estimate the empirical type distribution as that which minimizes the distance between the predicted and observed choice distributions. We calculate this distance using not only each individual menu (first moments) but also each pair of menus (second moments).[22] Formally, we use the generalized method of moments

---

[18] Unlike Eliaz and Spiegler (2020), we impose no requirement that relates the action distribution to subjects' choice frequency over the available interpretations.

[19] At least a handful of Swiss Francs in terms of implied spending on the action wherever possible.

[20] Appendix B.1.3 shows the distances between any pair of types.

[21] We identify the use of choice criteria in the aggregate because simulations revealed that individual-level classification yields unreliable results given the number of types we seek to estimate and the amount of noise we expect to be present in the data. Unlike individual-level classification, our aggregate identification approach lets us average out decision noise across subjects.

[22] To understand the need for second moments, suppose there are only two menus, each with two options, A and B, and three types. Type 1 chooses option A in both menus, type 2 chooses option B in both, and type 3 uniformly

Table 2: Choice sets in Experiment 1

Menu	DGP	Comp. 1	Comp. 2
1	A Z / ↓↘↑ / X→Y	A Z / ↘↑ / X→Y	
Promise	17.50	22.50	
Action	0.10	10.13	
2	A Z / ↓↘↑ / X←Y	A Z / ↓ ↑ / X→Y	
Promise	17.50	22.50	
Action	12.50	5.58	
3	A→Z / ↓ / X→Y	A→X / ↓↙ / Z→Y	
Promise	22.50	17.50	
Action	12.50	5.35	
4	A→Z / ↓ / X→Y	A→X / ↓↙ / Z→Y	
Promise	17.50	22.50	
Action	12.50	5.35	
5	A→Z / ↓↗ / X→Y	A X / ↓↗ / Z→Y	
Promise	17.50	22.50	
Action	12.48	3.45	
6	A Z / ↓ ↑ / X→Y	A X / ↓ ↑ / Z→Y	
Promise	22.50	17.50	
Action	12.50	5.13	
7	A Z / ↓ ↑ / X→Y	A X / ↓ ↑ / Z→Y	
Promise	17.50	22.50	
Action	12.50	5.13	
8	A→Z / ↘↑ / X→Y	A→X / ↘↑ / Z→Y	
Promise	22.47	17.56	
Action	8.00	0.55	
9	A→Z / ↘↑ / X→Y	A→X / ↘↑ / Z→Y	
Promise	17.50	22.51	
Action	8.00	0.55	
10	A→Z / ↘↑ / X→Y	A→X / ↘↑ / Z→Y	
Promise	17.53	22.44	
Action	3.78	10.17	
11	A→Z / ↘↑ / X→Y	A→X / ↘↑ / Z→Y	
Promise	22.50	17.50	
Action	3.78	10.16	
12	A→Z / ↘↑ / X→Y	X→Z / ↘↓ / A→Y	
Promise	22.50	17.50	
Action	12.50	3.12	
13	A→Z / ↘↑ / X→Y	X→Z / ↘↙ / A→Y	
Promise	17.50	22.50	
Action	12.50	3.12	
14	A→Z / ↘↑ / X→Y	X→Z / ↘↓ / A→Y	
Promise	22.50	17.50	
Action	3.12	12.50	
15	A Y / ↓↙ / X→Z	A Z / ↓ ↑ / X→Y	
Promise	17.72	22.25	
Action	0.00	8.38	
16	A Z / ↓↘↑ / X→Y	A→Z / ↓ / X→Y	
Promise	17.50	22.53	
Action	4.96	11.50	
17	A Z / ↓↘↑ / X→Y	A→Z / ↓ / X→Y	
Promise	22.50	17.52	
Action	4.96	11.50	
18	A Z / ↓↘↑ / X→Y	A Z / ↓ ↑ / X→Y	
Promise	22.50	17.50	
Action	3.12	5.96	
19	A Z / ↓↘↑ / X→Y	A Z / ↓ ↑ / X→Y	A X / ↓ ↑ / Z→Y
Promise	24.10	16.82	20.01
Action	12.50	2.42	7.20
20	A Z / ↓↘↑ / X→Y	A Z / ↓ ↑ / X→Y	A Z / ↘↑ / X→Y
Promise	21.19	19.83	20.48
Action	16.39	6.12	3.78
21	A Z / ↓↘↑ / X→Y	A X / ↓ ↑ / Z→Y	A Z / ↘↑ / X→Y
Promise	19.03	26.62	17.75
Action	7.03	0.68	12.50
22	A→Y / ↓↗ / X→Z	A Y / ↓↗ / X→Z	A Y / ↓↙ / X→Z
Promise	12.68	13.81	25.15
Action	11.28	6.05	0.00
23	A→Y / ↓↗ / X→Z	A Y / ↓↗ / X→Z	A Y / ↓↙ / X→Z
Promise	17.12	18.38	21.00
Action	3.12	9.73	0.00
24	A Z / ↓ ↑ / X→Y	A X / ↓ ↑ / Z→Y	A Y / ↓ ↑ / Z→X
Promise	21.52	19.76	20.44
Action	10.12	4.12	0.98
25	A Z / ↓ ↑ / X→Y	A X / ↓ ↑ / Z→Y	A Y / ↓ ↑ / Z→X
Promise	15.52	17.42	18.04
Action	10.12	6.27	5.45

Notes: Each menu shows the DGP and one or two competitor interpretations labeled 'Comp. 1' and 'Comp. 2,' along with the recommended spending on the action and the promised net payout. When combined with correlations generated by the DGP, this is all information that is required to apply each of our choice criteria. In the subjects' interface the variables of any DAG are positioned in the same way as in this table, except that subjects' DAGs are rotated clockwise by 45°.

Figure 3: Fingerprints of selected choice criteria
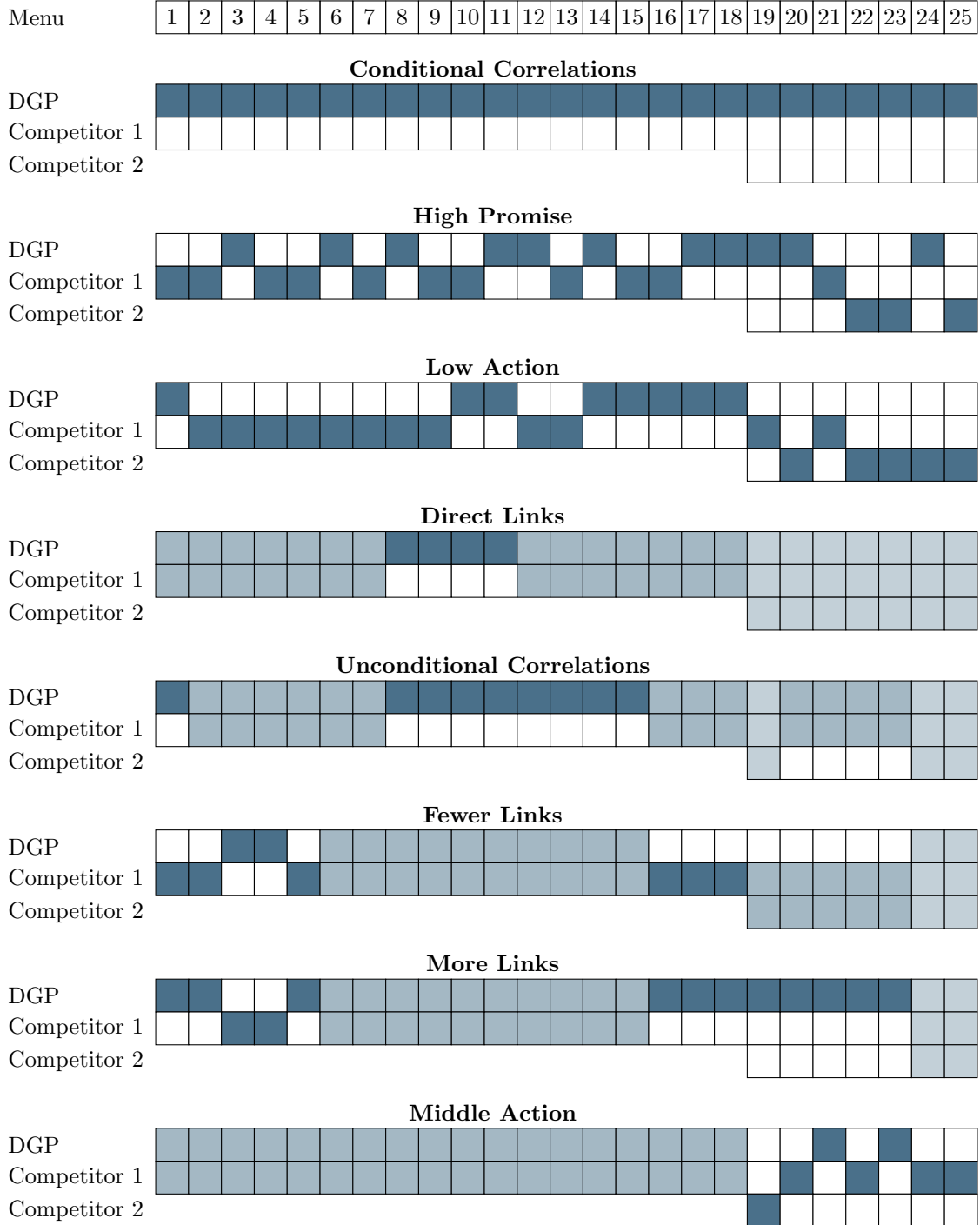
Notes: Each column corresponds to a menu. We use dark shading if the criterion chooses the corresponding option, and no shading if the rule does not choose the option. Intermediate shades indicate the number of tied options.

(GMM) with the optimal weighting matrix and heteroscedasticity-robust standard errors. Appendix B.1.1 presents details.

We allow for stochastic choice by assuming that, in any given round, a subject either chooses a type-consistent option (with probability $(1-q)$) or uniformly and independently randomizes across all available options (with probability q).[23] While we permit heterogeneity in noise probabilities across subjects, we assume that the mean noise probability is the same within each type. The stochastic choice probability and the proportion of the type that chooses uniformly randomly throughout are not separately identifiable. We interpret our results under the assumption that the latter proportion is zero.

Our mixture model has two attractive features. Both derive from the fact that, holding fixed the stochastic choice probability q, the predicted aggregate choice distribution is a linear function of the type frequencies. Linearity allows us to analytically prove identification: a sequence of menus identifies the vector of type weights if this linear function has full rank.[24] The second attractive feature is the fact that, for any fixed noise probability, the GMM objective function is a quadratic form. Hence, numerical optimization is rapid and evades the risk of converging to merely local optima, even with large type sets.[25]

3.2 Experiment 2: Withholding elements of the decision problem

In Experiment 2, we estimate the distributions of data- and advice based criteria by revealing only selected elements of the subjects' decision interface, while precluding structure-based criteria throughout.

Table 3 displays the choice sets we use. The menus in Panel A serve to estimate the frequency of data-based criteria. In each of them, we withhold advice and recommendations. Subjects generally have access to all correlational information. Subjects using the Direct Links criterion will identify the correct interpretation only in menus D1-D3. Those using Unconditional Correlations will also detect it in menus U1 and U2. Subjects who apply the Conditional Correlations criterion will pinpoint the correct interpretation throughout. Compared to Experiment 1, menus C1 and C2 make it more challenging to identify the correct interpretation in two ways. First, they give the correlational

---

randomizes. Suppose we observe that A is chosen 50% of the time in each menu (first moments). From this information alone we cannot determine whether the data was generated by a population consisting of 50% type 1 and 50% type 2 individuals, by a population consisting only of type 3 individuals, or by a convex combination of these two possibilities. This identification problem disappears once we also consider second moments. Because only type 3 may choose A in menu 1 but B in menu 2, the joint choice distribution across this pair of menus determines the proportion of type 3 individuals.

[23] Costa-Gomes and Crawford (2006); Ambuehl and Bernheim (2021) base their inference on similar assumptions. Some types generate a larger number of tied options than others. While the estimator is consistent, simulations show that in small samples, this heterogeneity may cause the estimator to assign excess weight to types with fewer ties. We choose sufficiently large study samples to render this issue negligible.

[24] We prove identification holding the noise probability fixed. We use numerical simulations to verify that our estimator correctly recovers the noise probability in synthetic data.

[25] Because these properties only hold given a fixed noise probability, we designate a grid of starting values for the noise probability, run the minimization algorithm for each of these starting points, and select the overall optimum.

implications of conditioning on a v-collider a prominent role, and second, they forestall access to the unconditional correlations. The latter would eliminate the need to interpret conditional correlations in menus involving v-colliders (as is the case throughout in Experiment 1). These menus allow us to more precisely characterize the boundaries to subjects' inferential abilities. It addresses the ceiling effects that might arise in Experiment 1 if many subjects consistently identify the correct interpretation. Menus NV1 and NV2, by contrast, are at the same level of difficulty that we use to identify the Conditional Correlations criterion in Experiment 1.

To interpret choices given the additional inferential challenges in this experiment (menus C1 and C2), we introduce the *Conditional Correlations (No V-colliders)* subcriterion. As the Conditional Correlations criterion, it rules out interpretations whose implied conditional or unconditional correlations (or absence thereof) are inconsistent with the data, except in menus that involve v-colliders and prevent access to unconditional correlational information.

The menus in Panel B identify advice-based criteria. They withhold correlational information and let subjects only observe the model structures (which, by design, cannot aid in decision-making) and the advice. In each pair of menus A1-A3, we vary whether the high-promise interpretation coincides with the high or low recommendation. The inferences we can draw from these decisions differ from those in Experiment 1. First, Experiment 2 reveals advice-based choices for all subjects, whereas Experiment 1 only reveals them for subjects who do not consistently select the correct interpretation. This fact prevents a direct comparison of the distributions of advice-based criteria across the two experiments.[26] Second, we can only infer the difference in the fraction of subjects using the High versus Low Promise criteria, but not the proportion of subjects using any single one of these criteria. The reason is that while we only observe the overall fraction of subjects choosing one rather than the other interpretation, displaying the promises draws subjects following the High Promise criterion toward the high-promise interpretation but pushes those following the Low Promise criterion away from it. In case of an equal number of High and Low Promise subjects, for instance, the overall choice frequency of the high-promise interpretation will thus not differ from the random benchmark. For a parallel reason, we only observe the difference in the proportions of subjects using the High versus Low Action criteria, but not the proportions themselves.

We formally estimate the frequency of the data-based criteria by assuming that t_D, t_U, t_{NV} and t_C describe the frequencies with which subjects use the Direct Links, Unconditional Correlations, Conditional Correlations (No v-Colliders), and Conditional Correlations criteria, respectively, and that each subject chooses uniformly randomly with probability q in any given menu. Letting p_D, p_U, p_{NV} and p_C denote the observed frequencies with which subjects choose the correct interpretation in menus D, U, NV and C, respectively, we deduce the frequencies of data-based criteria as follows (see Appendix B.2):

---

[26]Experiment 2 does not provide a straightforward way of conditioning on subjects who do not use the Conditional Correlations criterion as the two corresponding binary choices in menus C1 and C2 do not yield sufficient statistical precision for reliable individual-level classification.

Table 3: Choice sets in Experiment 2

A. Identification of data-based criteria

Menu	DGP	Competitor	Predicted correlations	
			DGP	Competitor

Identification of the Direct Links criterion

Menu	DGP	Competitor	DGP	Competitor
D1	$A \to Z$ ↘↑ $X \to Y$	$A \to X$ ↘↑ $Z \to Y$	$\text{cov}(A,X) = 0$ $\text{cov}(A,Z) \neq 0$	$\text{cov}(A,X) \neq 0$ $\text{cov}(A,Z) = 0$
D2	$A \to X$ ↘↑ $Z \to Y$	$A \to Z$ ↘↑ $X \to Y$	$\text{cov}(A,X) \neq 0$ $\text{cov}(A,Z) = 0$	$\text{cov}(A,X) = 0$ $\text{cov}(A,Z) \neq 0$
D3	$A \quad Z$ ↓↙↓ $X \to Y$	$A \quad X$ ↓↙↓ $Z \to Y$	$\text{cov}(A,X) \neq 0$ $\text{cov}(A,Z) = 0$	$\text{cov}(A,X) = 0$ $\text{cov}(A,Z) \neq 0$

Identification of the Unconditional Correlations criterion

Menu	DGP	Competitor	DGP	Competitor
U1	$X \to Z$ ↘↑ $A \to Y$	$Z \to X$ ↘↑ $A \to Y$	$\text{cov}(A,X) = 0$ $\text{cov}(A,Z) \neq 0$	$\text{cov}(A,X) \neq 0$ $\text{cov}(A,Z) = 0$
U2	$Z \to X$ ↘↑ $A \to Y$	$X \to Z$ ↘↑ $A \to Y$	$\text{cov}(A,X) \neq 0$ $\text{cov}(A,Z) = 0$	$\text{cov}(A,X) = 0$ $\text{cov}(A,Z) \neq 0$

Identification of the Conditional Correlations (No v-Colliders) criterion

Menu	DGP	Competitor	DGP	Competitor
NV1	$A \to Z$ ↓ $X \to Y$	$A \to X$ ↓ $Z \to Y$	$\text{cov}(Z,Y|A) = 0$ $\text{cov}(Z,Y|X) = 0$ $\text{cov}(A,Y|X) = 0$ $\text{cov}(X,Y|A) \neq 0$ $\text{cov}(X,Y|Z) \neq 0$ $\text{cov}(A,Y|Z) \neq 0$	$\text{cov}(Z,Y|A) \neq 0$ $\text{cov}(Z,Y|X) \neq 0$ $\text{cov}(A,Y|X) \neq 0$ $\text{cov}(X,Y|A) = 0$ $\text{cov}(X,Y|Z) = 0$ $\text{cov}(A,Y|Z) = 0$
NV2	$A \quad Z$ ↓ ↑ $X \to Y$	$A \quad X$ ↓ ↑ $Z \to Y$	$\text{cov}(A,X|Z) \neq 0$ $\text{cov}(A,Y|Z) \neq 0$ $\text{cov}(A,X|Y) \neq 0$ $\text{cov}(A,Y|X) = 0$ $\text{cov}(A,Z|X) = 0$ $\text{cov}(A,Z|Y) = 0$	$\text{cov}(A,X|Z) = 0$ $\text{cov}(A,Y|Z) = 0$ $\text{cov}(A,X|Y) = 0$ $\text{cov}(A,Y|X) \neq 0$ $\text{cov}(A,Z|X) \neq 0$ $\text{cov}(A,Z|Y) \neq 0$

Identification of the Conditional Correlations criterion

Menu	DGP	Competitor	DGP	Competitor
C1	$A \quad X$ ↘↑ $Z \to Y$	$A \quad Z$ ↘↑ $X \to Y$	$\text{cov}(A,X|Y) = 0$ $\text{cov}(A,Z|Y) \neq 0$	$\text{cov}(A,X|Y) \neq 0$ $\text{cov}(A,Z|Y) = 0$
C2	$A \quad Z$ ↘↑ $X \to Y$	$A \quad X$ ↘↑ $Z \to Y$	$\text{cov}(A,X|Y) \neq 0$ $\text{cov}(A,Z|Y) = 0$	$\text{cov}(A,X|Y) = 0$ $\text{cov}(A,Z|Y) \neq 0$

B. Identification of advice-based criteria

Menu	DGP			Competitor		
	Model	Promise	Recommendation	Model	Promise	Recommendation
A1a	$A \to Z$ ↓↗ $X \to Y$	p_H	a_H	$A \to X$ ↓↗ $Z \to Y$	p_L	a_L
A1b		p_L	a_H		p_H	a_L
A2a	$A \quad Z$ ↓↗↑ $X \to Y$	p_H	a_H	$A \quad X$ ↓↗↑ $Z \to Y$	p_L	a_L
A2b		p_L	a_H		p_H	a_L
A3a	$A \quad Z$ ↓ ↓ $X \to Y$	p_H	a_H	$A \to Z$ ↓ $X \to Y$	p_L	a_L
A3b		p_L	a_H		p_H	a_L

Notes: In the choice sets in panel A, subjects do not have access to promises and recommendations. It displays all correlations for which the implications of the two models differ and that are accessible to subjects. In the choice sets in panel B, subjects do *not* have access to correlational information from the DGP.

$$\begin{bmatrix} t_C \\ t_{NV} \\ t_U \\ t_D \end{bmatrix} = \frac{2}{1-q} \begin{bmatrix} p_C - \frac{1}{2} \\ p_{NV} - p_C \\ p_U - p_{NV} \\ p_D - p_U \end{bmatrix}. \tag{1}$$

For econometric inference about advice-based criteria, we let Δ_P denote the difference in the prevalence of the High and Low Promise criteria, and Δ_A the difference in that of the Low and High Action criteria. We assume that subjects who use neither of these criteria randomize uniformly across these menus. Letting p_a denote the choice proportion of the high-promise interpretation when it coincides with the high action (menus $A1a, A2a, A3a$), and p_b that in the remaining menus ($A1b, A2b, A3b$), we infer (see Appendix B.2)

$$\begin{bmatrix} \Delta_P \\ \Delta_A \end{bmatrix} = \frac{1}{1-q} \begin{bmatrix} p_a - p_b \\ 1 - p_a - p_b \end{bmatrix}. \tag{2}$$

While the noise parameter q affects our inference, Experiment 2 does not identify it. Hence, we first infer criterion frequencies assuming $q = 0$ (no errors) and then scale the inferred frequencies using the estimate of q from Experiment 1.

3.3   Experiment design details

Subjects make choices in the interface displayed in subsection 2.1. The data charts subjects can access do not indicate any statistical uncertainty because they display expected values and because the public discourse rarely features such information. The data dashboard shows links to data charts either overtly or behind the link 'show remaining charts.' Subjects know that overt links list all correlations for which the current rounds' advisors' causal models have different implications; hidden links show the remaining data. A link 'explanation correlation and causation' opens a page with intuitive explanations of the information listed in Observation 1; Appendix E.1 reproduces it in full. (To examine the behavioral effect of these simplifying elements, Experiment 3 in Section 5 includes a treatment in which all links are overtly shown and the latter page is not available.) Further links describe the causal structure in words ('show in words'), explain how the observed correlations inform the recommendation ('show explanation,' reproduced in Appendix E.1), detail the structure of the unconditional and conditional data charts ('how to read these charts'), and display a graph with the cost of each action ('explanation costs').

The instructions stress that exactly one advisor is correct in each round, that the data do not affect advisors' model specifications, that recommendations and promises derive from fitting the model to the

data, that there are no errors in fitting any model to the data, but that promises and recommendations based on misspecified models are wrong nonetheless. They also explain that any exogenous nodes are mutually independent, and that the action is always exogenous in the data displayed in the charts.[27]

A comprehension check that is hard to pass by chance ensures that subjects understand all elements of the experiment. Subjects can proceed with the experiment only if they answer correctly. We ask subjects who fail the check to revisit the instructions until they can pass. Because subjects in a pilot study spent twice as long on their first decision as on later ones, the experiment begins with two preliminary rounds (not identified as such) whose data we do not analyze.

We randomize elements of the experiment to increase the plausibility of our stochastic choice assumptions. By having subjects proceed through the menus in individually randomized order, we ensure, for instance, that even if subjects stop paying attention after some time, this lack of attention is uniformly distributed across the menus. Randomization of the order in which the interpretations appear on screen guarantees that behaviors such as consistently selecting the advisor on the right appear as uniform randomization across interpretations.

Toward the end of the study, we elicit an array of individual characteristics that we relate to decision criteria use. These include information that lets us classify each subject's field of study as STEM, economics and business, or other, as well as information on subjects' familiarity with concepts underlying probabilistic causal inference (completing the aphorism "correlation does not...", writing the name of the mathematical object $P(A|B)$ in words, spelling out the acronym 'DAG,' and reporting whether they have ever taken a class on causal statistical inference).[28] We also elicit risk preferences using the approach of Eckel and Grossman (2008) (incentivized), administer an extended version of the Cognitive Response Test (Frederick, 2005; Toplak et al., 2014), and measure subjects' beliefs in pseudoscience (Torres et al., 2020). Subjects report their gender and the Swiss political party they consider closest to their own views, which we assign a position on the political spectrum based on Jolly et al. (2022).[29] Finally, subjects describe in their own words how they typically made decisions in the main rounds of the experiment.

Subjects are free to leave once they have finished the study, at which time they receive a completion payment of Fr. 20, as well as the payoff from one randomly selected round of the study (including risk-preference elicitation).[30] We choose DGP parameters such that advisors' promises range from Fr. 10 to Fr. 30 (though recommendation based on a misspecified model can lead to a far lower payout than that model promises), for incentive payments ranging from Fr. 1.40 to Fr. 24.10 in Experiment 1 and from Fr. 8.50 to Fr. 22.50 in Experiment 2.

---

[27]To avoid overwhelming subjects, we do not explicitly communicate that DGPs are linear Gaussian. Linearity of the relations between the variables X, Y, and Z becomes apparent to subjects inspecting the data charts. Charts involving the action plot spending $c(A)$ on the horizontal axis, which is concavely related to the remaining variables.

[28]We score the first three questions by whether they include the strings 'caus', 'conditional' or 'given,' and 'acyc,' respectively.

[29]That survey does not include a score for the Swiss communist party ("Partei der Arbeit"). We assign it a score of 0 (leftmost possible).

[30]At the time of the study, 1 Fr. = \$1.12.

All further design details are listed in Appendix C.1.

4 Analysis

We begin by explaining our data collection and sample characteristics (subsection 4.1). Subsection 4.2 contains our main estimates of the frequency with which subjects' use the various decision criteria in Experiment 1. Subsection 4.3 demonstrates the robustness of our results using out-of-sample prediction exercises. Subsection 4.4 shows that we estimate a highly similar distribution of data-based criteria using a different identification strategy with a separate sample of subjects in Experiment 2. Subsection 4.5 searches for parsimonious models whose predictive accuracy rivals that of our unrestricted model with 111 types. Subsection 4.6 relates decision-making in our setting to subjects' educational background, political preferences, and demographic and psychological traits.

4.1 Data collection and preliminary analysis

We ran Experiment 1 with 485 subjects earning a mean payment of Fr. 38.30 and Experiment 2 with 279 subjects earning a mean payment of Fr. 37.50 in April and May 2023 at the Laboratory for Experimental and Behavioral Economics at the University of Zurich.[31] Subjects took between 50 minutes and 2 hours to complete the experiment, with a median completion time of around 75 minutes. Per round, the median subject spend 38 seconds in Experiment 1 and 20 seconds in Experiment 2 (means 56 and 40 seconds, respectively). The shorter decision times in Experiment 2 may be due to the fact that rounds in which less information is available can be completed more quickly.[32]

Approximately two-thirds of our subjects are enrolled in a STEM major, another 13% are enrolled in economics or business, and 18% report having taken a class on statistical causal inference. Yet, only 55% can complete the aphorism 'Correlation does not...,' one quarter can name '$P(A|B)$' and 9% can spell out 'DAG.' Appendix D.1 lists further sample characteristics.

Subjects use the data dashboard frequently. In Experiment 1, 90.5% of subjects access it at least once (excluding the two preliminary rounds). In an average round, 65.9% of subjects view at least one chart. While any single one of the overtly displayed charts can be used to determine the correct interpretation in each two-alternative menu, the average subject views 2.86 charts per round.[33]

---

[31]We preregistered a target sample size of 700 subjects, https://www.socialscienceregistry.org/trials/11336.

[32]Appendix D.2 examines order effects. While there are minor order effects in terms of time spent per round, these appear to reflect learning rather than fatigue, as subjects view data charts at the same rate throughout the experiment.

[33]One potential concern is the possibility that subjects determine their choice simply by looking at the unconditional correlation between the action and the outcome, which equals the true causal effect of the action. In fact, in Experiment 1, there is not a single round in which a subject viewed the unconditional correlation between A and Y, but did not view any other chart.

4.2 Distribution of decision criteria: Experiment 1

Figure 4 shows the overall distribution of choices for each menu. It provides initial insights about the nature and quality of the data. Three features stand out. First, while choice distributions place significant weight on all interpretations in most menus, some are far from uniform (e.g. 11, 21 and 25). Hence, the dispersion in choices present in some menus reflects individual heterogeneity rather than general inattention to the experiment. Second, aggregate patterns anticipate some of our overall conclusions. When the choices of the High Promise, Low Action, and Conditional Correlation criteria coincide (menus 11, 14, 17, and 18; see Figure 3), choices concentrate substantially on the corresponding prediction, suggestive of great popularity of these three criteria. Third, subjects do not simply resort to a choose-the-middle-action heuristic even when they could. In fact, in menus 24 and 25, the interpretation with the median recommendation is chosen least often by a wide margin.

Figure 4: Aggregate choice distribution in Experiment 1

Menu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
DGP	72	49	56	43	56	65	48	68	55	68	84	61	51	73	66	74	85	82	60	49	54	50	45	45	38
Competitor 1	28	51	44	57	44	35	52	32	45	32	16	39	49	27	34	26	15	18	18	21	39	20	18	16	15
Competitor 2																			23	30	7	30	37	39	47

Notes: Each column corresponds to a menu, listed in the same order as in Figure 3. Numbers list the percentage of subjects choosing a given option. Shades of blue reflect the percentages.

We next use our mixture model to estimate criterion frequencies. We consider the fraction of subjects who make use of a given criterion at some stage in their decision process. Accordingly, we sum the estimated weights of all types that use a given criterion regardless of any other criteria they may combine it with.

Before summing across types, we inspect the weights assigned to the three most common individual types. Accounting for a total weight of 57.7%, each of them uses a single criterion that they do not combine with other criteria. The most common type (37.4%) pinpoints the correct narrative throughout.[34] The second most common type (13.1%) consistently chooses the advisor recommending the lowest action. The third most common type (7.2% of subjects) chooses the high-promise interpretation without using any data- or structure-based criteria. Appendix D.3 lists the estimated frequency of each type. To characterize the choice behavior of all subjects, including the 42.3% not described by the three most popular types, we now turn to criteria distributions.

Panel A in Figure 5 displays the popularity of each data-based criterion; Table 4 lists the numerical estimates (column 1 of Panel A). While the modal subject (46.1%) does not use any data-based criterion, the 37.4% who pinpoint the correct interpretation throughout (the Conditional Correlations criterion) constitute the majority of those who use data-based criteria (53.9%). Another 5.8% make

---

[34]Experiment 2 shows that this estimate partly reflects a ceiling effect.

Figure 5: Distribution of decision criteria

A. Data-based criteria

B. Advice-based criteria

C. Structure-based criteria
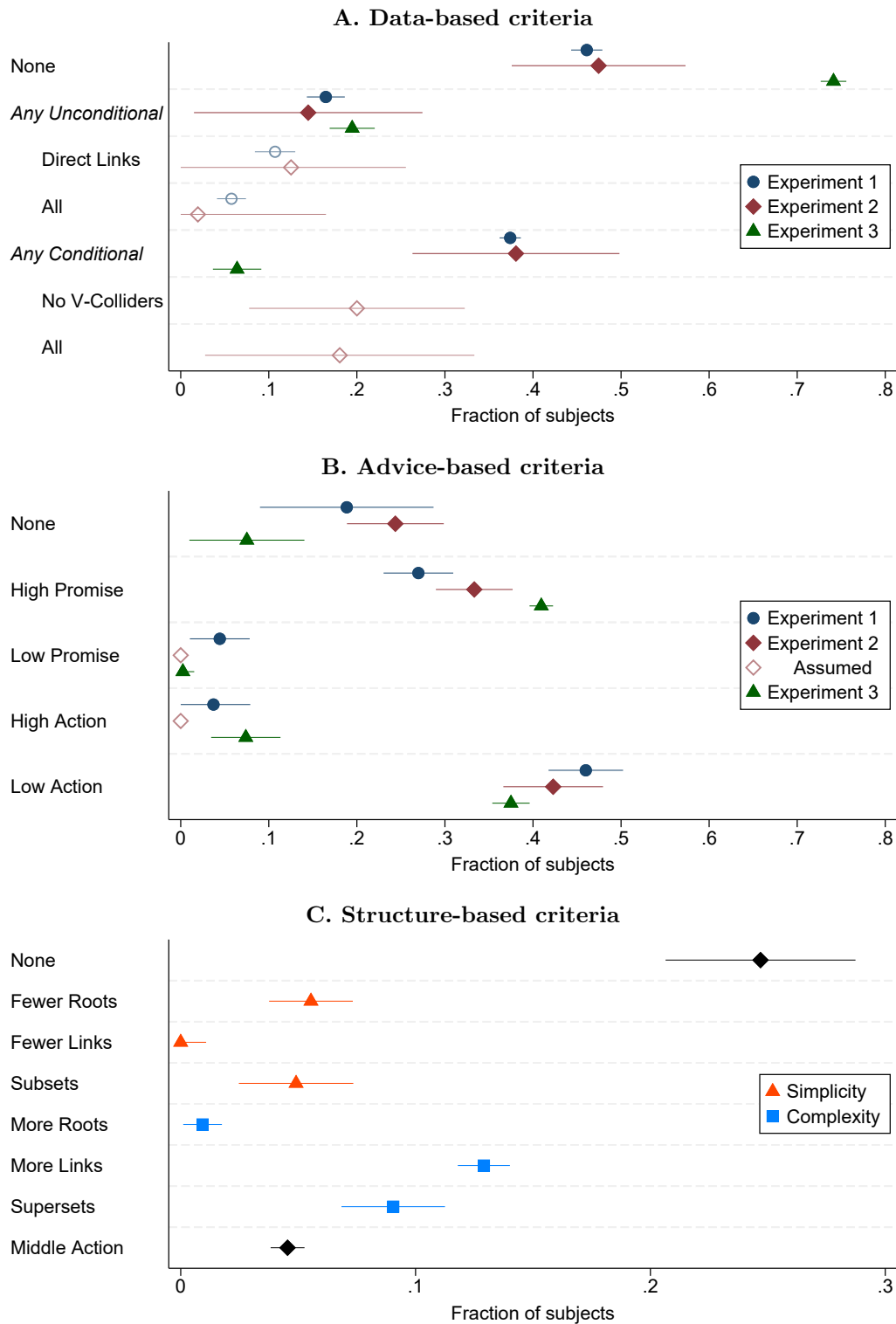
Notes: Whiskers show 95%-confidence intervals, truncated at 0. *Panel A:* Experiments 1 and 3 do not separately identify the Conditional Correlations (No v-Colliders) and Conditional Correlations criteria. Experiment 3 does not separately identify the Direct Links and Unconditional Correlations criteria. Experiment 2 estimates of data-based criteria are adjusted for noisy choice using $q = 0.266$. *Panel B:* Estimates of advice-based criteria in Experiment 1 are shown conditional on not using the Conditional Correlations criterion. Experiment 2 only identifies the difference between the High and Low Promise criteria and between the High and Low Action criteria, not the four levels. We set the prevalence of the Low Promise and High Action criteria to zero and plot estimates using $q = 0$. *Panel C:* Only Experiment 1 identifies structure-based criteria. The 37.4% of subjects who use the Conditional Correlations criterion do not reveal structure-based criteria.

data-informed choices as long as unconditional correlations suffice to do so (the Unconditional Correlations criterion), and an additional 10.7% do so if the Direct Links criterion is sufficient to discard misspecified interpretations. Hence, not only is there limited fit-checking among a significant portion of individuals, but there is also substantial heterogeneity in these limits.

Panels B of the Figure and Table showcase the distribution of advice-based criteria. The frequencies in the table sum to 62.6% because the 37.4% of subjects who consistently identify the correct interpretation do not reveal any advice-based criteria; those in the figure condition on not using the Conditional Correlations criterion. Ostensibly cautious choice is popular: the modal advice-based criterion is to choose the advisor whose action recommendation is lowest (28.8%). In spite of the formal connection between the Low Promise and *maximin* criteria, a vanishing fraction of subjects (2.8%) use that criterion.[35] Even fewer participants systematically prefer high actions (2.3%). The High Promise criterion describes 16.9% of our subjects. This fraction of maxi*max*-decisions is considerable from the viewpoint of the literature on cautious choice and from the viewpoint of a literature that rarely evidences behavior consistent with anticipatory utility in laboratory experiments (reviewed in Engelmann et al., forthcoming). From the viewpoint of the narrative equilibria literature (Eliaz and Spiegler, 2020), which assumes that this criterion describes everyone, the fraction is low. As Section 5 shows, the High Promise criterion garners far more support in Experiment 3.

Panels C of the Figure and Table exhibit the relative popularity of structure-based criteria. These, too, are observable only for the 62.6% of subjects who do not consistently choose the correct interpretation.[36] A mere 10.4% of subjects prefer simpler interpretations in the form of fewer roots (5.5%) and subsets of links (4.9%), but not in the form of fewer links (0.0%). More than twice as many (22.8%) favor more complex interpretations, especially those with more links (12.9%) and those whose links are a superset of an alternative interpretation (9%). The difference between the proportion of subjects with a complexity preference and those with a simplicity preference is highly statistically significant ($p < 0.01$). While this result is unexpected in light of Occam's razor, it may reflect the conceptual appeal of the Superset criterion outlined in Section 2.3 which the More Links criterion heuristically approximates.[37] Conceptually, caution about imposing potentially incorrect model restrictions does not support a preference for more roots, which, tellingly, subjects do not endorse (0.9% of subjects). Finally, we estimate that only 4.6% of subjects systematically choose the interpretation with the median action recommendation.

Due to randomness, a subject's choices do not always conform to their type. The estimated random choice parameter of 26.6%, however, indicates that participants paid attention to their choices and

---

[35]The formal connection between the *maximin* and Low Promise criteria is not obvious. Our interface does not provide subjects with the information to apply the *maximin* criterion directly. Hence, the low support for the Low Promise criterion should not be construed as evidence against maximin behavior in general.

[36]Moreover, as detailed in Section 2.3, the Fewer Roots and More Roots criterion cannot be used to discard interpretations once interpretations based on ill-fitting unconditional correlations have been rejected. Therefore, we set the fraction of types that combine the Unconditional Correlations criterion with one of the Root-based criteria to zero.

[37]Relatedly, Marsh et al. (2022) provide evidence that conspiracy theories tend to be disproportionately complex.

Table 4: Distribution of criteria use

Criteria	(1) Experiment 1	(2) Experiment 2	(3) Experiment 2
Adjusted for noise	-	No	Yes
A. Data-based			
None	0.461	0.614	0.474
	(0.009)	(0.037)	
Any Unconditional	0.165	0.106	0.145
	(0.011)	(0.049)	
Direct Links	0.107	0.092	0.125
	(0.012)	(0.049)	
All	0.058	0.014	0.020
	(0.008)	(0.054)	
Any Conditional	0.374	0.280	0.381
	(0.006)	(0.044)	
No V-Colliders	-	0.133	0.181
		(0.057)	
All	-	0.147	0.200
		(0.046)	
B. Advice-based			
None	0.118	-	-
	(0.020)		
High Promise	0.169	0.333	0.454
	(0.008)	(0.025)	
Low Promise	0.028	$0^{\dagger}$	$0^{\dagger}$
	(0.007)		
High Action	0.023	$0^{\dagger}$	$0^{\dagger}$
	(0.009)		
Low Action	0.288	0.423	0.576
	(0.009)	(0.029)	
C. Structure-based			
None	0.247	-	-
	(0.021)		
Simplicity			
Fewer Roots	0.055	-	-
	(0.009)		
Fewer Links	0.000	-	-
	(0.006)		
Subsets	0.049	-	-
	(0.012)		
Complexity			
More Roots	0.009	-	-
	(0.004)		
More Links	0.129	-	-
	(0.006)		
Supersets	0.090	-	-
	(0.011)		
Middle Action	0.046	-	-
	(0.004)		
Random choice probability	0.266	$0^{\dagger}$	$0.266^{\dagger}$
	(0.008)		
Subjects	475	279	-
Observations	11875	4185	-

Notes: † indicates imposed values. Dashes indicate that the corresponding experiment or model does not estimate the listed criterion frequency. Standard errors in parentheses, heteroskedasticity-robust in column 1, and clustered by subject in column 2. Estimates in column 1 represent the output of a single estimation. Estimates in column 2 reflect the output of two regressions, one for data-based criteria (OLS, 2511 observations from 279 subjects) and one for advice-based criteria (two-equation stacked OLS, 1674 observations from 279 subjects), with $q = 0$. Estimates in column 3 present the corresponding estimates using $q = 0.266$.

suggests that out-of-sample predictability will substantially exceed a random choice benchmark, as we will examine in Subsection 4.3.

Overall, these results document, first, pronounced dispersion in the willingness and ability to assess the fit between interpretations and observable data, second, some support for the core assumption of the narratives equilibria literature and substantial support for a notion of caution not yet considered in the literature (the Low Action criterion), and, third, a considerable preference for more complex over simpler interpretations.

One potential concern with these findings is that our experiment may not measure subjects' natural comprehension and use of the correlational implications of causal structures because it provides an explainer about these linkages. Experiment 2 presents evidence indicating that this concern is unwarranted—subjects rarely open the explainers, and they do not open them more often in more difficult problems. Experiment 3 dispels the concern in a different subject population using a treatment that does not provide this explainer to subjects, and that, in addition, overtly displays links to all data charts regardless of whether they help distinguish between the available interpretations.

A second concern questions whether subjects' misunderstandings are confined to the correlational implications of causal structures (which we explicitly model) or whether they reflect more fundamental misperceptions about the decision environment in our study (from which we abstract). While our mandatory comprehension checks minimize the latter possibility, subjects' answers to the open-ended question of how they typically chose between advisors provide additional insight. Many subjects describe approaches consistent with our criteria.[38] Yet, 33 subjects (6.8%) appear confused about the fact that a promise or recommendation based on a misspecified model carries no information. These subjects commonly describe a strategy of comparing the difference in promised outcomes across the interpretations to the difference in the costs of the recommended actions, in an apparent attempt at cost-benefit analysis.[39] We obtain highly similar estimates when we estimate our finite mixture model excluding these subjects, as Appendix D.4 shows.

These results abstract from priors about the plausibility of the models in a menu. Appendix D.5 considers decisions in three different real-world framings and shows that such beliefs would indeed confound the identification of our decision criteria. The same Appendix section also considers the effects of conveying models verbally rather than graphically in two selected menus. While the verbal

---

[38]For instance: Conditional Correlations (*"I tried to look at all the given graphs. If there were a a1 → a2 → a3 path, I especially looked at the graph where a2 was fixed"*), Unconditional Correlations (*"I tried to find quantities which do not influence each other. I prefered the realtion between pairs instead of the fixing third quantities. Then I tried to find data which does not fit into the statement of one of the advisors. I never looked at the amount of money."*), High Promise (*"I only glanced at the data provided. In most rounds i chose the advisor by looking whose theory would result in the highest expected payment if they were right."*), High Action (*"[I] never choose the actions that where the cheapeast because theirs a spanisch saying that allways the cheapeast becomes the expensiest."*), or choosing randomly after partial data checking (*"sometimes i choose randomly because i didnt know the answer"*).

[39]Example statements are *"I always looked at how much the action would cost in relation to the bonus that the advisor expected"* and *"I was comparing the amount I spend with the bonus I eventually get"*. A research assistant classified each subject's response according to whether it shows definite evidence of the misunderstanding or potential evidence. We report the numbers that include merely potential evidence for misunderstanding. In Experiment 2, 25 subjects (9.0%) report such strategies.

presentation substantially raises response times, it does not have statistically significant effects on the likelihood with which subjects choose any given interpretation at the 5% level.

4.3   Out-of-sample predictions

The large number of types in our mixture model may raise questions about overfitting. We address these by testing the out-of-sample predictive power of our model. We designate a training set of menus, which we use to estimate the type distribution, as well as a test set, which we use to evaluate predictive success. We score predictions by the Euclidian distance between the predicted and actual choice distributions on the test set. We designate training- and test-sets using the leave-one-out approach: one menu constitutes the test set and all remaining menus make up the training set. While there is one menu (menu 21 of Table 2) that must be included in every training set to guarantee the identification of all types, we repeat the procedure using each of the remaining menus as test set once, and average the predictive scores.

Figure 6 shows the results. Each blue circle represents the alternative predicted to be more popular in a two-alternative menu. Its horizontal position indicates the predicted choice frequency, and its vertical position indicates the observed frequency. Each three-option menu is represented by two symbols: red triangles for the interpretation predicted to be most popular, and green rectangles for the one predicted to be second-most popular. We observe tight clustering around the diagonal, which demonstrates substantial out-of-sample predictive power.

Figure 6: Out-of-sample predictive power
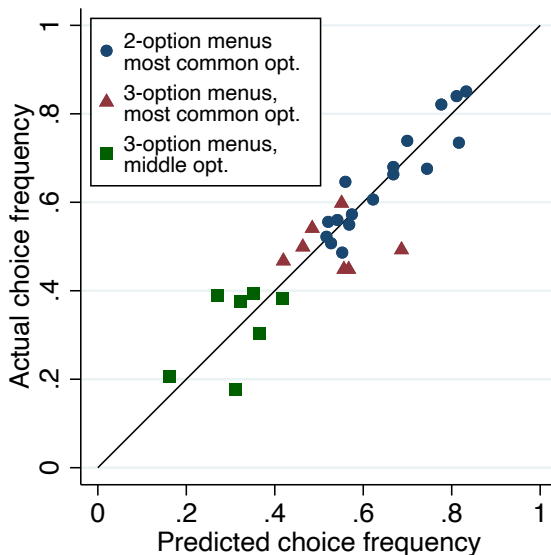
Notes: Each blue circle represents a two-option menu with the predicted choice frequency of the option predicted to be more popular on the horizontal axis and the corresponding observed choice frequency on the vertical axis. Red triangles correspond to the option predicted to be most popular in three-option menus and green squares to the option predicted to be second-most popular.

Panel A of Table 5 shows the formal results. We attain an average distance of 0.031 between predicted and actual choice distributions. Benchmarked against the distance of 0.101 between the uniformly random and actual choice distributions, our model predictions correspond to an improvement of 69.3%. This substantive out-of-sample predictive power indicates that our estimates in Subsection 4.2 are not simply a consequence of overfitting.

4.4 Distribution of decision criteria: Experiment 2

Experiment 2 provides a robustness check of the results from Experiment 1. It also lets us more precisely characterize the bounds of subjects' inferential abilities. In contrast to Experiment 1, it excludes structure-based criteria and it only identifies the popularity of criteria rather than types (combinations of criteria).

Table 4 lists the estimates of criterion frequencies we obtain from Experiment 2. Column 2 assumes zero noise; column 3 uses the noise parameter of 0.226 estimated in Experiment 1. We first consider data-based criteria, plotted in Panel A of Figure 5, which shed light on the boundaries of subjects' inferential abilities. Of the subjects who use conditional correlational information effectively at the level of difficulty that corresponds to Experiment 1 (38.1%), only about half (20% of all subjects) also manage to do so in menus that involve v-colliders and prevent access to unconditional correlations. One

candidate explanation for the lower rate of correct choices in those menus is the mistaken belief that uncorrelated parents of a collider remain uncorrelated even when conditioning on the collider. Under this misperception, a participant facing menu $C1$ will interpret the observed statistic $\mathrm{cov}(A, X|Y) = 0$ as consistent with both available interpretations, $\mathrm{cov}(A, Z|Y) \neq 0$ as equally inconsistent with both, and will thus choose randomly. This finding shows that subjects struggle not only with reading data charts but also with deducing a theory's correlational implications and connecting these to the data.

We study the robustness of our main results by comparing the estimates from Experiment 1 with the noise-adjusted estimates from Experiment 2.[40] They coincide to a surprising extent; the largest discrepancy is the marginally lower weight of the Unconditional Testing criterion in Experiment 2. The agreement between the estimates suggests that our set of types in Experiment 1 does not omit empirically relevant structure- and advice-based criteria. Otherwise, Experiment 1, but not Experiment 2, would misattribute the choices of omitted types to the included types, likely bringing the estimated distributions of advice-based criteria out of alignment across the experiments.[41]

The juxtaposition of the estimates from the two experiments also addresses whether subjects in Experiment 1 use advice- and structure-based criteria as cognitively inexpensive substitutes for fit-checking. If so, Experiment 2 would yield higher rates of data-based decision-making than Experiment 1, contrary to our findings. Along with the result that Experiment 2 has a higher percentage of subjects checking data charts in any given round than Experiment 1 (76.6% (s.e. 2.0%) vs. 65.9% (s.e. 1.9%)), the comparison of the use of data-based criteria across the experiments also suggests that the failure to apply data-based criteria is at least partly due to limited ability, not solely unwillingness. While the foregoing argument relies on a comparison of decisions across different sets of menus, the same results emerge when we focus on the two menus included in both experiments (see Appendix D.9). Experiment 3 obtains a qualitatively similar result through a treatment/control comparison.

Our estimates of data-based criteria are not driven by the explanations provided in the study interface concerning the correlational implications of causal structures. In any given round that provides access to correlational data, only a small number of subjects (16.1%) view these explanations. Moreover, if subjects checked the explanations whenever they encountered a model whose implications they did not understand, we would observe data-checking to increase in more challenging problems, contrary to our evidence (see Appendix D.7).[42]

As explained in Section 4.4, the distributions of advice-based criteria cannot be directly compared across the experiments because the distribution in Experiment 1 excludes subjects who apply the Conditional Correlations criterion, whereas Experiment 2 includes them. Ignoring these selection effects, and assuming that there is no support for the Low Promise and High Action criteria, we

---

[40]The alignment of estimates between the two experiments is not a consequence of the noise adjustment, since this adjustment does not alter the relative frequencies of the criteria.

[41]Appendix D.6 shows that estimating the mixture model excluding structure-based criteria, excluding advice-based criteria, or both leads to estimates of data-based criteria that substantially differ from both those we obtain from the full model and from those in Experiment 2.

[42]This absence of an effect may arise if subjects fail to realize when they misunderstand the correlational implications of a causal structure.

observe that low actions are more popular than high promises (column 2 of Panel B of Table 4). This qualitative feature mirrors Experiment 1, but at a greater magnitude. Accounting for noise using the corresponding estimate from Experiment 1 further magnifies the differences between the two experiments' criteria distributions.[43]

4.5 Parsimonious approximations

So far, our analysis considers a larger number of types than is practical for applications. Can more parsimonious models explain our data similarly well as our unrestricted model?

To answer this question, we consider the leave-one-out predictive power of all possible one-, two-, and three-type combinations, measured by the Euclidian distance between predicted and observed choice distributions, averaged across leave-out rounds. We benchmark the results against the full model (mean distance 0.031), which we treat as a lower bound,[44] and against uniform choice (mean distance 0.101), which we treat as an upper bound. Panel A of Table 5 lists the benchmarks. Panels B, C, and D show the three most predictive one-, two-, and three-type models, respectively, along with their predictive scores. The best one-type model covers a mere 47.1% of the distance between the two benchmarks. The best two-type model does substantively better, covering 82.9% of the distance. The best three-type model rivals the performance of the full model. It covers a whopping 95.7% of the distance. The most predictive three-type combination features precisely the three single-criterion types to which the full model assigns the largest weight when maximizing in-sample fit: High Promise, Low Action, and Conditional Correlations. This three-type model best fits the data in-sample with type frequencies of 15.5%, 30.8%, and 53.8%, respectively. While it rivals the full models' out-of-sample predictive power, the fact that its in-sample estimated noise parameter of 42.9% substantially exceeds the full model's 26.6% suggests that its parsimony comes at the cost of excluding some empirically relevant types.

4.6 Individual characteristics

Heterogeneity is a key characteristic of subjects' approaches to choice under conflicting causal interpretations. This fact raises the question of whether it is possible to predict the way in which a subject approaches these decisions. To answer this question, we first study which individual characteristics predict the choice of the correct, the low-action, and the high-promise interpretations. We then estimate an extended version of our mixture model that lets the weights of a restricted set of types depend on predictor variables.

In our reduced-form analysis, we regress binary indicators for selecting interpretations with a given property (e.g. high promise) on a vector of individual characteristics. We control for the

---

[43]Since the provision of correlational information arguably increases problem difficulty, it is plausible that subjects implement advice-based criteria in Experiment 2 with less noise than in Experiment 1.

[44]The full model does not necessarily minimize the distance between predicted and empirical choice distributions, as it may overfit.

Table 5: Best n-type models

	Out-of-sample	In-sample	
	L1O Euclidian distance	Type frequency	Noise parameter
A. All types			
Full model	0.031		0.266
Uniformly random	0.101		1.000
B. Best single types			
Combination 1	0.068		0.699
A. High Promise, Unconditional Correlations, More Links		1.000	
Combination 2	0.069		0.697
A. Unconditional Correlations, Supersets		1.000	
Combination 3	0.069		0.671
A. Unconditional Correlations, More Links		1.000	
C. Best two-type combinations			
Combination 1	0.043		0.494
A. Low Action		0.395	
B. Conditional Correlations		0.605	
Combination 2	0.053		0.488
A. High Action, Unconditional Correlations, More Links		0.385	
B. Low Action		0.615	
Combination 3	0.055		0.531
A. High Action, Unconditional Correlations, More Links		0.388	
B. Low Action, Supersets		0.612	
D. Best three-type combinations			
Combination 1	0.034		0.429
A. High Promise		0.155	
B. Low Action		0.308	
C. Conditional Correlations		0.538	
Combination 2	0.035		0.440
A. High Promise, Supersets		0.164	
B. Low Action		0.298	
C. Conditional Correlations		0.539	
Combination 3	0.036		0.443
A. High Promise, Middle Action		0.148	
B. Low Action		0.314	
C. Conditional Correlations		0.538	

correlations between the correct, high-promise, and low-action interpretations that arise across menus. For example, when the dependent variable is an indicator for choosing the correct interpretation, controls consist of indicators for whether the correct interpretation coincides with the high-promise interpretation, the low-action interpretation, or both. We construct parallel sets of control indicators when the dependent variable indicates choosing the high-promise or the low-action interpretation.

Column 1 of Table 6 shows how individual characteristics affect the propensity to choose the correct interpretation. Reassuringly, we find that higher CRT scores and greater background knowledge of statistics and causal inference both positively predict that propensity ($p < 0.01$ in both cases). So does studying a STEM field ($p < 0.01$), and—although insignificantly so—studying economics or business. The effects are sizeable, especially compared to the random choice benchmark.[45] Given our highly educated subject sample, these effects suggest that our laboratory results represent a ceiling on subjects' fit-checking abilities; a hypothesis we examine in Section 5.

---

[45]To derive the random choice benchmark, notice that the experiment uses 7 three-option rounds and 18 two-option rounds. Assuming that in each round fraction p of subjects choose the correct consistently and fraction $(1-p)$ randomize uniformly across all options, we expect fraction of subjects $\frac{7}{25}(p + (1-p)\frac{1}{3}) + \frac{18}{25}(p + (1-p))\frac{1}{2}) = 0.5467p$ to choose the correct interpretation. Therefore, an effect of Δq on the probability of choosing the target corresponds to an effect of $\Delta q/0.5467$ of choosing the target for reasons other than randomness.

Table 6: Predictors of choice: demographics and subject background

VARIABLES	(1) Correct DAG	(2) Data viewed	(3) High Promise	(4) Low Action	(5) High Promise	(6) Low Action	(7) p-value difference
	Reduced-form estimates				Structural estimates		
Constant	0.398***	0.234**	0.351***	0.419***	-1.546***	-0.684***	0.000***
	(0.049)	(0.091)	(0.040)	(0.046)	(0.058)	(0.079)	
Female	-0.050**	-0.071*	0.039***	0.015	0.843***	0.170**	0.015**
	(0.021)	(0.040)	(0.014)	(0.019)	(0.219)	(0.080)	
CRT score (0 to 1)	0.156***	0.521***	-0.093***	-0.097**	-1.496***	-1.069***	0.389
	(0.038)	(0.077)	(0.031)	(0.040)	(0.508)	(0.138)	
Knowledge index (0 to 1)	0.133***	0.141**	-0.024	-0.069**	-0.660*	-0.559***	0.834
	(0.034)	(0.062)	(0.024)	(0.031)	(0.349)	(0.190)	
Field: STEM	0.069***	0.140***	-0.011	-0.053***	-0.016	-0.361***	0.199
	(0.020)	(0.046)	(0.017)	(0.020)	(0.277)	(0.057)	
Field: Econ. or business	0.042	0.155***	-0.030	-0.042	-0.564	-0.434***	0.747
	(0.032)	(0.059)	(0.022)	(0.030)	(0.366)	(0.111)	
Pseudoscience score (0 to 1)	-0.161**	-0.325***	0.080	0.078	2.298***	0.884***	0.056*
	(0.062)	(0.113)	(0.050)	(0.057)	(0.603)	(0.248)	
Political position (0 to 1)	-0.007	0.096	0.016	-0.020	-0.026	-0.599***	0.162
	(0.031)	(0.060)	(0.023)	(0.030)	(0.340)	(0.131)	
Risk aversion perc. rank (0 to 1)	-0.001	0.105*	-0.056**	0.024	-1.796***	-0.072	0.000***
	(0.032)	(0.061)	(0.025)	(0.030)	(0.390)	(0.125)	
Observations	11965	11965	11965	11965		11725	
Subjects	479	479	479	479		469	

Notes: Columns 1 to 4 report coefficient estimates from OLS regressions. Columns 5 and 6 represent estimated odds ratio from a single GMM estimation. Knowledge index is the number of the following questions a subject can answer correctly, normalized to the unit interval: 1. Name of $P(A|B)$, 2. Complete "Correlation does not…" 3. Spell out 'DAG'. Omitted category for gender is male. Samples exclude subjects who identify as neither male nor female (2 and 6 in Experiments I and S, respectively). Omitted category for field of study is 'other.' Political position is the position of the preferred political party according to Jolly et al. (2022), with higher values indicating a more right-wing orientation. Pseudoscience scale (Torres et al., 2020) is higher the more an individual believes in pseudoscience. Coefficient estimates for the structural model show the effect of the predictor on the log-odds of being the specified type rather than the conditional tester. p-values in column 7 reflect Wald tests of the joint hypothesis that the two estimates on a given predictor equal each other (1 degree of freedom). Each regression in columns 1-4 includes 25 observations for each of the 479 subjects in Experiment 1 who have provided complete demographic characteristics, with the exception of 10 subjects in the first session who were not shown round 25. The model in columns 5-7 excludes the 10 subjects from the first session entirely. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A greater belief in pseudoscience negatively predicts choosing the correct interpretation ($p < 0.05$). Yet, contrary to common expectations, we do not find an effect of political position.[46] This result appears less surprising when considering that both sides of the political spectrum tend to doubt their political opponents' capacity to make rational inferences from observations (*cf. naïve realism*; Griffin and Ross, 1991). We also find that females choose the correct interpretation less often ($p < 0.05$, though this effect vanishes in Experiment 3) and that risk preferences do not predict this choice. Neither effect admits a straightforward interpretation.[47]

Column 2 examines a potential mechanism underlying these findings by using an indicator for whether a subject viewed a data chart in the current round as the dependent variable. The effects generally parallel those of column 1. This suggests that the effects on correct choice do not exclusively reflect differences in the ability to draw inferences from data but also differences in the willingness to check data.

The dependent variable in column 3 is an indicator for choosing the high-promise interpretation. We find that women more frequently choose it ($p < 0.01$), whereas more risk-averse individuals ($p < 0.05$) and individuals scoring higher on the CRT ($p < 0.01$) choose it less often. The latter also select the low-action interpretation less often, as column 4 demonstrates ($p < 0.05$). Topically related characteristics such as knowledge about statistical causal inference ($p < 0.05$) and studying a STEM field ($p < 0.01$) exert similar effects on the propensity to choose the low-cost action.

The foregoing regressions only utilize a subset of the moments that our mixture model employs. To estimate the effect of individual characteristics on types using the full set of moments, we estimate a version of our mixture model that employs a multinomial logit formulation of type probabilities.[48] For tractability, we use the three-type model with the best out-of-sample predictive power (Subsection 4.5), which consists of the single-criterion types Conditional Correlations, High Promise, and Low Action. We estimate the model with GMM using heteroscedasticity-robust standard errors. Columns 5 and 6 of Table 6 display the estimated effects of the individual characteristics on the High Promise and Low Action type probabilities, respectively, using the Conditional Correlations type as the benchmark category. The estimates are consistent with those from the reduced form regressions but generally provide higher statistical precision. In fact, tests of the joint hypothesis that the effect of a characteristic on both high-promise and low-action types is zero are rejected at $p < 0.01$ throughout. The effect of a belief in pseudoscience on choosing the high-promise interpretation is particularly large and

---

[46]We also do not find an effect when we replace the position on the political spectrum by its square, to capture political extremism, or by the extent of populism (quantified by Meijers and Zaslove, 2021) of the political party closest to one's views. Moreover, replacing subjects' political positions with their responses to the question "most political issues are inherently complex" (strongly disagree...strongly agree) has no effect either.

[47]Our setting incorporates different types of risk with countervailing effects. The risk of selecting a wrong interpretation can be mitigated by spending more effort checking data charts, but expending that effort is itself risky for a subject unsure about her ability to identify the correct interpretation based on the charts. Experiment 3 (Section 5) also tests the effect of ambiguity attitudes.

[48]Specifically, given reference type K, the probability of type k is given by $\log\left(\frac{t_k}{t_K}\right) = \boldsymbol{\beta}_k \boldsymbol{X}$, where $\boldsymbol{X} = [X_{i,k}]_{i,k}$ is a matrix with one column for each demographic and psychological predictor and a constant term and one row for each subject. All other aspects of the model remain unchanged from Section 4.2.

larger than that on choosing the low-action recommendation ($p < 0.1$). Only gender and risk aversion, however, differentially affect the probability of being a high-promise and low-action type at $p < 0.05$, as column 7 shows.[49]

5 General population

Experiments 1 and 2 use a sample of students in predominantly technical fields, which raises the question of whether a broader subject population would behave similarly. Experiment 3 is a simplified version of Experiment 1 that we administer to a general population sample.

Experiments 1 and 2 also raise the question of the extent to which subjects' choices reflect specific design choices, especially the highlighting of data charts that differentiate between the available interpretations and the explainer of the correlational implications of causal structures. We address this question by including these elements in the *Explanation Treatment* of Experiment 3, administered to half our subjects, and excluding them in the *Control*, administered to the other half.

Table 7 displays the menus we use in Experiment 3. To simplify the experiment, we only use three-node DAGs. To abbreviate it, we reduce the number of rounds. To effectively discern data- and advice-based criteria, we design the experiment to preclude the influence of structure-based criteria. This approach, however, limits our ability to differentiate between the Direct Link and Unconditional Testing criteria. Consequently, we identify 10 different types, each of which combines one advice-based criterion with one data-based criterion (including 'none' in both cases). In order to identify the Conditional Correlations criterion in this simplified setting, Menus 9 and 10 do not display advisers' promises and recommendations. Each subject encounters menus 1 to 8 twice, with slightly different parametrizations each time, for a total of 18 choice problems, shown in individually randomized order and preceded by 2 practice menus. We ensure that the high and low promises amount to approximately $10 and $6, respectively, in each round.

We conducted the study on December 6 and 7, 2023, on prolific.com, targeting a sample representative in gender, age, and political affiliation. A total of 789 subjects completed the study; an additional 408 subjects started the study but did not complete it. In terms of education and age, there are no statistically significant differences between subjects who completed the survey and those who did not. Our sampling restrictions further ensure that our final sample is gender-balanced even though men drop the study at a lower rate. See Appendix D.1 for details. The median subject completed the study in 50 minutes using 15 seconds per round and earned an average incentive payment of $6.60 and a completion payment of $10.[50]

---

[49] Appendix D.8 performs a parallel analysis using subjects' agreement with various statements that loosely relate to our decision criteria.

[50] This duration is substantially shorter than the corresponding figure in Experiment 1. Possible causes include population differences, differences in the complexity of the experiments, and stakes.

Table 7: Menus used in Experiment 3

Menu	1	2	3	4	5	6	7	8	9	10
DGP	A ↙ $X \longleftarrow Y$		A ↘ $X \longleftarrow Y$		A ↙ ↘ $X \quad Y$		A ↘ $X \longrightarrow Y$		A ↘ $X \longleftarrow Y$	A ↙ ↘ $X \quad Y$
Competitor	A ↘ $X \longrightarrow Y$		A ↙ $X \longrightarrow Y$		A ↙ $X \longrightarrow Y$		A ↙ $X \longleftarrow Y$		A ↙ $X \longrightarrow Y$	A ↙ $X \longrightarrow Y$
Advice withheld	No		No		No		No		Yes	Yes
Independence										
In the data	$A \perp\!\!\!\perp Y$		$A \perp\!\!\!\perp X\vert Y$		$X \perp\!\!\!\perp Y\vert A$		$A \perp\!\!\!\perp X$		$A \perp\!\!\!\perp X\vert Y$	$X \perp\!\!\!\perp Y\vert A$
Implied by comp.	$A \perp\!\!\!\perp X$		$A \perp\!\!\!\perp Y\vert X$		$A \perp\!\!\!\perp Y\vert X$		$A \perp\!\!\!\perp Y$		$A \perp\!\!\!\perp Y\vert X$	$A \perp\!\!\!\perp Y\vert X$
High-promise DAG	Comp.	DGP	Comp.	DGP	Comp.	DGP	Comp.	DGP	DGP	Comp.
Low-action DAG	DGP	DGP	Comp.	Comp.	Comp.	Comp.	Comp.	Comp.	Comp.	Comp.

Notes: Each DAG in this table implies precisely one conditional or unconditional independence relationship which we list under the heading *Independence*. Any other conditional or unconditional correlation is generically nonzero. Each menu 1-8 is presented twice, with two slightly different parametrizations.

Figure 7: Aggregate choice distribution in Experiment 3

Menu	1		2		3		4		5		6		7		8		9	10
Parameters	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b		
DGP	40	42	75	78	22	21	46	48	21	24	55	59	66	68	67	67	44	52
Competitor	60	58	25	22	78	79	54	52	79	76	45	41	34	32	33	33	56	48

Notes: Each column corresponds to a menu from Table 7. Numbers list the percentage of subjects choosing a given option. Shades of blue reflect the percentages.

Figure 7 shows the aggregate distribution of choices pooled across the Explanation and Control treatments. The fact that we observe similar distributions in each pair of equivalent menus (up to minor parametric variation) indicates that subjects paid attention to the study.

We fit our mixture model to infer the popularity of the individual choice criteria. Table 8 shows the results. Column 1 pools across the Explanation and Control treatments (results plotted in Figure 5); columns 2 and 3 show the corresponding treatment-specific estimates. Two features stand out in the pooled analysis. First, at 6.4%, the fraction of subjects who consistently identify the correct interpretation is less than a fifth of the corresponding figure in the laboratory sample. Second, at 40.9%, the prevalence of the High Promise criterion—a preference to view the world through rose-tinted glasses—is more than double the rate observed in the laboratory sample. While we observe a higher fraction of subjects using the High Action criterion (7.4%) than in the laboratory sample (2.3%), the fraction of subjects using the Unconditional Correlations or Direct Link criteria is comparable to

the laboratory sample (19.5% and 16.5%, respectively), as is the virtual absence of subjects described by the Low Promise criterion (0.2% and 2.3%, respectively).

The impact of the Explanation Treatment is minor. It decreases the popularity of the High Promise criterion by 4 percentage points ($p < 0.01$), and increases the estimated probability of random choice by 5.2 percentage points ($p < 0.01$). As column 4 shows, however, it does not statistically significantly alter the prevalence of any data-based criterion, including 'none.' The latter effect is surprising considering that the treatment nearly doubles the fraction of subjects who view a chart in any given round (from 17.8% to 33.8%; $p < 0.01$). These effects are consistent with the hypothesis that the Explanation Treatment encourages subjects to try to make more data-based decisions. Yet, the slight changes in choices it causes do not reflect an improvement.

Table 8: Distribution of criteria use in general population sample

	(1)	(2)	(3)	(4)
Sample				*p*-value treatment effect
Explanation treatment	✓		✓	
Control	✓	✓		
Data-based criteria				
None	0.741	0.766	0.745	0.069
	(0.007)	(0.009)	(0.008)	
Unconditional	0.195	0.188	0.201	0.532
	(0.013)	(0.015)	(0.014)	
Conditional	0.064	0.046	0.054	0.687
	(0.014)	(0.016)	(0.014)	
Advice-based criteria				
None	0.075	0.020	0.053	0.538
	(0.033)	(0.032)	(0.044)	
High Promise	0.409	0.448	0.407	0.000
	(0.007)	(0.007)	(0.008)	
Low Promise	0.002	0.008	0.005	0.827
	(0.007)	(0.007)	(0.008)	
High Action	0.074	0.078	0.098	0.538
	(0.020)	(0.019)	(0.027)	
Low Action	0.375	0.401	0.383	0.295
	(0.011)	(0.010)	(0.014)	
Random choice probability	0.311	0.268	0.320	0.000
	(0.007)	(0.007)	(0.008)	
Subjects	789	387	402	
Observations	14202	6966	7236	

Notes: Column 4 shows *p*-values of Wald tests that the frequency estimates in the Explanation and Control treatments equal each other.

As in Experiment 1, we find that a three-type version of our full model achieves out-of-sample predictive power (0.021) akin to that of the full model (0.020). We use a leave-out-one approach based on the Euclidian distance between predicted and actual choice distributions for all subsets of

types.[51] While the most predictive three-type subset includes the single-criteria types High Promise and Low Action, just as in Experiment 1, it is unsurprising that the Conditional Correlations criterion is not included in that subset given its low incidence in Experiment 3. Instead, the third type uses the Unconditional Correlations criterion and resolves indeterminacies by selecting the low-action interpretation. The best two-type model achieves substantially lower power (0.026), as does the best single-type model (0.040).

Finally, we use the larger variation in individual characteristics in the general population sample to examine the predictors of choice under conflicting causal interpretations. We extend the set of predictors with ambiguity attitudes measured using the method of Dimmock et al. (2015), as well as age and personal income defined as household income divided by the square root of household size (OECD, 2022). Table 9 displays the results. Consistent with our interpretation of Experiments 1 and 2 as an upper bound on inferential abilities, we find that individuals with a graduate degree significantly more often choose the correct interpretation ($p < 0.01$). We also replicate the effect of knowledge of statistical causal inference on correct choice ($p < 0.01$), though we no longer detect an effect of CRT scores. We further replicate the null effect of political position even in this politically more diverse sample, though belief in pseudoscience is no longer predictive. The effects of risk and ambiguity aversion countervail each other and are statistically weak. Age and income exert negative effects.[52] The frequency of chart-viewing in column 2 provides suggestive evidence about mechanisms. Consistent with an ability-channel, subjects with a graduate degree do not check the data more often, but still choose the correct interpretation more often. CRT scores exhibit the opposite pattern. In spite of more frequent chart-checking ($p < 0.01$), higher-scoring subjects do not make better choices. Neither do subjects with a greater belief in pseudoscience, though they check data less frequently ($p < 0.01$). Columns 3 and 4 show the effects of individual characteristics on the propensity to choose the high-promise and low-action interpretations, respectively. Mirroring results from the laboratory, subjects with a graduate degree less often choose the low-action interpretation ($p < 0.05$), and those with more knowledge about statistical causal inference less often choose the high-promise interpretation ($p < 0.01$). Higher-earning individuals more often choose either ($p < 0.01$ in both cases). There are no strong or consistent effects of risk and ambiguity attitudes.[53]

To estimate the effect of individual characteristics on the type distribution using the full set of moments, we again estimate our mixture model with a multinomial logit formulation of type probabilities, restricting attention to the most predictive three-type model. In contrast to the corresponding analysis in the laboratory sample, our benchmark type combines the Unconditional Correlation and Low Action criteria. The estimates achieve greater statistical precision than those from the reduced-form

---

[51]To achieve the globally best predictive power (0.018), the best-fitting three-type model needs to be extended to a four-type model using the High Action single-criterion type.

[52]The effect of income is consistent with the hypothesis that identifying the correct interpretation has cognitive costs and higher-earning individuals have a lower marginal utility of money.

[53]In light of existing studies on the cross-domain predictive power of ambiguity attitudes, the latter result is not entirely surprising (Trautmann and Van De Kuilen, 2015).

analysis. Column 5 shows that being more politically conservative and earning a higher income both significantly increase the likelihood of being a high-promise type ($p < 0.05$ in both cases), whereas a higher CRT score and greater knowledge about statistical causal inference decrease it ($p < 0.01$ in both cases). Column 6 shows that multiple factors raise the probability of being a low-action type (each with $p < 0.01$): being politically more conservative, being female, earning more, and being more risk-averse. Education- and knowledge-related factors (undergraduate degree, graduate degree, CRT, and knowledge score) decrease it, as, surprisingly, does greater ambiguity aversion ($p < 0.01$). In each of these cases, the two coefficients on the High Promise and Low Action types are statistically jointly significant. The difference in the effects is only statistically significant for the case of possessing a graduate degree ($p < 0.05$, column 7).

Overall, the results from the broader population in Experiment 3 differ in two key ways from those obtained from the laboratory sample in Experiments 1 and 2:[54] a significantly higher popularity of the High Promise criterion and a substantially lower incidence of the Conditional Correlations criterion. Other results emerge in both populations. There is pronounced heterogeneity in decision-making with conflicting causal interpretations, the criteria High Promise and Low Action both receive substantial support, as does one of the data-based criteria, and a three-type model suffices to achieve out-of-sample predictive power that rivals the performance of the full model, while simpler models do not.

---

[54]Because Experiments 1 and 3 also differ in terms of design and incentive amounts, we cannot exclude the possibility that factors other than subject-pool differences contribute to these effects.

Table 9: Predictors of choice: demographics and subject background

VARIABLES	(1) Correct DAG	(2) Data viewed	(3) High Promise	(4) Low Action	(5) High Promise	(6) Low Action	(7) p-value difference
	Reduced-form estimates				Structural estimates		
Constant	0.327***	0.286***	0.634***	0.454***	0.903***	0.337***	0.000***
	(0.037)	(0.076)	(0.041)	(0.044)	(0.110)	(0.129)	
Female	-0.014	-0.004	-0.020	0.018	-0.046	0.380***	0.052*
	(0.013)	(0.027)	(0.013)	(0.015)	(0.128)	(0.118)	
Age (10 year units)	-0.008	-0.020**	0.001	-0.002	0.113**	0.092*	0.792
	(0.005)	(0.010)	(0.005)	(0.005)	(0.050)	(0.051)	
Income (USD 100k units)	-0.056**	-0.065	0.023	0.070**	0.651**	1.009***	0.404
	(0.024)	(0.059)	(0.027)	(0.029)	(0.276)	(0.291)	
Undergraduate degree	0.019	0.020	-0.002	-0.014	-0.210	-0.371***	0.548
	(0.015)	(0.033)	(0.016)	(0.018)	(0.163)	(0.141)	
Graduate degree	0.049***	-0.039	0.028	-0.045**	0.262*	-0.457**	0.021**
	(0.018)	(0.039)	(0.020)	(0.021)	(0.158)	(0.189)	
CRT score (0 to 1)	0.017	0.234***	-0.028	-0.010	-0.649**	-0.578**	0.863
	(0.024)	(0.049)	(0.025)	(0.030)	(0.251)	(0.255)	
Knowledge index (0 to 1)	0.060**	0.198***	-0.086***	-0.021	-1.734***	-1.220***	0.327
	(0.026)	(0.057)	(0.029)	(0.033)	(0.468)	(0.326)	
Pseudoscience score (0 to 1)	-0.035	-0.244***	0.009	-0.050	0.350	0.319	0.958
	(0.033)	(0.072)	(0.034)	(0.039)	(0.347)	(0.303)	
Political position (0 to 1)	-0.015	-0.065*	0.016	0.017	0.539***	0.569***	0.925
	(0.018)	(0.037)	(0.021)	(0.023)	(0.203)	(0.213)	
Risk aversion perc. rank (0 to 1)	-0.038*	-0.045	-0.016	0.032	0.280	0.656***	0.292
	(0.021)	(0.043)	(0.022)	(0.025)	(0.213)	(0.203)	
Ambiguity aversion perc. rank (0 to 1)	0.037*	-0.063	-0.009	-0.049*	-0.237	-0.506**	0.476
	(0.021)	(0.043)	(0.022)	(0.027)	(0.215)	(0.218)	
Help treatment	0.013	0.135***	-0.013	-0.010	-0.129	-0.098	0.885
	(0.013)	(0.027)	(0.013)	(0.015)	(0.127)	(0.111)	
Observations	13500	13500	13500	13500	13500		
Subjects	750	750	750	750	750		

Notes: The estimates exclude 39 subjects, of which 24 report a political party preference of 'other,' 18 identify as neither male nor female, and 3 do both. The omitted category for education is having neither a graduate nor undergraduate degree. Effective income is measured in units of $100,000. Age is measured in units of 10 years. Columns 1 to 4 report coefficient estimates from OLS regressions. Columns 5 and 6 represent estimated odds ratio from a single multinomial logit regression. Knowledge index is the number of the following questions a subject can answer correctly, normalized to the unit interval: 1. Name of $P(A|B)$, 2. Complete "Correlation does not..." 3. Spell out 'DAG'. Omitted category for gender is Male. Pseudoscience scale (Torres et al., 2020) is higher the more an individual believes in pseudoscience. Coefficient estimates for the structural model show the effect of the predictor on the log-odds of being the specified type rather than the conditional tester. Asterisks in columns 1 to 4 reflect tests of the null hypothesis that the corresponding parameter value is zero. Asterisks in columns 5 and 6 reflect tests of the null hypothesis that the corresponding odds ratio is one. p-values in column 7 reflect Wald tests of the joint hypothesis that the two-parameter estimates on a given predictor equal each other (1 degree of freedom). Each regression in columns 1-4 includes 18 observations for each of the 750 general population subjects. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

6 Conclusion

In this paper, we have experimentally studied the common problem of selecting actions according to their causal impact when only correlational data and some hypotheses about underlying causal mechanisms are available. Our data feature three prominent types: those who seek to determine the objective truth (with varying levels of success), those who select the interpretation that promises the rosiest outlook, and those who minimize spending on the action. More detailed analysis reveals that subjects' abilities to determine the correct interpretation range from merely interpreting unconditional correlations to accurately identifying correct interpretations, even amid complexities such as the presence of v-colliders and the absence of data on unconditional correlations. Among subjects who use structure-based criteria, a preference for complexity is more than twice as common as a preference for simplicity. While the popularity of some criteria varies substantially across student subjects and the U.S. general population, the arguably small group of criteria garnering significant support is consistent across both samples.

Our results confirm some of the key assumptions in the Narrative Equilibria and Model Persuasion literatures, but add important qualifications, such as the prevalence of a preference for low actions that is likely driven by caution. The diversity in inferential abilities we document introduces intriguing screening challenges to the Model Persuasion literature. Beyond theoretical implications, our findings also help predict when individuals' behavioral tendencies yield the greatest losses. This typically occurs when the true causal mechanism demands substantial investments but an alternative interpretation promises better outcomes with lesser investment, especially when the discrepancies between the alternative interpretation and the data are hard to discern. The debate on anthropogenic climate change is an example.

Our current analysis, facilitated by a stylized decision environment, prompts exploration in richer settings. These include situations in which no interpretation completely aligns with the DGP, situations influenced by prior beliefs and attachments to specific advisors, and situations involving the added challenge of small-sample uncertainty. Based on results about the dependence of optimal model complexity on sample size (Montiel Olea et al., 2022) and on evidence concerning domain-specific priors over model structures (Tenenbaum et al., 2011; Johnson et al., 2019), we anticipate that these factors will particularly impact the support for structure-based criteria, as may settings that feature a need to memorize and internalize model structures. We leave these questions for further research.

References

Aina, Chiara, "Tailored Stories," *Unpublished*, 2024.

Alysandratos, Theodore, Aristotelis Boukouras, Sotiris Georganas, and Zacharias Maniadis, "The Expert and The Charlatan: an Experimental Study in Economic Advice," *Available at SSRN 3644653*, 2020.

Ambuehl, Sandro and B Douglas Bernheim, "Interpreting the will of the people: a positive analysis of ordinal preference aggregation," *NBER working paper*, 2021.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective models of the macroeconomy: Evidence from experts and representative samples," *The Review of Economic Studies*, 2022, *89* (6), 2958–2991.

_ , **Ingar Haaland, Christopher Roth, and Johannes Wohlfart**, "Narratives about the Macroeconomy," 2023.

_ , **Philipp Schirmer, and Johannes Wohlfart**, "Mental models of the stock market," 2023.

Angrisani, Marco, Anya Samek, and Ricardo Serrano-Padial, "Competing Narratives in Action: An Empirical Analysis of Model Adoption Dynamics," *Unpublished*, 2023.

Barron, Kai and Tilman Fries, "Narrative persuasion," *WZB Discussion Paper*, 2023.

Benjamin, Daniel J, "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, *2*, 69–186.

Camerer, Colin F, *Behavioral game theory: Experiments in strategic interaction*, Princeton University Press, 2003.

Camuffo, Arnaldo, Alfonso Gambardella, and Andrea Pignataro, "Theory-driven strategic management decisions," *CEPR DP 17664v2*, 2023.

Costa-Gomes, Miguel A and Vincent P Crawford, "Cognition and behavior in two-person guessing games: An experimental study," *American Economic Review*, 2006, *96* (5), 1737–1768.

Denzau, Arthur T and Douglass C North, "Shared mental models: ideologies and institutions," *Kyklos*, 1994, *47*, 3–31.

Dimmock, Stephen G, Roy Kouwenberg, Olivia S Mitchell, and Kim Peijnenburg, "Estimating ambiguity preferences and perceptions in multiple prior models: Evidence from the field," *Journal of Risk and Uncertainty*, 2015, *51*, 219–244.

Eckel, Catherine C and Philip J Grossman, "Men, women and risk aversion: Experimental evidence," *Handbook of experimental economics results*, 2008, *1*, 1061–1073.

Eliaz, Kfir and Ran Spiegler, "A model of competing narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

_ , **Simone Galperti, and Ran Spiegler**, "False Narratives and Political Mobilization," *arXiv preprint arXiv:2206.12621*, 2022.

Engelmann, Jan, Maël Lebreton, Peter Schwardmann, Joel J van der Weele, and Li-Ang Chang, "Anticipatory anxiety and wishful thinking," *American Economic Review*, forthcoming.

Epley, Nicholas and Thomas Gilovich, "The mechanics of motivated reasoning," *Journal of Economic Perspectives*, 2016, *30* (3), 133–140.

Felin, Teppo and Todd R Zenger, "The theory-based view: Economic actors as theorists," *Strategy Science*, 2017, *2* (4), 258–271.

Frechette, Guillaume, Emanuel Vespa, and Sevgi Yuksel, "Extracting Models From Data Sets: An Experiment Using Notes-to-Self," *Unpublished*, 2023.

Frederick, Shane, "Cognitive reflection and decision making," *Journal of Economic Perspectives*, 2005, *19* (4), 25–42.

Gilboa, Itzhak and David Schmeidler, "Maxmin expected utility with non-unique prior," *Journal of mathematical economics*, 1989, *18* (2), 141–153.

Griffin, Dale W and Lee Ross, "Subjective construal, social inference, and human misunderstanding," in "Advances in experimental social psychology," Vol. 24, Elsevier, 1991, pp. 319–359.

Horz, Carlo and Korhan Kocak, "How to keep citizens disengaged: Propaganda and causal misperceptions," *OSF Preprints*, 2022.

Ichihashi, Shota and Delong Meng, "The Design and Interpretation of Information," *Available at SSRN 3966003*, 2021.

Izzo, Federica, Gregory J Martin, and Steven Callander, "Ideological Competition," *American Journal of Political Science*, 2023, *67* (3), 687–700.

Jain, Atulya, "Informing agents amidst biased narratives," *Unpublished*, 2023.

Johnson, Samuel GB, JJ Valenti, and Frank C Keil, "Simplicity and complexity preferences in causal explanation: An opponent heuristic account," *Cognitive psychology*, 2019, *113*, 101222.

Jolly, Seth, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova, "Chapel Hill expert survey trend file, 1999–2019," *Electoral studies*, 2022, *75*, 102420.

Kahneman, Daniel, *Thinking, fast and slow*, Macmillan, 2011.

Kendall, Chad W and Constantin Charles, "Causal Narratives," *Unpublished*, 2022.

Klusowski, Joowon, Deborah A Small, and Joseph P Simmons, "Does choice cause an illusion of control?," *Psychological Science*, 2021, *32* (2), 159–172.

Kunda, Ziva, "The case for motivated reasoning.," *Psychological bulletin*, 1990, *108* (3), 480.

Langer, Ellen J, "The illusion of control.," *Journal of personality and social psychology*, 1975, *32* (2), 311.

Levy, Gilat, Ronny Razin, and Alwyn Young, "Misspecified politics and the recurrence of populism," *American Economic Review*, 2022, *112* (3), 928–962.

Marsh, Jessecae K, Cayse Coachys, and Samantha Kleinberg, "The compelling complexity of conspiracy theories," in "Proceedings of the Annual Meeting of the Cognitive Science Society," Vol. 44 2022.

Meijers, Maurits J and Andrej Zaslove, "Measuring populism in political parties: appraisal of a new approach," *Comparative political studies*, 2021, *54* (2), 372–407.

Molavi, Pooya, "Macroeconomics with Learning and Misspecification: A General Theory and Applications," 2019.

_ , Alireza Tahbaz-Salehi, and Andrea Vedolin, "Model Complexity, Expectations, and Asset Prices," *NBER working paper*, 2021.

Olea, José Luis Montiel, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat, "Competing models," *The Quarterly Journal of Economics*, 2022, *137* (4), 2419–2457.

Organization for Economic Co-operation and Development, "What are equivalence scales," https://web.archive.org/web/20220831233525/https://www.oecd.org/els/soc/OECD-Note-EquivalenceScales.pdf 2022.

Schumacher, Heiner and Heidi Christina Thysen, "Equilibrium contracts and boundedly rational expectations," *Theoretical Economics*, 2022, *17* (1), 371–414.

Schwartzstein, Joshua and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

_ and _ , "Shared Models in Networks, Organizations, and Groups," *Unpublished*, 2022.

Shiller, Robert J, "Narrative economics," *American Economic Review*, 2017, *107* (4), 967–1004.

Spiegler, Ran, "Bayesian networks and boundedly rational expectations," *The Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

Stefan, Simona and Daniel David, "Recent developments in the experimental investigation of the illusion of control. A meta-analytic review," *Journal of Applied Social Psychology*, 2013, *43* (2), 377–386.

Tenenbaum, Joshua B, Charles Kemp, Thomas L Griffiths, and Noah D Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, 2011, *331* (6022), 1279–1285.

Toplak, Maggie E, Richard F West, and Keith E Stanovich, "Assessing miserly information processing: An expansion of the Cognitive Reflection Test," *Thinking & Reasoning*, 2014, *20* (2), 147–168.

Torres, Marta N, Itxaso Barberia, and Javier Rodríguez-Ferreiro, "Causal illusion as a cognitive basis of pseudoscientific beliefs," *British Journal of Psychology*, 2020, *111* (4), 840–852.

Trautmann, Stefan T and Gijs Van De Kuilen, "Ambiguity attitudes," *The Wiley Blackwell handbook of judgment and decision making*, 2015, *2*, 89–116.

Waldmann, Michael, *The Oxford handbook of causal reasoning*, Oxford University Press, 2017.

ONLINE APPENDIX

Choosing Between Causal Interpretations: An Experimental Study

Sandro Ambuehl, Heidi C. Thysen

Table of Contents

A Theory

A.1 Generalization of Observation 1

In this section, we provide the formal extension of Observation 1 detailing the independence assumptions (unconditional and conditional on a single node) between any pair of nodes embedded in a DAG $G = (N, E)$.

We use the following definition:

Definition 5. *Consider a DAG $G = (N, E)$.*

(i) *A path in G is a sequence of distinct nodes $\{X_1, \ldots, X_n\}$ such that for every $k < n$, $(X_k, X_{k+1}) \in E$ or $(X_{k+1}, X_k) \in E$.*

(ii) *A node L is said to be a descendent of K if there exists a sequence of nodes $\{X_1, \ldots, X_n\}$ with $X_1 = K$ and $X_n = L$ such that $(X_k, X_{k+1}) \in E$ for every $k < n$.*

(iii) *A path p is said to be blocked by $Z \subset N$ if and only if*

    (a) *p contains a sequence $I \to K \to J$ or $I \leftarrow K \leftarrow J$ such that $K \in Z$.*

    (b) *p contains a sequence $I \to K \leftarrow J$ such that $K \notin Z$ and no descendent of K is in Z.*

    A set $Z \subset N$ is said to d-separate I and J if and only if Z blocks every path from I to J.

Observation 2. *Consider a DAG $G = (N, E)$, with $I, J, K \in N$ and $Z \subset N$.*

(i) *If $I \to J$, then generically $\mathrm{cov}(I, J) \neq 0$.*

(ii) (a) *If $Z = \emptyset$ does not d-separate I and J, then generically $\mathrm{cov}(I, J) \neq 0$.*

    (b) *If $Z = \emptyset$ d-separate I and J, then $\mathrm{cov}(I, J) = 0$.*

(iii) (a) *If $Z = \{K\}$ d-separate I and J, then $\mathrm{cov}(I, J|K) = 0$.*

    (b) *If $Z = \{K\}$ does not d-separate I and J, then generically $\mathrm{cov}(I, J|K) \neq 0$.*

This observation follows directly from Theorem 5.2.1 in Pearl (2009).

A.2 The Low Promise criterion and $\max\min$ choices

We characterize the choices of subjects who choose actions according to the $\max\min$ criterion proposed by Gilboa and Schmeidler (1989) in two-option menus, using notation and results from Appendix A.3. As Lemma 1 formalizes, such a subject always chooses the interpretation that makes the lowest promise.

Lemma 1. *Let $\mathcal{G}$ be the set of DAGs the interpretations under consideration is based on. If $|\mathcal{G}| = 2$, DAG $G^* \in \mathcal{G}$ implies the lowest promise if and only if*

$$G^* \in \arg\max_{G' \in \mathcal{G}} \min_{G \in \mathcal{G}} V_G(a^{G'}).$$

Proof. Consider two interpretations based on the DAGs G and G'. The expected payoff predicted by model G if action recommendation $a^{G'}$ is implemented given by:

$$
\begin{aligned}
V_G(a^{G'}) &= \hat{\alpha}_G + \hat{\alpha}_A^G a^{G'} - \frac{c}{2}\left(a^{G'}\right)^2 \\
&= \alpha^* + (\alpha_A^* - \hat{\alpha}_A^G)\mathbb{E}[A] + \hat{\alpha}_A^G a^{G'} - \frac{c}{2}\left(a^{G'}\right)^2 \\
&= \alpha^* + (\alpha^* - c \cdot a^G)\mathbb{E}[A] + c \cdot a^G a^{G'} - \frac{c}{2}\left(a^{G'}\right)^2,
\end{aligned}
$$

where the second equality follows from Lemma 2 (through the steps used in the proof of Lemma 3), and the third equality follows from $a^G = \frac{\hat{\alpha}_A^G}{c}$.

The expected payoff according to G when action recommendation $a^{G'}$ is implemented is higher than the expected payoff according G' when the action recommendation a^G is implemented if and only if

$$V_G(a^{G'}) - V_{G'}(a^G) = c \cdot (a^{G'} - a_A^G)\mathbb{E}[A] + \frac{c}{2}\left(\left(a^G\right)^2 - \left(a^{G'}\right)^2\right) \geq 0,$$

or equivalently,

$$c \cdot (a^G - a^{G'})\left(\frac{a^G + a^{G'}}{2} - \mathbb{E}[A]\right) \geq 0.$$

By Lemma 3, $V_G(a^G) > V_{G'}(a^{G'})$ if and only if $c \cdot (a^G - a^{G'})\left(\frac{a^G + a^{G'}}{2} - \mathbb{E}[A]\right) > 0$. Which in turn is equivalent to $V_G(a^{G'}) > V_{G'}(a^G)$. Furthermore, by definition of $a^{G'}$, and $a^{G'} \neq a^G$, we have $V_{G'}(a^{G'}) \geq V_{G'}(a^G)$. Hence, $V_{G'}(a^G) \leq \min\{V_G(a^{G'}), V_{G'}(a^{G'})\}$, this completes the proof. $\square$

A.3 Tools to construct menus that identify criteria distributions

In this section, we first find the optimal action recommendation given a model fitted to the data (section A.3.1). Second, we separate the four-node DAGs into equivalence classes, such that given any DGP, any two DAGs from the same equivalence class recommend the same action (section A.3.2). In section A.3.3 we show that any recursive system of linear Gaussian equations (with a constant term) when fit to the data correctly predicts the unconditional mean of each variable. We use this result in section A.3.4 to show that for any two interpretations with divergent action recommendations, there

is a cut-off, $\bar{A}$, such that when the mean of the action in the data is below (above) the cut-off then the interpretation with the higher (lower) action recommendation makes the higher promise.

Throughout, we will use the following definitions.

Definition 6. *Consider a DAG $G = (N, E)$.*

(i) *For $I \in N$, $G(I) = \{J \in N \mid (J, I) \in E\}$ is the set of* Parents *of node I.*

(ii) *For $I, J, K \in N$ we call the triple (I, J, K) a v-collider if $I, J \in G(K)$, but $I \notin G(J)$, and $J \notin G(I)$.*

In the model represented by G, the set of parents $G(I)$ is the set of right-hand-side variables of the regression equation for I. That is, it is the set of variables that model G posits to directly influence on I. A v-collider arises whenever two variables appear on the right-hand-side in the regression equation of a given third variable, but neither is a right-hand-side variable in the respective other variable's regression equation.

Furthermore, let $V_G(a) = \mathbb{E}_G[Y \mid A = a] - \frac{c}{2}a^2$ denote the expected payoff according to the interpretation based on G, when the action a is implemented.

A.3.1 Optimal action recommendation

Consider a linear system of equations with DAG representation G and $G(A) = \emptyset$ fitted to the data using ordinary least squares. Due to the linear structure, we can write the mean of Y conditional on A as:[55]

$$\mathbb{E}_G[Y|A] = \hat{\alpha}^G + \hat{\alpha}^G_A A,$$

When G is consistent with the DGP we let α^*, and α^*_A denote the intercept and slope coefficients.

Therefore, the optimal action recommendation associated with DAG G is the solution to the following maximization problem:

$$a^G = \arg\max_a \hat{\alpha}^G + \hat{\alpha}^G_A a - \frac{c}{2}a^2.$$

That is, the interpretation based on the DAG G recommends the action $a^G = \frac{\hat{\alpha}^G}{c}$.

A.3.2 Equivalence classes

In order to separate the four-node DAGs under consideration into 15 classes that always yield the same action recommendation, we start by noting that only variables that are posited to (directly or indirectly) influence the variable Y can affect the estimated effect of the action of the outcome.

---

[55]Notice that for every G with $G(A) = \emptyset$, we have $\mathbb{E}_G[Y|do(a)] = \mathbb{E}_G[Y|a]$, where $\mathbb{E}_G[Y|do(a)]$ gives the conditional mean of Y induced by deleting the equation for A, and substituting $A = a$ in the remaining equations.

Note that the equivalence classes we consider here are defined differently than the Markov equivalence classes commonly used in the Bayesian Networks literature.

We characterize the set of DAGs in each equivalence class and report the estimated effect of the action on the bonus for the DAGs in that class. The variance and covariance operators refer to the DGP. Recall that for any two variables I and J, the estimated slope coefficient of the regression of J on I is given by $\frac{\text{cov}(J,I)}{\text{var}(I)}$. The slope coefficient of I in a regression of a variable J on variables I and K is given by $\frac{\text{cov}(I,J)\text{var}(K)-\text{cov}(I,K)\text{cov}(K,J)}{\text{var}(I)\text{var}(K)-\text{cov}(I,K)^2}$.

Class 1 consists of all DAGs in which A does not (directly or indirectly) influence Y. Regardless of the DGP, the any system of linear regressions represented by a DAG in this group will predict that the action does not influence the outcome. Hence, for any DAG G in this class, we have $\hat{\alpha}^G = 0$.

Class 2 consists of all DAGs with $A \in G(Y)$, and there is no $I \in N$ such that (A, I, Y) is a v-collider. That is, A has a direct influence on Y, and no other variable has a direct influence on Y. While some of the system of linear regressions represented by a DAG in this class might calculate the total effect of A on Y as the sum of the direct effect of A on Y and the indirect effect of A on Y through one or more of the covariates, the total predicted effect of A on Y is the same. This follows directly from Proposition 2 in Spiegler (2020). For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A,Y)}{\text{var}(A)}.$$

Class 3 consists of all DAGs with $G(X) = \{A\}$ and $G(Y) = \{X\}$. That is, A does not have a indirect influence on Y, but a direct influence on Y through X, and Z does not (directly or indirectly) influence Y. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A,X)}{\text{var}(A)}\frac{\text{cov}(X,Y)}{\text{var}(X)}.$$

Class 4 consists of all DAGs with $G(Z) = \{A\}$ and $G(Y) = \{Z\}$. This class parallels Class 3 with the positions of X and Z switched. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A,Z)}{\text{var}(A)}\frac{\text{cov}(Z,Y)}{\text{var}(Z)}.$$

Class 5 consists of the single DAG $G : A \to X \to Z \to Y$. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,X)}{\text{var}(A)}\frac{\text{cov}(X,Z)}{\text{var}(X)}\frac{\text{cov}(Z,Y)}{\text{var}(Z)}.$$

Class 6 consists of the single DAG $G : A \to Z \to X \to Y$. It parallels Class 5 with the positions of X and Z switched. We have

$$\hat{\alpha}^G = \frac{\operatorname{cov}(A,Z)}{\operatorname{var}(A)} \frac{\operatorname{cov}(Z,X)}{\operatorname{var}(Z)} \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}.$$

Class 7 consists of all DAGs that contain the v-collider (A, X, Y) and no other v-colliders. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\operatorname{cov}(A,Y)\operatorname{var}(X) - \operatorname{cov}(A,X)\operatorname{cov}(X,Y)}{\operatorname{var}(A)\operatorname{var}(X) - \operatorname{cov}(A,X)^2}.$$

Class 8 consists of all DAGs that contain the v-collider (A, Z, Y) and no other v-colliders. It parallels Class 7 with the positions of X and Z switched. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\operatorname{cov}(A,Y)\operatorname{var}(Z) - \operatorname{cov}(A,Z)\operatorname{cov}(Z,Y)}{\operatorname{var}(A)\operatorname{var}(Z) - \operatorname{cov}(A,Z)^2}.$$

Class 9 consists of all DAGs for which $G(Y) = \{A, X, Z\}$ and $A \notin G(X), G(Z)$. For any DAG G in this class,

$$
\begin{aligned}
\hat{\alpha}^G \;=\; & \frac{\operatorname{cov}(AY)(\operatorname{var}(X)\operatorname{var}(Z) - \operatorname{cov}(X,Z)^2)}{\operatorname{var}(A)\operatorname{var}(X)\operatorname{var}(Z) + 2\operatorname{cov}(A,X)\operatorname{cov}(A,Z)\operatorname{cov}(X,Z) - \operatorname{cov}(X,Z)^2\operatorname{var}(A) - \operatorname{cov}(A,X)^2\operatorname{var}(Z) - \operatorname{cov}(A,Z)^2\operatorname{var}(X)} \\[4pt]
& - \frac{\operatorname{cov}(A,X)(\operatorname{cov}(X,Y)\operatorname{var}(Z) - \operatorname{cov}(X,Z)\operatorname{cov}(Z,Y)) + \operatorname{cov}(A,Z)(\operatorname{cov}(Z,Y)\operatorname{var}(X) - \operatorname{cov}(X,Z)\operatorname{cov}(X,Y))}{\operatorname{var}(A)\operatorname{var}(X)\operatorname{var}(Z) + 2\operatorname{cov}(A,X)\operatorname{cov}(A,Z)\operatorname{cov}(X,Z) - \operatorname{cov}(X,Z)^2\operatorname{var}(A) - \operatorname{cov}(A,X)^2\operatorname{var}(Z) - \operatorname{cov}(A,Z)^2\operatorname{var}(X)},
\end{aligned}
$$

which is the slope coefficient on A in the regression of Y on the three regressors A, X, Z.

Class 10 consists of the single DAG $G : A \to X \to Y \leftarrow Z$. We have

$$\hat{\alpha}^G = \frac{\operatorname{cov}(A,X)}{\operatorname{var}(A)} \frac{\operatorname{cov}(X,Y)\operatorname{var}(Z) - \operatorname{cov}(X,Z)\operatorname{cov}(Z,Y)}{\operatorname{var}(X)\operatorname{var}(Z) - \operatorname{cov}(X,Z)^2}.$$

Class 11 consists of the single DAG $G : A \to Z \to Y \leftarrow X$. It parallels Class 10 with the positions of X and Z switched. We have

$$\hat{\alpha}^G = \frac{\operatorname{cov}(A,Z)}{\operatorname{var}(A)} \frac{\operatorname{cov}(Z,Y)\operatorname{var}(X) - \operatorname{cov}(X,Z)\operatorname{cov}(X,Y)}{\operatorname{var}(X)\operatorname{var}(Z) - \operatorname{cov}(X,Z)^2}.$$

Class 12 consists of the single DAG $\begin{smallmatrix} A & & Z \\ \downarrow & \swarrow & \\ X & \to & Y \end{smallmatrix}$. We have

$$\hat{\alpha}^G = \frac{\operatorname{cov}(A,Z)\operatorname{var}(X) - \operatorname{cov}(A,Z)\operatorname{cov}(X,Z)}{\operatorname{var}(A)\operatorname{var}(Z) - \operatorname{cov}(A,Z)^2} \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}.$$

5

Class 13 consists of the single DAG $\begin{smallmatrix} A & \rightarrow & Z \\ & \nearrow & \downarrow \\ X & & Y \end{smallmatrix}$. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,X)\text{var}(Z) - \text{cov}(A,X)\,\text{cov}(X,Z)}{\text{var}(A)\text{var}(X) - \text{cov}(A,X)^2}\frac{\text{cov}(Z,Y)}{\text{var}(Z)}.$$

Class 14 consists of the single DAG $\begin{smallmatrix} A & & Z \\ \downarrow & \swarrow & \downarrow \\ X & \rightarrow & Y \end{smallmatrix}$. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,Z)\text{var}(X) - \text{cov}(A,Z)\,\text{cov}(X,Z)}{\text{var}(A)\text{var}(Z) - \text{cov}(A,Z)^2}\frac{\text{cov}(X,Y)\text{var}(Z) - \text{cov}(X,Z)\,\text{cov}(Z,Y)}{\text{var}(X)\text{var}(Z) - \text{cov}(X,Z)^2}.$$

Class 15 consists of the single DAG $\begin{smallmatrix} A & \rightarrow & Z \\ & \nearrow & \downarrow \\ X & \rightarrow & Y \end{smallmatrix}$. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,X)\text{var}(Z) - \text{cov}(A,X)\,\text{cov}(X,Z)}{\text{var}(A)\text{var}(X) - \text{cov}(A,X)^2}\frac{\text{cov}(Z,Y)\text{var}(X) - \text{cov}(X,Z)\,\text{cov}(X,Y)}{\text{var}(X)\text{var}(Z) - \text{cov}(X,Z)^2}.$$

This is a comprehensive list of all DAGs we consider for the experiment. We exclude the DAG $\begin{smallmatrix} A & \rightarrow & Z \\ \downarrow & & \downarrow \\ X & \rightarrow & Y \end{smallmatrix}$ because one of its characteristic independence relationships, $A \perp\!\!\!\perp Y|(X,Z)$, involves conditioning on two variables simultaneously, which is not information that we provide to subjects.

A.3.3 Unconditional means predicted by misspecified models

Next, we turn to the unconditional means predicted by potentially misspecified models. Given a system of linear equations represented by the DAG G in which A is an exogenous variable, the predicted conditional mean might deviate from the observed conditional mean. However, any variable's expected value (according to the DGP) of the unconditional mean predicted by G equals the true unconditional mean, as 2 shows. An immediate implication of this lemma is that any interpretation predicts the same mean bonus if the action equals its mean in the data. This observation proves useful below, when we compare the predicted payoffs of the recommended actions across interpretations.

Lemma 2. *Consider a system of linear equations represented by the DAG $G = (N, E)$, where $G(A) = \emptyset$. For every DGP and $I \in N$, we have*

$$\mathbb{E}[\mathbb{E}_G[I \mid A]] = \mathbb{E}[I].$$

Proof. We prove this statement by induction. To anchor the induction, consider any variable $I \in N$ for which $G(I) = \emptyset$. If $I = A$, this holds trivially. If $I \neq A$, then $\mathbb{E}_G[I \mid A] = \mathbb{E}[I]$, since G treats I and A as exogenous variables, and therefore as independent. Hence, $\mathbb{E}[\mathbb{E}_G[I \mid A]] = \mathbb{E}[I]$.

Next, consider any node J and suppose that the induction hypothesis $\mathbb{E}[\mathbb{E}_G[I \mid A]] = \mathbb{E}[I]$ holds for every $I \in G(J)$. Let $\hat{\beta}_{IJ}$ denote the slope coefficient on variable I in the OLS regression of J on

all its parents. Then, the constant term in that regression, $\hat{\beta}_J$ is given by

$$\hat{\beta}_J = \mathbb{E}[J] - \sum_{I \in G(J)} \hat{\beta}_{IJ} \mathbb{E}[I]. \tag{3}$$

Furthermore, applying the conditional expectation operator $E_G[\cdot|A]$ to the regression equation that defines J according to G yields

$$\mathbb{E}_G[J \mid A] = \hat{\beta}_J + \sum_{I \in G(J)} \hat{\beta}_{IJ} \mathbb{E}_G[I \mid A]. \tag{4}$$

Substituting 3 into equation 4, and taking the expectation over the action, we obtain:

$$\mathbb{E}[\mathbb{E}_G[J \mid A]] = \mathbb{E}[J] + \sum_{I \in G(J)} \hat{\beta}_{IJ}(\mathbb{E}[\mathbb{E}_G[I \mid A]] - \mathbb{E}[I])$$

By the induction hypothesis, the term in parentheses is zero. Hence, $\mathbb{E}[\mathbb{E}_G[J \mid A]] = \mathbb{E}[J]$ as was to be shown. $\qquad\square$

A.3.4 Pairwise comparison of promises

We next demonstrate how to select the mean of the action in the DGP to change which of two given models yields the higher promise. Lemma 3 shows that the interpretation with the lower action recommendation make the higher promise if and only if the mean action exceeds some threshold. An immediate consequence of the lemma is that in any three-option menu, the interpretation associated with the median action recommendation can never make the highest promise.

Lemma 3. *Consider two interpretations based on the models, G and G', where $G(A) = G'(A) = \emptyset$. Let a_G and $a_{G'}$ denote the corresponding action recommendations, and suppose $a_G > (<)a_{G'}$. Then $V_G(a^G) \geq V_{G'}(a^{G'})$ if and only if $\mathbb{E}[A] \leq (\geq)\frac{a^G + a^{G'}}{2}$.*

Proof. First, recall that for every DAG, G, we can write the predicted conditional mean of the bonus as a linear function of the action, a, specifically, $\mathbb{E}_G[Y \mid A = a] = \hat{\alpha}^G + \hat{\alpha}_A^G a$. By Lemma 2, we thus have

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}_G[Y \mid A]] = \hat{\alpha}^G + \hat{\alpha}_A^G \mathbb{E}[A]$$
$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid A]] = \alpha^* + \alpha_A^* \mathbb{E}[A].$$

Combining the two equations and solving for $\hat{\alpha}^G$ yields:

$$\hat{\alpha}_G = \alpha^* + (\alpha_A^* - \hat{\alpha}_A^G)\mathbb{E}[A]. \tag{5}$$

We use (5) to express the interpretation's promise as a function of $\mathbb{E}[A]$:

$$
\begin{aligned}
V_G(a^G) &= \hat{\alpha}^G + \hat{\alpha}_A^G a^G - \frac{c}{2} \left(a^G\right)^2 \\
&= \alpha^* + (\alpha_A^* - \hat{\alpha}_A^G)\mathbb{E}[A] + \hat{\alpha}_A^G a^G - \frac{c}{2}\left(a^G\right)^2 \\
&= \alpha^* + (\alpha_A^* - c \cdot a^G)\mathbb{E}[A] + \frac{c}{2}\left(a^G\right)^2,
\end{aligned}
$$

where the third equality uses $a^G = \frac{\hat{\alpha}_A^G}{c}$. We use this expression to write the difference between the promises associated with models G and G', respectively, as follows:

$$
V_G(a^G) - V_{G'}(a^{G'}) = c \cdot (a^{G'} - a^G)\mathbb{E}[A] + \frac{c}{2}\left(\left(a^G\right)^2 - \left(a^{G'}\right)^2\right) = c \cdot (a^G - a^{G'})\left(\frac{a^G + a^{G'}}{2} - \mathbb{E}[A]\right).
$$

This concludes the proof. $\qquad\square$

B Identification and estimation

B.1 Experiment 1

In this section we present the details of our GMM estimator (section B.1.1), perform Monte Carlo simulations to explore its finite-sample properties (section B.1.2), and exhibit the distances between any pair of types (section B.1.3).

B.1.1 GMM estimation

Let $\boldsymbol{t} = (t_1, \ldots, t_{n-1})$ denote a $(n-1)$-simplex which identifies the distribution over the n types. Each type i is associated with a matrix T^i with elements $(T^i_{c,m})$ that indicates the probability with which type i chooses c from menu m, where $T^i_{3,m} = 0$ for every menu with only two options. Hence, if i makes a unique choice on menu m, then $T^i_{c,m} = 1$ if c is the chosen option and 0 otherwise. If i randomizes across k options, then $T^i_{c,m} = 1/k$ for each chosen option c, and 0 for each unchosen option. We let N denote the vector consisting of elements $N_{c,m} = \frac{1}{|m|}$ where $|m|$ denotes the number of alternatives in menu m.

We use the generalized method of moments to obtain an estimate $\hat{t}$ of the type vector and an estimate $\hat{q}$ of the noise probability. We choose these to make the model match all first moments and all second moments of the choice distribution. Specifically, the estimated parameters $(\hat{t}, \hat{q})$ should match the marginal distribution of choices across all menus, i.e. the vector of observed probabilities $p_{c,m}$ with which option c is chosen on menu m in our subject sample. In addition, the estimated parameters should match the products of these probabilities, i.e. $p_{c,m} p_{c',m'}$ for all menus m, m' and for all options c, c' from the corresponding menus.

To state the optimization problem formally, and to prove identification of our model, let $\tilde{p}^i_{c,m}(q) = (1-q)T^i_{c,m} + qN_{c,m}$ denote the probability type i chooses option c and menu m given the noise probability q. Similarly, $\tilde{p}^i_{(c,m),(c',m')}(q) = (1-q)^2 T^i_{c,m} T^i_{c',m'} + q^2 N_{c,m} N_{c',m'} + q(1-q)(T^i_{c,m} N_{c',m'} + N_{c,m} T^i_{c',m'})$ denote the probability type i chooses option c from menu m *and* option c' from menu m'. Given (t, q), our model then predicts first moments $\tilde{p}_{c,m} = \sum_{i=1}^n t_i \tilde{p}^i_{c,m}$, and second moments $\tilde{p}_{(c,m)(c',m')} = \sum_{i=1}^n t_i \tilde{p}^i_{(c,m),(c',m')}$. Note that some of these choice probabilities are redundant, since (conditional) choice probabilities across all options in a menu must sum to one. Therefore, we remove the last option in all menus.

Let $\boldsymbol{M_1(q)} = [(\tilde{p}^i_{c,m_j}(q))_{(c,m_j),i}]$ (where $c = A$ whenever $|m_j| = 2$ and $c \in \{A, B\}$ otherwise) denote the matrix with n columns (each column corresponds to a type) and $\sum_j (|m_j| - 1)$ rows. The rows list the probability with which the types choose all but the last option in all menus. Similarly, let $\boldsymbol{M_2(q)} = (\tilde{p}^i_{(c,m_j)(c',m_{j'})}(q))_{((c,m_j),(c',m_{j'}),i}$ (where $c = A$ whenever $|m_j| = 2$ and $c \in \{A, B\}$ otherwise, and $m_{j'} > m_j$). The rows list the probability with which the types chooses an option in two menus, where the last option is omitted in both menus.

Letting $\boldsymbol{M}(q) = \begin{bmatrix} M_1(q) \\ M_2(q) \end{bmatrix}$, the vector of theoretically predicted moments is $\boldsymbol{M}(q) \cdot \bar{\boldsymbol{t}}$. Let $\tilde{\boldsymbol{\mathcal{E}}}$ denote the corresponding empirical moments, and $\bar{\boldsymbol{t}}$ be an n-column vector, where the first $n-1$ columns corresponds to the simplex $\boldsymbol{t}$ and the n'th column is given by $1 - \sum_{i=1}^{n-1} t_i$. Thus, $\bar{\boldsymbol{t}}$ specifies the entire distribution over the types. Our estimator is then defined as

$$(\hat{\boldsymbol{t}}, \hat{q}) = \arg\min_{\boldsymbol{t},q} \left(\boldsymbol{M}(q)\bar{\boldsymbol{t}} - \tilde{\boldsymbol{\mathcal{E}}} \right) \boldsymbol{W} \left(\boldsymbol{M}(q)\bar{\boldsymbol{t}} - \tilde{\boldsymbol{\mathcal{E}}} \right)^{\mathsf{T}} \text{ s.t. } \boldsymbol{t} \text{ is a simplex} \tag{6}$$

For the weighting matrix $\boldsymbol{W}$ we use both the identity matrix and the optimal weighting matrix derived from two-stage feasible GMM. For any given q, the objective function is a quadratic form in $\bar{\boldsymbol{t}}$. Hence, it has a unique global minimum to which any numerical estimator quickly converges even in case of a large number of types.

Identification Fixing a noise parameter q, the type frequencies are identified only if $\boldsymbol{M}(q)\bar{\boldsymbol{t}} = \boldsymbol{M}(q)\bar{\boldsymbol{t}}'$ implies $\bar{\boldsymbol{t}} = \bar{\boldsymbol{t}}'$, that is, if the nullspace of the linear map $\boldsymbol{M}(q)$ from the type space to the moment space is empty, a condition that is easy to check. We constructed the set of menus in the Main treatment (Table 2) based on this insight. To determine identification with endogenous q, we rely on the intuition that any distribution of moments obtained from noise parameter q and type distribution $\bar{\boldsymbol{t}}$ can be replicated by setting $q = 0$ but including a type that uniformly randomizes on each menu. We then check that the nullspace of $M(0)$ in the redefined problem is empty. Numerically, we start the estimation procedure on a grid of initial values for q that spans the unit interval and check the local identification condition at the resulting estimates.

B.1.2 Monte Carlo Simulation

It is well known that GMM is consistent but potentially biased in finite samples. To check for such bias in our specific setting, we conduct Monte Carlo studies. We consider two data-generating processes, one whose parameters equal the estimates in column 1 of Table 4, the other consisting of pure noise. In each cases, we generate 1000 samples of 475 individuals each.

Figure B.1 plots the estimated criterion frequencies averaged across the simulations. For our first data-generating process, the estimates closely match the assumed true distribution of criteria use. While the procedure overestimates the fraction of subjects using none of the advice-based or structure-based criteria, these discrepancies are within a few percentage points. Noise is slightly underestimated, which is an expected and inevitable consequence of in-sample fitting procedures (purely noisy choice will coincide with some type in some cases by pure chance). Overall, the results suggest that the finite-sample bias of our estimator is small.

10

Figure B.1: Distribution of decision criteria in Monte Carlo Simulations

A. Data-based criteria

B. Advice-based criteria

C. Structure-based criteria

For our second data-generating process (pure noise), we find, reassuringly, that the three criteria that are most prominent in our empirical data (Conditional Correlation, High Promise, Low Action) do not receive substantial weight. Our result on preferences for simpler versus more complex model is also unlikely an artifact of the small-sample properties of our estimator. This is evidenced by the fact that our classification of pure noise does not show a bias towards or against a preference for more complex interpretation. While we do observe the unavoidable underestimation of noise, the estimated parameter of 85.4% is still very high.

Overall, our results are unlikely due to small-sample bias or misclassification of noise.

B.1.3   Distance between pairs of types

Here, we show the distance between any pair of types on the choice sets of Experiment 1. We measure the distance between types t and t' as $d(t, t') = (\boldsymbol{M}(0)\boldsymbol{I}_t)\,(\boldsymbol{M}(0)\boldsymbol{I}_{t'})'$ where $\boldsymbol{M}(0)$, as defined in section B.1.1, is the matrix of theoretically predicted moments when the probability of noisy choices equals $q = 0$ and $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t'}$ are column vectors that have entries one in positions t and t', respectively, and zero everywhere else.

Figure B.2 places types in lexicographic order along the axes. Figure B.3 displays the same information with types placed in an order that places types with smaller distances in closer proximity.

B.2   Experiment 2: Inferring type frequencies from choice frequencies

We derive type frequencies from observable choice frequencies.

To estimate the distribution of data-based criteria, we let $\boldsymbol{t} = [t_C, t_{NV}, t_U, t_D]'$, $\boldsymbol{p} = [p_C, p_{NV}, p_U, p_D]$, $\boldsymbol{I}_1 = [1, 1, 1, 1]'$, and

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Thus, if subjects choose without random errors (i.e., $q = 0$), then for the sample of subjects who use some data-based criterion, we have $\boldsymbol{p} = \boldsymbol{A}\boldsymbol{t}$. Hence, the fraction of subjects unable to identify the correct interpretation in menus C, NV, U, and D, respectively, is $\boldsymbol{I}_1 - \boldsymbol{A}\boldsymbol{t}$. By assumption, subjects unable to identify the correct interpretation in a given menu uniformly randomize across the available interpretations. Thus, we have

$$\boldsymbol{p} = (1 - q)\left(\boldsymbol{A}\boldsymbol{t} + \frac{1}{2}(\boldsymbol{I}_1 - \boldsymbol{A}\boldsymbol{t})\right) + q\frac{1}{2}\boldsymbol{I}_1$$

12

Figure B.2: Distances between types, ordered lexicographically
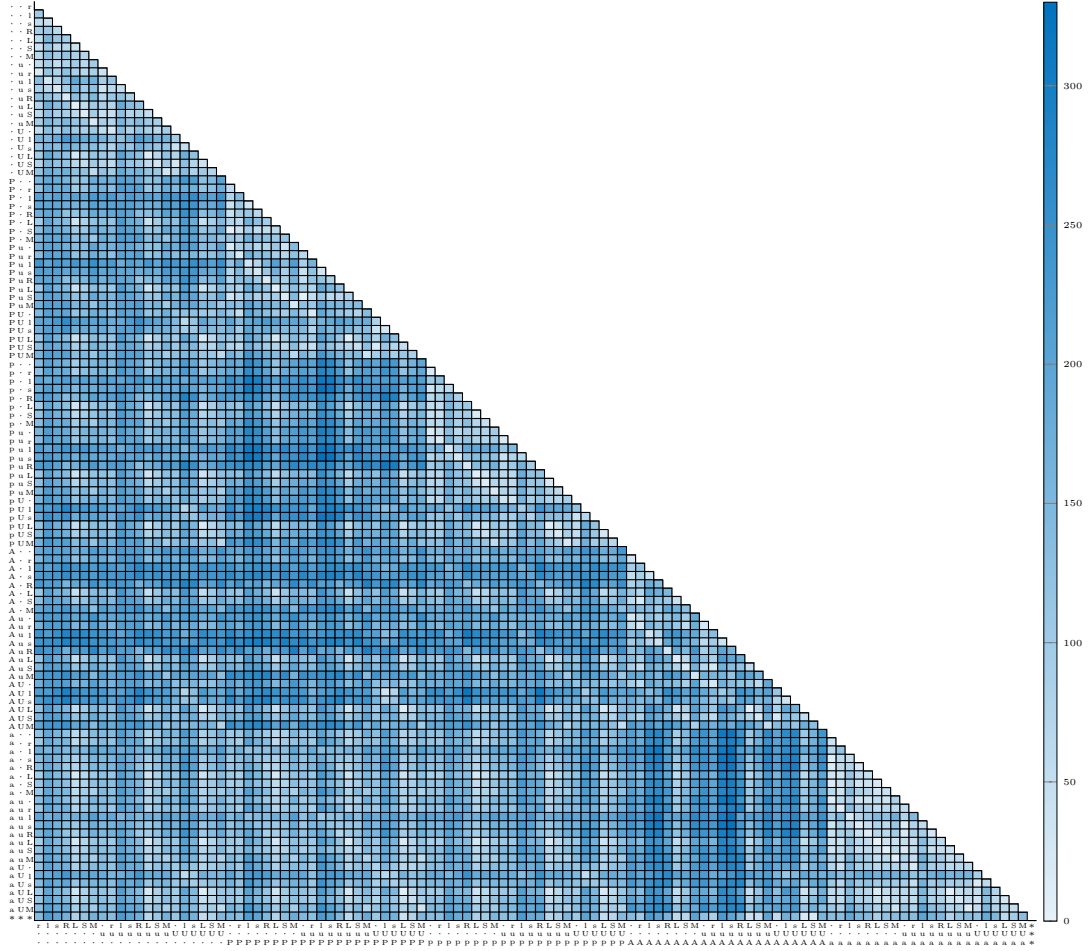
Notes: The color of a cell indicates the distance between the two types defining that cell. Legend given in the bar on the right. Each type is listed as a triple of three criteria (advice-based, data-based, structure-based). In each class, a period (.) stands for 'none.' ∗∗∗ indicates the correct interpretation, which a subject performing conditional tests always chooses. The remaining criteria are encoded as follows. *Advice-based*: *P*: High Promise, *p*: Low Promise, *A*: High Action, *a*: Low Action. *Data-based*: *u*: Direct Links, *U*: Unconditional (All), ∗: Conditional. *Structure-based*: *r*: Fewer Roots, *l*: Fewer Links, *s*: Subsets, *R*: More Roots, *L*: More Links, *S*: Supersets, *M*: Middle Action.

Figure B.3: Distances between types, ordered by closeness
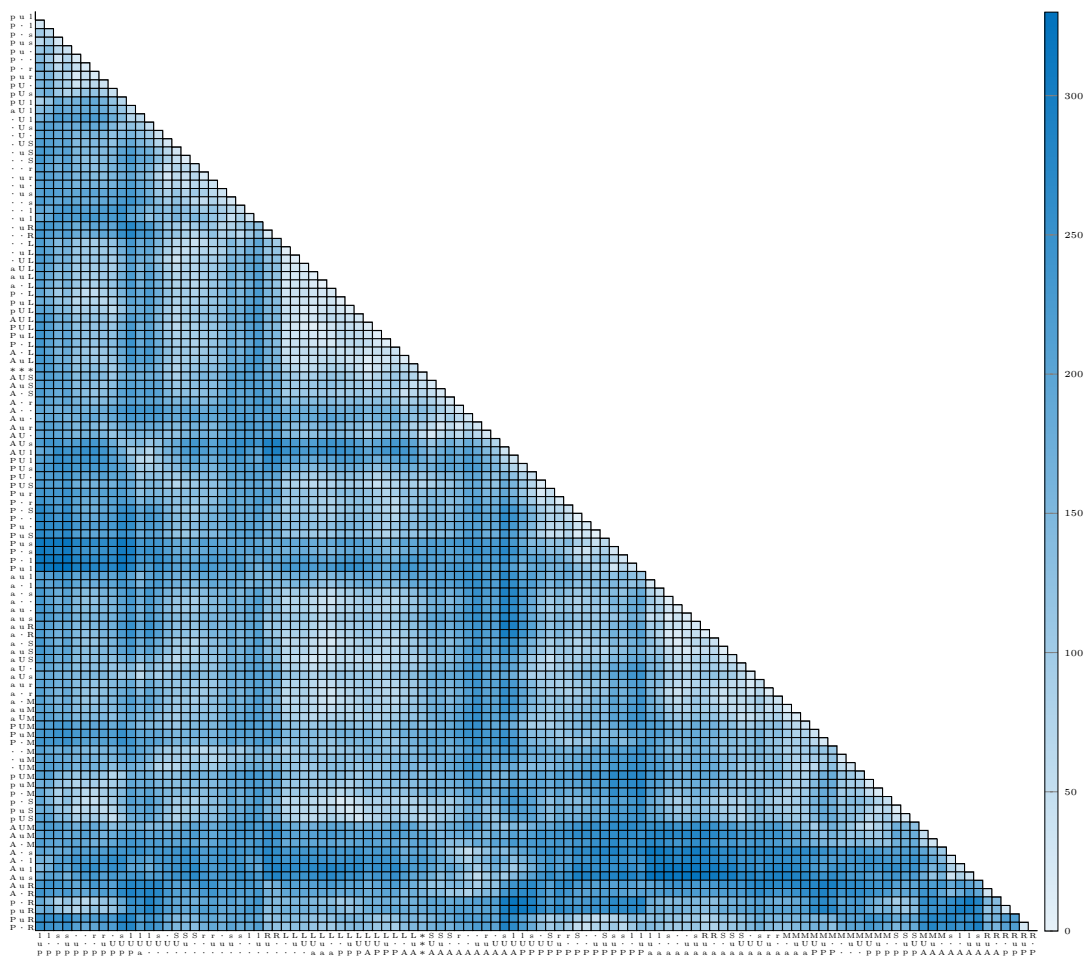
Notes: The color of a cell indicates the distance between the two types defining that cell. Legend given in the bar on the right. Each type is listed as a triple of three criteria (advice-based, data-based, structure-based). In each class, a period (.) stands for 'none.' $*\,*\,*$ indicates the correct interpretation, which a subject performing conditional tests always chooses. The remaining criteria are encoded as follows. *Advice-based*: P: High Promise, p: Low Promise, A: High Action, a: Low Action. *Data-based*: u: Direkt Links, U: Unconditional (All), $*$: Conditional. *Structure-based*: r: Fewer Roots, l: Fewer Links, s: Subsets, R: More Roots, L: More Links, S: Supersets, M: Middle Action.

Rearranging yields

$$\frac{2}{1-q} \boldsymbol{A}^{-1} \left(\boldsymbol{p} - \frac{1}{2} \boldsymbol{I}_1 \right) = \boldsymbol{t} \qquad (7)$$

which is equation (1).

To estimate the distribution of advice-based criteria, we let $\boldsymbol{s} = [s_{HP}, s_{LP}, s_{HA}, s_{LA}]'$ denote the fraction of subjects following the High Promise, Low Promise, High Action, and Low Action criteria, respectively, and we let $\Delta_P = s_{HP} - s_{LP}$, $\Delta_A = s_{LA} - s_{HA}$, $\boldsymbol{\Delta} = [\Delta_P, \Delta_A]'$, $\boldsymbol{p} = [p_a, p_b]'$, $\boldsymbol{I}_1 = [1, 1]'$, and

$$\boldsymbol{B} = \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}$$

Recall that the High Promise criterion chooses the correct interpretation in menus a, but not in menus b, and that the Low Action criterion never chooses the correct interpretation. Therefore,

$$
\begin{aligned}
\boldsymbol{p} &= (1-q) \left(\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \boldsymbol{s} + \frac{1}{2} \left(\boldsymbol{I}_1 - \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \boldsymbol{s} \right) \right) + q\frac{1}{2} \boldsymbol{I}_1 \\
&= \frac{(1-q)}{2} \begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 \end{bmatrix} \boldsymbol{s} + \frac{1}{2} \boldsymbol{I}_1 \\
&= \frac{(1-q)}{2} \boldsymbol{B} \boldsymbol{\Delta} + \frac{1}{2} \boldsymbol{I}_1
\end{aligned}
$$

Therefore, using $\boldsymbol{B}^{-1} = \frac{1}{2} \boldsymbol{B}$,

$$\boldsymbol{\Delta} = \frac{1}{1-q} \boldsymbol{B} \left(\boldsymbol{p} - \frac{1}{2} \boldsymbol{I}_1 \right),$$

which is equation (2).

C Experimental design

C.1 Experiments 1 and 2

Table C.1 shows an overview of the structure of the study, which was coded in Qualtrics and javascript. Here, we list details about each of the stages.

Table C.1: Experiment structure

---

1. Instructions and comprehension check

2. Choice between causal interpretations

    (a) Preliminary rounds

    (b) Main decisions

    (c) Framing manipulation

3. Additional decisions

    (a) Risk preference elicitation

    (b) Self-classification of decision style

    (c) Cognitive Response Test

    (d) Pseudoscience scale

    (e) Explanation of own decision-making

    (f) Educational background and demographics

Notes: The experiment proceeds in the order listed. Within each part of section 2, the order of rounds is randomized at the individual level. The two rounds of risk elicitation are also shown in individually randomized order.

Instructions and comprehension check We display all instructions on screen. The entire experiment is in English. A good command of English is a curricular requirement for all students in our subject pool. The instructions include a part that sequentially highlights each of the interactive display elements and requires subjects to interact with them.

Our statistical inference relies on the assumption that subjects have understood specific aspects of the decision environment, in particular: (i) what part of an advisor's recommendation responds to data and what part does not, (ii) how, qualitatively, DGP and action co-determine the payment, (iii) that each round is independent of all other rounds. We repeatedly emphasize these points in the instructions. Moreover, each subject can continue with the experiment only once they pass a comprehension check that presents them with ten statements that are either true or false. The statements refer to the three crucial points above. Subjects need to simultaneously label each statement as true or false and can continue only once no error is left. In case of an error, subjects only learn that one of the statements was labeled incorrectly, but not which one. Because there are 1024 possible ways of

labeling the statements, it is unlikely that a subject passes by chance. A subject who is unable to pass can raise their hand. The experimenter first reminds them to check the summary of the instructions on the previous page. If the subject is still unable to continue, the experimenter checks which of the statements was labeled incorrectly and explains the point to the subject.

Preliminary rounds   Table C.2 shows the menus used for the preliminary rounds. From the subjects' point of view, preliminary rounds are indistinguishable from other rounds of the study.

<div align="center">

Table C.2: Preliminary rounds

</div>

	Experiment 1			Experiment 2	
Menu	DGP	Competitor	Menu	DGP	Competitor
P1	$A \quad Z$ $\searrow \uparrow$ $X \to Y$	$A \quad X$ $\searrow \uparrow$ $Z \to Y$	P1	$A \quad Z$ $\searrow \uparrow$ $X \to Y$	$A \quad X$ $\searrow \uparrow$ $Z \to Y$
Promise	17.50	22.50	Promise	17.50	22.50
Action	8.00	0.07	Action	12.48	3.45
P2	$X \to Z$ $\searrow \uparrow$ $A \to Y$	$Z \to X$ $\searrow \uparrow$ $A \to Y$	P2	$Z \to X$ $\searrow \uparrow$ $A \to Y$	$X \to Z$ $\searrow \uparrow$ $A \to Y$
Promise	22.50	17.50	Promise	17.50	22.50
Action	8.82	1.13	Action	0.10	10.13

Main rounds   We graphically position the nodes of each DAG in a diamond shape. We assign nodes to positions with the following objectives, in the given priority order: (i) Arrows never cross, (ii) Whenever a menu contains two DAGs such that one can be obtained from the other by exchanging the covariates, the display of these two DAGs only differs by switching the symbols representing the covariates, (iii) If a DAG can be obtained from another DAG by deleting a link, then the nodes are in the same position in the two DAGs, and (iv) the symbol representing the action is on top and the symbol representing the bonus is at the bottom. If condition (iv) is inconsistent with the previous conditions, we place the symbol representing the action on the left.

We randomize the following display elements: (i) The order in which the list of charts is displayed. For each individual, we randomly select an order of charts displaying unconditional relations that we keep constant for the entire experiment.[56] For each unconditional relation, there is a pair of conditional relations that condition on either of the remaining variables, respectively. For each subject, we display these pairs of conditional relations in the same order as the unconditional variables. (ii) The position of the interpretations in a menu in each round for each individual. (iii) The position of theory and recommendations within the advisor speech bubbles. This is kept constant for a given subject. For

---

[56] All subjects in the first sessions saw the same order of charts.

half of the subjects, the theory is on top, for the other half the recommendations are on top. (iv) We randomly redraw the colors of the advisors in each round for each individual.

Table C.3 displays the parameters used in each round.

Framing manipulation   We test our assumption that using real-world contexts instead of our abstract frame would induce plausibility considerations that hinder the identification of the decision criteria we seek to study. Subsection D.5 describes details and shows evidence of such framing effects.

Risk preference elicitation   There are two rounds. In the first round, subjects chose one of the six 50/50 lotteries with payments (14, 14), (18, 12), (22, 10), (26, 8), (30, 6), (35, 1), where all amounts are in Swiss Francs. In the second round, payments are (21, 21), (27, 18), (33, 15), (39, 12), (45, 9), (52.5, 1.5). We randomize whether the lotteries are ordered safe to risky or risky to safe.

Self-classification of decision style   Subjects indicate their agreement with each of several statements that describe how they approach real-world decisions in the case of conflicting causal interpretations. Appendix Section D.8 presents details and results.

Demographics and other characteristics   In addition to the characteristics listed in Section 3.3, we also elicit the following characteristics: (i) native language, (ii) monthly spending, (iii) religiosity, (iv) eligibility to vote in political elections in Switzerland, (v) extent of agreement with the position of the political party they are closest to, (vi) degree level to which the subject is working towards, (vii) final grades in the university admission exam in mathematics and in their main language. Subjects also indicate their agreement with each of the following statements about political issues (such as immigration, unemployment, income inequality, social insurance, healthcare, etc.): "Most political issues are simple in principle. They have straightforward solutions." and "Most political issues are inherently complex. They do not have straightforward solutions."

C.2   Experiment 3

In addition to the differences outlined in Section 5, Experiment 3 differs from Experiment 1 in the following ways: (i) We elicit a subset of demographic information at the beginning of the study in order to understand attrition. (ii) We do not ask subjects to self-classify their decision style. We do not include rounds with real-world or verbal framing. (iii) We elicit risk preferences by letting subjects make one choice from the 50/50 lotteries $(7, 7)$, $(9, 6)$, $(11, 5)$, $(13, 4)$, $(15, 3)$, $(17.5, 0.5)$ and another choice from the 50/50 lotteries $(10.5, 10.5)$, $(13.5, 9)$, $(16.5, 7.5)$, $(19.5, 6)$, $(22.5, 4.5)$, $(16.25, 0.75)$ (shown in this order; all amounts in USD).

In the Control treatment, in which we do not provide access to explanations about the correlational implications of causal structures, we still provide descriptions of the archetypical causal structures.

Table C.3: DGP parameters

Round	β_A	β_X	β_Y	β_Z	β_{AX}	β_{AY}	β_{AZ}	β_{XY}	β_{XZ}	β_{YZ}	σ_A	σ_X	σ_Y	σ_Z
A. Experiment 1														
1	6.20	3.75	14.02	2.00	-0.90	0.90	0.00	0.90	0.00	1.00	3.09	5.50	0.50	5.50
2	35.94	3.00	5.00	3.00	0.50	1.00	0.00	2.00	0.00	3.00	2.50	4.50	0.50	2.50
3	6.22	5.00	5.00	10.00	1.00	0.00	1.00	1.00	0.00	0.00	2.75	1.00	0.50	2.00
4	35.13	5.00	0.00	10.00	1.00	0.00	1.00	1.00	0.00	0.00	2.75	1.00	0.50	2.00
5	29.62	3.00	0.53	1.00	0.67	0.00	0.60	1.50	0.60	0.00	5.50	1.50	0.50	5.50
6	6.59	5.00	5.00	7.00	1.00	0.00	0.00	1.00	0.00	0.90	2.75	1.00	0.50	2.00
7	34.44	5.00	0.00	7.00	1.00	0.00	0.00	1.00	0.00	0.90	2.75	1.00	0.50	2.00
8	4.30	10.00	4.47	0.00	0.00	0.80	-0.70	1.00	0.00	4.00	2.15	2.50	0.50	2.50
9	21.12	3.00	6.50	3.00	0.00	0.80	-0.70	1.00	0.00	4.00	2.15	2.50	0.50	2.50
10	4.20	3.20	5.75	10.00	0.00	0.55	-1.00	2.50	0.00	1.05	2.10	2.10	0.50	2.10
11	32.36	4.00	8.72	20.00	0.00	0.55	-1.00	2.50	0.00	1.05	2.50	2.50	0.50	2.50
12	8.75	3.00	7.00	3.00	0.00	1.00	-0.75	1.00	0.00	2.00	2.50	1.50	0.50	0.50
13	28.75	3.00	2.00	3.00	0.00	1.00	-0.75	1.00	0.00	2.00	2.50	0.50	0.50	0.50
14	28.75	3.00	16.38	33.00	0.00	0.50	-1.50	1.00	0.00	1.00	2.50	2.50	0.50	0.50
15	4.70	3.00	17.72	3.00	6.00	0.00	0.00	6.00	1.00	0.00	2.35	0.50	5.00	2.50
16	4.60	3.00	9.24	5.00	1.30	-0.80	0.00	1.10	0.00	1.00	2.30	2.50	0.50	2.50
17	35.00	3.00	14.24	5.00	1.30	-0.80	0.00	1.10	0.00	1.00	2.30	2.50	0.50	2.50
18	41.15	0.00	19.38	5.00	1.50	-1.00	0.00	1.00	0.00	2.00	5.50	4.00	0.50	4.50
19	5.00	3.00	9.80	3.00	0.50	0.70	0.00	0.60	0.00	1.50	2.50	2.50	0.50	2.50
20	20.00	4.00	2.00	4.00	0.85	0.55	0.00	0.70	0.00	1.00	2.00	3.50	3.00	2.50
21	27.00	4.00	14.00	7.00	0.50	1.00	0.00	-0.50	0.00	0.50	5.50	5.50	5.50	5.50
22	25.00	2.00	0.00	2.00	0.50	0.60	0.00	0.70	0.90	0.00	3.50	1.50	1.50	2.50
23	14.00	3.00	11.00	4.00	1.00	-0.50	0.00	1.00	1.00	0.00	2.50	4.50	0.50	2.50
24	13.00	6.00	6.00	7.00	1.00	0.00	0.00	0.90	0.00	0.50	2.00	0.50	1.90	1.00
25	30.00	6.00	0.00	7.00	1.00	0.00	0.00	0.90	0.00	0.50	2.00	0.50	0.50	0.50
B. Experiment 2														
D1	21.12	3.00	6.50	3.00	0.00	0.80	-0.70	1.00	0.00	4.00	2.15	2.50	0.50	2.50
D2	21.12	3.00	6.50	3.00	-0.70	0.80	0.00	4.00	0.00	1.00	2.15	2.50	0.50	2.50
D3	4.30	3.00	5.48	3.00	0.91	0.00	0.00	0.91	1.00	1.00	2.15	1.00	0.50	2.50
U1	5.00	2.00	11.68	2.00	0.00	0.84	0.00	1.00	1.00	1.50	2.50	2.50	1.50	0.50
U2	22.00	2.00	6.68	2.00	0.00	0.84	0.00	2.00	1.00	1.00	2.50	0.50	1.50	2.50
NV1	35.15	5.00	0.00	10.00	1.00	0.00	1.00	1.00	0.00	0.00	2.75	1.00	0.50	2.00
NV2	6.59	5.00	5.00	7.00	1.00	0.00	0.00	1.00	0.00	0.90	2.75	1.00	0.50	2.00
C1	5.00	2.00	11.47	5.00	0.00	0.85	0.00	1.00	0.00	0.60	2.50	2.50	2.50	0.50
C2	17.40	5.00	6.47	2.00	0.00	0.85	0.00	0.60	0.00	1.00	2.50	0.50	2.50	2.50
A1a	4.72	3.00	8.49	2.00	0.85	0.00	0.00	1.00	0.00	1.00	2.36	0.50	0.50	2.50
A1b	16.50	3.00	3.47	2.00	0.85	0.00	0.00	1.00	0.00	1.00	2.35	0.50	0.50	2.50
A2a	29.18	3.00	2.00	3.00	1.00	0.00	1.00	1.00	1.00	0.00	2.50	1.50	0.50	5.50
A2b	8.65	3.00	7.00	3.00	1.00	0.00	1.00	1.00	1.00	0.00	2.50	1.50	0.50	5.50
A3a	29.32	3.00	2.00	3.00	1.00	0.00	0.00	1.00	1.00	1.00	2.50	0.50	0.50	5.00
A3b	8.61	3.00	7.00	3.00	1.00	0.00	0.00	1.00	1.00	1.00	2.50	0.50	0.50	5.00

Notes: Each DAG can uniquely be represented as a system of linear Gaussian equations. For any variable i, β_i denotes the constant term in the equation corresponding to endogenous variable i, and σ_i is the standard deviation of the corresponding error term. For any endogenous variable i that depends on some other variable j, β_{ij} is the slope coefficient on variable j in the equation corresponding to endogenous variable i.

Table C.4: DGP parameters in Experiment 3

Round	β_A	β_X	β_Y	β_{AX}	β_{AY}	β_{XY}	σ_A	σ_X	σ_Y
Experiment 3									
1a	1.33	5.00	6.10	-0.83	0.00	0.70	2.80	1.00	1.50
1b	14.14	2.00	10.00	-0.63	0.00	1.00	2.80	0.50	1.50
2a	25.25	2.00	-0.30	0.00	0.71	0.60	2.00	1.00	1.00
2b	1.75	2.00	3.70	0.00	0.71	0.60	2.00	1.00	1.00
3a	27.93	5.00	-1.03	0.60	0.75	0.00	2.00	1.00	1.00
3b	1.90	5.00	2.97	0.60	0.75	0.00	2.00	1.00	1.00
4a	14.17	5.00	1.50	0.00	0.60	0.90	2.00	1.00	1.00
4b	1.97	5.00	4.72	0.00	0.65	1.00	2.00	1.00	1.00
5a	1.45	5.00	6.00	0.85	0.00	-0.80	2.80	1.00	1.50
5b	14.21	5.00	10.00	0.85	0.00	-0.80	2.80	1.00	1.50
6a	27.46	2.00	-1.03	0.00	0.75	0.65	2.00	1.00	1.00
6b	2.15	2.00	2.97	0.00	0.75	0.65	2.00	1.00	1.00
7a	22.55	35.00	-0.12	-0.43	0.70	0.00	2.00	0.99	1.00
7b	2.49	5.00	3.88	-0.43	0.70	0.00	2.00	0.99	1.00
8a	14.23	5.00	1.04	0.00	0.63	1.00	2.00	1.00	1.00
8b	1.53	5.00	5.04	0.00	0.63	1.00	2.00	1.00	1.00
9	3.68	2.00	2.40	0.00	0.78	0.60	2.00	1.00	1.00
10	22.55	35.00	-0.12	0.43	0.70	0.00	2.00	0.99	1.00

Notes: Each DAG can uniquely be represented as a system of linear Gaussian equations. For any variable i, β_i denotes the constant term in the equation corresponding to endogenous variable i, and σ_i is the standard deviation of the corresponding error term. For any endogenous variable i that depends on some other variable j, β_{ij} is the slope coefficient on variable j in the equation corresponding to endogenous variable i.

Appendix E.2 shows screenshots. In each prospect used for ambiguity elicitation, subjects can win \$7 or \$0. Table C.4 shows the parameters of the DGPs used in Experiment 3.

D Analysis

D.1 Summary statistics

Table D.5 shows summary statistics of our samples across the three experiments.

Experiments 1 and 2   Our sample skews female and to the political left, as is common for student subject pools. A slight majority of our subjects are from German-speaking Europe, and approximately 20% each are from the remaining Europe and from Asia. Mean monthly spending is Fr. 1208 and Fr. 1348 in Experiments 1 and 2, respectively.

Experiment 3   We elicited subjects' race, age, gender, and education at the beginning of the survey to test whether individuals who complete the survey differ systematically from those who give up at the comprehension check stage. We targeted a gender-balanced sample of 800 U.S. subjects.[57] 1197 individuals started the survey and continued until the attention check. 396 subjects did not pass the comprehension check (and did not take up our offer to help with it by email). 12 did not finish the study in spite of passing the comprehension check.

---

[57] On December 5, we ran a pilot study with 50 participants that did not include menus 17 and 18 with the purpose of checking that the Prolific subjects understood the instructions. We preregistered the study on https://www.socialscienceregistry.org/trials/12652.

Table D.5: Subject characteristics

	(1)	(2)	(3)	(4)	(5)
	Exp. 1	Exp. 2		Exp. 3	
			Completed study	Dropped at attention check	p-value attrition
Demographics					
Male	0.436	0.509	0.511	0.400	0.000
Female	0.564	0.491	0.489	0.600	0.000
Age	23.335	23.503	41.493	41.859	0.657
Origin					
German-speaking Europe	0.535	0.517	-	-	-
Other Europe	0.181	0.183	-	-	-
Asia	0.187	0.215	-	-	-
Other	0.097	0.086	-	-	-
Income			-	-	-
Monthly spending	1207.967	1347.790	-	-	-
Education					
Highschool	-	-	0.113	0.157	0.030
BA	0.436	0.406	0.636	0.652	0.591
MA	0.535	0.519	0.215	0.145	0.003
PhD and MD/JD	0.029	0.076	0.035	0.047	0.350
Field of study					
STEM	0.632	0.699	-	-	-
Economics or business	0.135	0.126	-	-	-
Other field	0.233	0.176	-	-	-
Knowledge of statistics and inference					
Can name $P(A|B)$	0.246	0.261	0.067	-	-
Can complete "Correlation does not..."	0.563	0.527	0.602	-	-
Can spell out DAG	0.087	0.097	0.071	-	-
Taken class on statistical causal inference	0.186	0.183	0.105	-	-
Psychological measures					
CRT score (0 to 7)	4.939	5.166	4.630	-	-
Pseudoscience score (20 to 100)	59.772	60.458	56.663	-	-
Religiosity (1 to 5)	1.785	1.708	2.414	-	-
Political party preference					
SVP	0.050	0.047	-	-	-
FDP	0.138	0.154	-	-	-
BDP	0.074	0.043	-	-	-
CVP	0.077	0.064	-	-	-
GPL	0.258	0.273	-	-	-
SP	0.222	0.226	-	-	-
Green	0.143	0.151	-	-	-
PdA	0.037	0.043	-	-	-
Republican	-	-	0.185	-	-
Independent	-	-	0.352	-	-
Democrat	-	-	0.432	-	-
Other party	-	-	0.000	-	-
Subjects	485	279	789	408	

Notes: For Experiments 1 and 2, educational categories refer to the program in which subjects are enrolled. For Experiment 3, educational categories reflect subjects' highest educational attainment. Parties listed in order of overall stance on the political spectrum, beginning with most conservative. CVP is the center party.

D.2 Order effects

Panel A of Figure D.4 plots the median response time against the position at which the subject made the corresponding decision in Experiment 1. Subjects take substantially longer or the first decision, presumably to familiarize themselves with the interface and to draw inferences for the first time that can be recalled in a lesser amount of time later. While response times decline across the entire experiment, this decline appears to reflect learning rather than decreased attention, as Panel B shows. For each decision position, it plots the fraction of subjects in Experiment 1 who viewed at least one data chart. Approximately two-thirds of all subjects do so throughout the main rounds, slightly fewer than in the preliminary rounds.

Figure D.4: Order effects

A. Response time
B. Data viewing
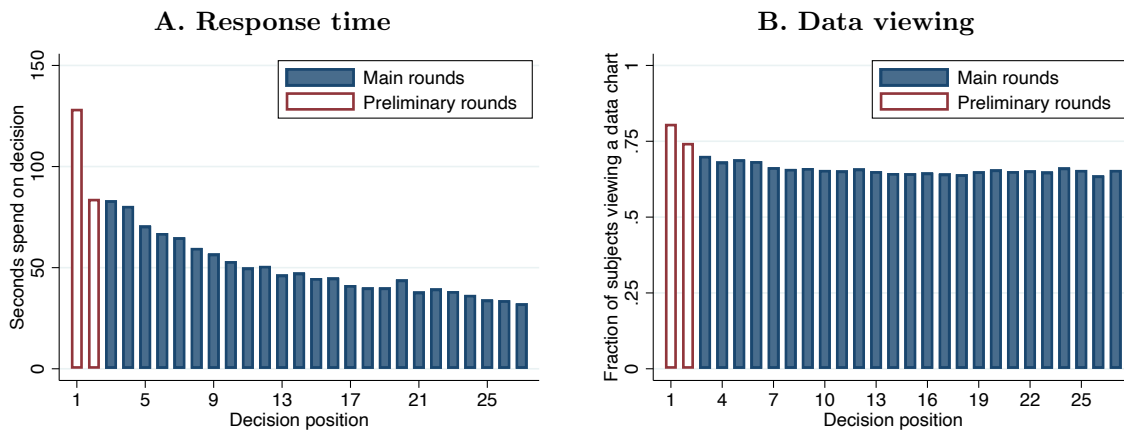
Notes: Data from Experiment 1 (subjects do not have access to the data charts in some rounds of Experiment 2).

D.3 Best-fitting types

Table D.6 lists the estimated frequencies of all 39 types that receive at least 0.1% weight, along with heteroscedasticity-robust standard errors.

D.4 Robustness to subjects affected by contingent reasoning failure

Table D.7 replicates Table 4 excluding subjects classified as falling prey to the failure of contingent reasoning based on their open responses, as discussed in Section 4.2.

Table D.6: Most common types

	Criterion																	Frequency	s.e.
	Advice-based					Data-based				Structure-based									
None	High Promise	Low Promise	High Action	Low Action		None	Direct Links	Unconditonal	Conditional	None	Fewer Roots	Fewer Links	Subsets	More Roots	More Links	Supersets	Middle Action		
									●									0.374	(0.006)
				●		○				○								0.131	(0.011)
	●					○				○								0.072	(0.012)
				●		○										●		0.064	(0.006)
○						○									●			0.045	(0.038)
○							●								●			0.045	(0.038)
	●					○								●				0.030	(0.009)
				●			●			○								0.024	(0.011)
	●					○					●							0.023	(0.006)
				●		○					●							0.020	(0.006)
	●					○											●	0.016	(0.005)
		●					●								●			0.014	(0.009)
○									●	○								0.013	(0.023)
				●					●							●		0.013	(0.005)
				●					●				●					0.012	(0.005)
				●		○									●			0.009	(0.010)
			●			○									●			0.009	(0.011)
				●		○											●	0.009	(0.007)
	●					○										●		0.009	(0.012)
	●								●								●	0.008	(0.005)
○						○						●						0.007	(0.032)
		●					●										●	0.006	(0.008)
○						○								●				0.005	(0.029)
	●								●	○								0.005	(0.004)
		●					●										●	0.004	(0.006)
	●								●							●		0.003	(0.004)
		●					●						●					0.003	(0.012)
				●			●								●			0.003	(0.010)
	●					○								●				0.003	(0.009)
		●					●								●			0.003	(0.012)
		●					●				●							0.003	(0.008)
		●				○							●					0.002	(0.010)
			●				●				●							0.002	(0.006)
		●				○				○								0.002	(0.010)
○						○											●	0.002	(0.022)
		●							●							●		0.002	(0.006)
		●							●						●			0.001	(0.007)
			●			○								●				0.001	(0.008)
		●				○							●					0.001	(0.012)

Notes: The symbol ● indicates that the corresponding criterion is being used, ○ indicates that no criterion from the corresponding class is being applied. The top row contains only a single symbol because the Conditional Correlations criterion prevents the identification of structure- and advice-based criteria. Heteroskedasticity-robust standard errors in parentheses.

Table D.7: Distribution of criteria use in Experiments 1 and 2 excluding subjects affected by contingent reasoning failure

	(1)	(2)	(3)
Criteria	Experiment 1	Experiment 2	Experiment 2
Adjusted for noise	-	No	Yes
A. Data-based			
None	0.433	0.594	0.443
	(0.011)	(0.038)	
Any Unconditional	0.171	0.106	0.146
	(0.013)	(0.050)	
Direct Links	0.115	0.106	0.146
	(0.014)	(0.050)	
All	0.056	0.000	0.000
	(0.010)	(0.056)	
Any Conditional	0.396	0.300	0.411
	(0.007)	(0.045)	
No V-Colliders	-	0.157	0.216
		(0.058)	
All	-	0.142	0.195
		(0.047)	
B. Advice-based			
None	0.109	-	-
	(0.023)		
High Promise	0.158	0.338	0.464
	(0.010)	(0.025)	
Low Promise	0.032	$0^{\dagger}$	$0^{\dagger}$
	(0.008)		
High Action	0.026	$0^{\dagger}$	$0^{\dagger}$
	(0.010)		
Low Action	0.278	0.409	0.562
	(0.010)	(0.029)	
C. Structure-based			
None	0.237	-	-
	(0.025)		
Simplicity			
Fewer Roots	0.047	-	-
	(0.011)		
Fewer Links	0.000	-	-
	(0.006)		
Subsets	0.048	-	-
	(0.014)		
Complexity			
More Roots	0.008	-	-
	(0.005)		
More Links	0.139	-	-
	(0.007)		
Supersets	0.081	-	-
	(0.013)		
Middle Action	0.044	-	-
	(0.004)		
Random choice probability	0.271	$0^{\dagger}$	$0.271^{\dagger}$
	(0.009)		
Subjects	453	267	-
Observations	11325	4005	-

Notes: This table replicates Table 4 excluding subjects affected by a contingent reasoning failure.

Table D.8 replicates Table 8 excluding the corresponding subjects in Experiment 3.

Table D.8: Distribution of criteria use in Experiment 3 excluding subjects affected by contingent reasoning failure

	(1)	(2)	(3)	(4)
Sample				*p*-value treatment effect
Explanation treatment	✓		✓	
Control	✓	✓		
Data-based criteria				
None	0.718	0.743	0.734	0.405
	(0.008)	(0.009)	(0.008)	
Unconditional	0.201	0.200	0.187	0.555
	(0.014)	(0.016)	(0.015)	
Conditional	0.081	0.057	0.079	0.319
	(0.015)	(0.018)	(0.014)	
Advice-based criteria				
None	0.070	0.003	0.066	0.279
	(0.036)	(0.035)	(0.045)	
High Promise	0.412	0.474	0.392	0.000
	(0.007)	(0.008)	(0.008)	
Low Promise	0.006	0.017	0.004	0.258
	(0.007)	(0.008)	(0.008)	
High Action	0.072	0.081	0.085	0.906
	(0.022)	(0.021)	(0.027)	
Low Action	0.359	0.369	0.374	0.772
	(0.012)	(0.011)	(0.015)	
Random choice probability	0.322	0.283	0.329	0.000
	(0.008)	(0.007)	(0.009)	
Subjects	707	338	369	
Observations	12726	6084	6642	

Notes: This table replicates Table 8 excluding subjects affected by a contingent reasoning failure.

D.5 Framing effects

We conduct our experiment in an abstract frame to prevent the influence of priors. Here, we show that framing the experiment in terms of social and biological systems causes subjects to introduce in plausibility considerations that would prevent the clean identification of decision criteria.

Framing treatments After completing the main rounds, subjects proceed through three rounds that have the same structure and parameters as previous rounds, but in which the decisions are presented in different frames. The interface no longer displays the links 'show in words' and 'show explanation.' While the theories in our main treatments only make statements about which quantities are causally connected, verbal explanations sound natural only if they also indicate whether the effect is positive or negative, which we include. We use different parametrizations and assignments of frame

26

elements to nodes. This variation occurs across subjects; each subject makes exactly one decision in each of the three frames.

Table D.9 outlines the frames. The first manipulation (column 1) keeps the decision context the same—decisions concern action, circles, squares, and bonus—but the theory part of the advisor speech bubble is replaced by a verbal description of the mechanism. The explanation differs from what subjects can see by clicking 'show in words' in the main treatment. One advisor's theory is presented as follows. "Circles are the mechanism through which the action affects the bonus. Squares are only a symptom, but not a cause of the bonus. That is, a higher action increases the number of circles, which then increases the bonus. And a higher bonus increases the number of squares, but that does not matter for your decision." For the other advisor, we exchange 'circles' and 'squares' in the foregoing explanation, but otherwise keep the explanation unchanged.

The second and third manipulations use a biological and social context, respectively, see columns 2 and 3. In the farming context, we present the interpretation corresponding to the DGP as follows: "Fertilizer directly increases rice growth. So does the presence of ladybugs. Both fertilizer and rice growth affect soil quality, growth depletes it, and fertilizer regenerates it. But soil quality is a symptom rather than a cause of rice growth in the current year. The presence of ladybugs does not depend on fertilizer use, soil quality, or rice growth." (We inform subjects that ladybugs feed on other insects, which renders causal influences between rice growth and the presence of ladybugs plausible.) The social context concerns a blended education technology, called BlendEd, presented as follows in the case of the competitor interpretation. "BlendEd enables better self-regulated learning, which, in turn, increases mathematics comprehension. BlendEd also affects computer skills, but computer skills neither affect nor depend on mathematics comprehension or on self-regulated learning." We reformulate these explanations to match the remaining DAGs.

When introducing these contexts we inform subjects that "If this round is selected for payment, you will be paid according to the rice growth that your choice of fertilizer generates (minus the costs of the fertilizer)" and "If this round is selected for payment, you will be paid according to the maths comprehension that your choice of BlendEd generates (minus the costs of the investment)." We highlight that all other aspects are unchanged from the previous rounds of the study.

Results Table D.10 analyzes behavior in the framed rounds. The six columns correspond to the three menus of Table D.9, each with two parameterizations, respectively. Each column uses data from a single parametrized menu along with the corresponding menu from the main rounds. It uses a frame indicator as a predictor in OLS regressions on an indicator for choosing the correct interpretation (panel A), an indicator for viewing any data charts (panel B), and on the number of seconds taken to make a choice (panel C).

Columns 1 and 2 show verbal descriptions of abstractly framed decisions slightly decrease the choice of the correct interpretation for one ($p < 0.1$) but not for the other parametrization, does not

Table D.9: Framing manipulation treatments

Menu	F1		F2		F3	
	DGP	Competitor	DGP	Competitor	DGP	Competitor
	$A \quad Z$ $\downarrow \quad \uparrow$ $X \rightarrow Y$	$A \quad X$ $\downarrow \quad \uparrow$ $Z \rightarrow Y$	$A \rightarrow Z$ $\searrow \uparrow$ $X \rightarrow Y$	$A \rightarrow X$ $\searrow \uparrow$ $Z \rightarrow Y$	$A \quad Z$ $\downarrow \searrow \uparrow$ $X \rightarrow Y$	$A \rightarrow Z$ $\downarrow$ $X \rightarrow Y$
Context						
C1			Farming		Education	
a			X: ladybugs, Z: soil quality			
b			X: soil quality, Z: ladybugs			
C2			Education		Farming	
a			X: computer skills Z: self-regulated learning			
b			Z: computer skills X: self-regulated learning			
Parameters						
P1		as in menu 6		as in menu 9		as in menu 17
High-promise DAG		DGP		Competitor		DGP
Low-action DAG		Competitor		Competitor		DGP
P2		as in menu 7		as in menu 10		as in menu 16
High-promise DAG		Competitor		Competitor		Competitor
Low-action DAG		Competitor		DGP		DGP

Notes: Menu labels refer to Table 2. For menu F2, we interchange the assignment of labels to nodes X and Z across subjects.

affect chart viewing but substantially increases the time subjects take to decide ($p < 0.01$ in both cases). One potential explanation for these results is that the verbal explanation is cognitively more difficult to process, though other explanations are possible.

In terms of data viewing and response times, the results in columns 3 to 6 parallel those of columns 1 and 2. There is no effect on chart viewing, and decision times drastically increase. In terms of choice, column 3 shows that choice can be substantially affected by a frame in otherwise identical choice problems. The effect applies to one but not the other real-world context, which suggests that it arises from domain-specific plausibility considerations rather than from the difference between abstract and real-world frames in general. Column 5 shows a similar though weaker effect.

We conclude that real-world frames presented verbally increase the difficulty of the problem for subjects, and introduce plausibility considerations that would impede clean identification of decision criteria in our main experiment.

Table D.10: Framing effects

	(1)	(2)	(3)	(4)	(5)	(6)
Menu	F1		F2		F3	
	DGP	Competitor	DGP	Competitor	DGP	Competitor
	$A \quad Z$ $\downarrow \quad \uparrow$ $X \to Y$	$A \quad X$ $\downarrow \quad \uparrow$ $Z \to Y$	$A \to Z$ $\searrow \uparrow$ $X \to Y$	$A \to X$ $\searrow \uparrow$ $Z \to Y$	$A \quad Z$ $\downarrow \searrow \uparrow$ $X \to Y$	$A \to Z$ $\downarrow$ $X \to Y$
Parameters	P1	P2	P1	P2	P1	P2
High-promise DAG	DGP	Comp.	Comp.	Comp.	DGP	Comp.
Low-action DAG	Comp.	Comp.	Comp.	DGP	DGP	DGP

A. Correct interpretation chosen

Context						
Symbols	-0.068*	-0.031				
	(0.038)	(0.056)				
Farming			0.040	-0.088	-0.111**	-0.117*
			(0.070)	(0.065)	(0.051)	(0.062)
Education			-0.208***	-0.012	-0.063	-0.106*
			(0.071)	(0.064)	(0.051)	(0.063)
Constant	0.648***	0.466***	0.555***	0.683***	0.833***	0.752***
	(0.027)	(0.039)	(0.041)	(0.037)	(0.022)	(0.036)

B. Any data charts viewed

Symbols	0.012	0.025				
	(0.037)	(0.052)				
Farming			0.045	0.026	-0.074	0.080
			(0.066)	(0.061)	(0.062)	(0.063)
Education			-0.032	0.085	0.018	0.031
			(0.067)	(0.060)	(0.062)	(0.064)
Constant	0.660***	0.671***	0.685***	0.708***	0.657***	0.652***
	(0.026)	(0.037)	(0.038)	(0.035)	(0.027)	(0.037)

C. Time to decide in seconds

Symbols	16.585***	18.463***				
	(3.489)	(4.662)				
Farming			51.905***	50.446***	31.154***	47.369***
			(9.273)	(7.544)	(6.927)	(8.159)
Education			50.025***	48.283***	38.393***	48.628***
			(9.358)	(7.451)	(6.850)	(8.261)
Constant	47.508***	45.325***	36.429***	41.624***	45.557***	42.646***
	(2.467)	(3.296)	(5.378)	(4.328)	(2.954)	(4.740)
Observations	648	322	292	322	470	322

Notes: Each column in each panel displays the estimates of an OLS regression of the dependent variable on context indicators using data from a single decision problem in different frames. Behavior in the abstract frame (main rounds) constitutes the baseline throughout. Standard errors in parentheses, clustered by subject. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

D.6 Estimates with restricted criteria spaces

Table D.11 shows that our mixture model yields substantially different distributions of data-based criteria when we exclude advice-based or structure-based criteria from our specification than when we estimate the full model. Column 1 reproduces the distribution of data-based criteria from column 1 of Table 4. Column 2 estimates the model excluding all types that make use of an advice-based criterion. The fraction of subjects estimated to use the Conditional Correlations criterion drops by 11.4 percentage points, and that corresponding to no data-based criteria increases by 10.6 percentage points. (While these directional changes may be surprising, recall that omitted variable bias yields inconsistencies in the estimates of all parameters in unpredictable directions even in OLS regressions.) Excluding structure-based criteria (column 3) increases the estimated fraction of subjects using the Conditional Correlations criterion by 9.6 percentage points, presumably because choices following a structure-based criterion are attributed to a data-based criterion when structure-based criteria are excluded. When we exclude both advice- and structure-based criteria (column 4), we set the fraction of subjects using no data-based criterion to zero due to the fact that we cannot identify an implementation noise parameter separately from the frequency of a pure noise type (see Section 3.1). While this mechanically increases the fraction of types attributed to a data-based criterion, the fact that the increase is much larger for the Conditional Correlations criterion than for the Unconditional Correlations criteria shows that the exclusion of the two classes of criteria substantially changes inference over data-based criteria.

D.7 Checking of explanations

Throughout the experiment, subjects had access to buttons displaying additional explanations they had seen in the instructions. Data on button usage addresses questions about the extent to which our results are driven by the availability of these explanations. We consider usage data from Experiment 2 because of the clear ranking of menus in terms of the data-based criteria required to identify the DGP.

Column 1 of Table D.12 lists the frequency with which subjects clicked the button to reveal the explanation concerning the correlational implications of causal structures. The first row shows that 22% of subjects check the explanation in at least one of the practice rounds. According to the second row, merely 16.1% check it in at least one of the main rounds. The fact that these rates are low suggests that the availability of the explanations is not a main driver of our results. The remaining rows show that subjects do not check the explanations more frequently in more difficult decisions. We see that 6.8%, 3.9%, 5.0%, and 3.2% check the explanation in at least one of the rounds in which the the Direct Links, Unconditional Correlations, Conditional Correlations (No v-Colliders), and Conditional Correlations criteria are necessary to identify the correct interpretation, respectively.

Table D.11: Distribution of data-based criteria use in restricted models

	(1)	(2)	(3)	(4)
	Full model	Excluded criteria class		
		Advice-based	Structure-based	Advice- and structure-based
Data-based criteria				
None	0.461	0.567	0.409	$0^{\dagger}$
	(0.009)	(0.005)	(0.003)	
Any Unconditional	0.165	0.173	0.121	0.200
	(0.011)	(0.006)	(0.004)	(0.002)
Direct Links	0.107	0.115	0.038	0.177
	(0.012)	(0.005)	(0.004)	(0.002)
All	0.058	0.058	0.083	0.024
	(0.008)	(0.005)	(0.003)	(0.003)
Any Conditional	0.374	0.260	0.470	0.800
	(0.006)	(0.004)	(0.002)	(0.002)
Random choice probability	0.266	0.187	0.375	0.647
	(0.008)	(0.006)	(0.005)	(0.002)
Subjects	475	475	475	475
Observations	11875	11875	11875	11875

Notes: † indicates imposed values. Heteroskedasticity-robust standard errors in parentheses. Estimates in column 1 represent the output of the full model. Estimates in column 2 reflect the output of a model that excludes advice-based criteria. Column 3 presents the corresponding estimates for a model excluding structure-based criteria. Estimates in column 4 reflect a model that excludes both advice- and structure-based criteria.

Column 2 shows the frequencies with which subjects check the explanation on how to read our data charts. Column 3 displays data on how often subjects check the explanation of the costs. The qualitative patterns mirror those of column 1, with no more frequent checking in more challenging problems.

Table D.12: Frequency of checking explanations

| | (1) | (2) | |
Explanation	Correlational implications of causal structures	Data charts	Costs
Preliminary rounds	0.222	0.348	0.237
	(0.025)	(0.029)	(0.025)
All data-based rounds	0.161	0.233	0.100
	(0.022)	(0.025)	(0.018)
D1, D2, D3	0.068	0.115	0.036
	(0.015)	(0.019)	(0.011)
U1, U2	0.039	0.061	0.022
	(0.012)	(0.014)	(0.009)
NV1, NV2	0.050	0.065	0.032
	(0.013)	(0.015)	(0.011)
C1, C2	0.032	0.043	0.018
	(0.011)	(0.012)	(0.008)

Notes: Each entry in the table shows the fraction of subjects in Experiment 2 who view the corresponding explanation at least once in the set of menus indicated in the first column.

D.8 Self-classification

Design We elicit subjects' self-reported decision-making strategies in 3 hypothetical scenarios, displayed in random order. The first scenario describes the decision to eat less food with additives E250 and E252 that have recently been linked to cancer. We describe a competing interpretation by which the relation between the consumption of these additives and cancer arises artifactually as a consequence of a generally unhealthy lifestyle. We intentionally choose additives labels that subjects are unlikely to know, so they approach the problem as novel. The second scenario describes the decision to consume more foods that contain lycopene, a substance that has been linked with health benefits, but in a way that may also reflect a general tendency for a healthy lifestyle. The third scenario considers a business setting in which action needs to be taken to reverse a decline in sales.

Subjects then indicate their agreement with each of the following statements, appropriately adapted to each scenario. (i) In situations like this, I tend to follow advice that gives me more hope about my health outcomes. (ii) Unfortunately, in situations like this, advice that gives me less hope is more often the right advice to take, so I tend to follow that. (iii) In situations like this, I tend to follow advice that enables me to take action, rather than sticking with the status quo. (iv) In situations like this, I tend to follow the advice that appears to be the easiest. (v) In this situation, I would

check the facts to determine whether there really is a correlation between the consumption of these additives and cancer risk. (vi) In this situation, I would look into the facts. I would only consider people with an otherwise healthy lifestyle, to determine whether a relation between the consumption of these additives and cancer risk is present for this population. (vii) In situations like this, reality is rarely simple. I am more likely to follow advice that is based on more comprehensive and elaborate theories. (viii) In situations like this, simpler explanations are usually better. I am more likely to follow advice based on simpler and more straightforward theories. (ix) In this situation, I would check the facts with a focus on determining which analyses are most methodologically sound to ensure any correlations are not spurious.

The statements are intended to capture the following criteria: (i) High Promise, (ii) Low Promise, (iii) High Action, (iv) Low Action, (v) Unconditional Correlations, (vi) Conditional Correlations, (vii) simplicity preference, (viii) complexity preference. Statement (ix) captures a general tendency for scrutinizing data.

Results We consider how subjects' self-descriptions of their information search behavior relate to their experimental choices, using the same econometric specifications as Table 6, except that we use self-reported decision strategies as predictor variables rather than demographics. The results need to be interpreted in light of existing research that shows that people are often inept at self-assessment (e.g. Kruger and Dunning, 1999). Some of the correlations we find appear intuitive, others counterintuitive. Amongst the former, we observe that a propensity for following advice that appears hopeful is negatively related to selecting the correct interpretation ($p < 0.05$) and to viewing data charts ($p < 0.05$), and positively related to choosing the advisor making the higher promise ($p > 0.1$). A self-reported propensity for checking conditional correlations is also consistent with our expectations; it positively predicts selecting the correct interpretation ($p < 0.05$), though without a strong effect on the propensity to view data charts. Relatedly, a propensity for checking the methodological soundness of studies positively predicts choosing the correct interpretation ($p < 0.1$) and the propensity to check data charts ($p < 0.01$). Finally, individuals with a preference for simpler theories check data less often ($p < 0.01$) and choose the correct interpretation less often ($p < 0.01$). Turning to counterintuive results, we observe that a self-reported propensity to follow the advice that appears easiest positively relates to choosing the correct interpretation ($p < 0.01$), and fails to predict choosing the cost-minimizing advisor. Moreover, a self-reported propensity for checking unconditional correlations negatively predicts choosing the correct interpretation and checking data charts ($p < 0.05$) but positively predicts taking the cost-minimizing action. Self-reported propensities to choose more pessimistic advice, advice that permits taking action, and a preference for more comprehensive and elaborate theories no not have statistically significant predictive power on any dependent variable in columns 1 to 4.

The foregoing analysis only employs a subset of the moments that our mixture model employs. We thus estimate the multinomial logit specification of the three-type version of our mixture model that we introduced in Section 4.6. The results are often consistent with the corresponding results from columns 1 to 4, and usually more highly statistically significant. Subjects with a self-reported propensity to adopt more hopeful advice more frequently correspond to the High Promise ($p < 0.01$) and Low Action ($p < 0.01$) types. These subjects are thus also less likely to be consistent with the Conditional Correlations type, our baseline category. We observe directionally similar effects for subjects who prefer advice that lets them take action, who report using unconditional correlations to determine the soundness of interpretations, and for subjects who prefer simpler to more complex interpretations. Subjects who report using conditional correlations to determine the soundness of interventions, by contrast, less often choose consistently with the High Promise ($p < 0.01$) and Low Action ($p < 0.01$) types (and thus more often with the Conditional Correlations type). A directionally similar result holds for subjects who report attempting to determine the methodological soundness of the analyses on which interpretations are based.

Table D.13: Self-classification and choice in Experiment 1

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Reduced-form estimates				Structural estimates		
	Correct DAG	Data viewed	High Promise	Low Action	High Promise	Low Action	p-value difference
Constant	0.662***	0.718***	0.471***	0.488***	-1.126***	-0.413***	0.000***
	(0.058)	(0.107)	(0.044)	(0.048)	(0.106)	(0.030)	
Advice that gives more hope	-0.088**	-0.190**	0.050*	0.029	1.091***	0.554***	0.101
	(0.040)	(0.084)	(0.029)	(0.038)	(0.220)	(0.207)	
Advice that gives less hope	0.013	-0.064	-0.030	-0.046	-0.121	-0.338	0.582
	(0.050)	(0.102)	(0.035)	(0.047)	(0.214)	(0.291)	
Advice that lets me take action	-0.031	0.003	0.010	0.021	0.574***	0.399*	0.610
	(0.047)	(0.089)	(0.032)	(0.042)	(0.211)	(0.240)	
Advice that appears easiest	0.125***	0.147	0.060*	0.002	0.071	-0.987***	0.009***
	(0.047)	(0.093)	(0.031)	(0.040)	(0.216)	(0.305)	
Correlation (unconditional)	-0.107**	-0.228**	0.063	0.092**	1.103***	0.868**	0.595
	(0.054)	(0.094)	(0.039)	(0.042)	(0.255)	(0.342)	
Correlation (conditional)	0.079**	0.058	-0.025	-0.046	-0.843***	-0.662***	0.538
	(0.039)	(0.073)	(0.025)	(0.036)	(0.189)	(0.206)	
Determine methodological soundness	0.088*	0.259***	-0.052	-0.027	-1.410***	-0.584**	0.045**
	(0.050)	(0.094)	(0.036)	(0.046)	(0.279)	(0.265)	
More comprehensive and elaborate theories	-0.030	0.084	-0.008	-0.048	0.451**	0.226	0.476
	(0.042)	(0.074)	(0.031)	(0.035)	(0.179)	(0.246)	
Simpler explanations usually better	-0.137***	-0.263***	0.039	0.048	1.336***	1.097***	0.456
	(0.042)	(0.082)	(0.034)	(0.038)	(0.228)	(0.205)	
Observations	12115	12115	12115	12115		11875	
Subjects	485	485	485	485		475	

Notes: Columns 1 to 4 report coefficient estimates from OLS regressions. Columns 5 and 6 represent estimated odds ratio from a single GMM estimation. Predictor variables are normalized to range from 0 to 1. Each variable is the average of a subject's answers to the corresponding questions across all three scenarios for which we asked the question. Coefficient estimates for the structural model show the effect of the predictor on the log-odds of being the specified type rather than the conditional tester. Asterisks in columns 1 to 4 reflect tests of the null hypothesis that the corresponding parameter value is zero. Asterisks in columns 5 and 6 reflect tests of the null hypothesis that the corresponding odds ratio is one. p-values in column 7 reflect Wald tests of the joint hypothesis that the two estimates on a given predictor equal each other (1 degree of freedom). Each regression in columns 1-4 includes 25 observations for each of the 485 subjects in Experiment 1 who have provided complete answers to all questions, with the exception of 10 subjects in the first session who were not shown round 25. The model in columns 5-7 excludes the 10 subjects from the first session entirely. Standard errors in parentheses, clustered by subject. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D.9 Effect of a fact-checking nudge in the laboratory sample

Section 4.4 argues that the comparison of the use of data-based criteria across the experiments suggests that the failure to apply data-based criteria is at least partly due to limited ability, not solely unwillingness, though in a comparison that uses different menus. Here, we show that this argument also applies when we focus on the two menus (6 and 10) included in both experiments.

Table D.14 shows the effect of making advice available. As columns 1 and 2 show, in both menus, it decreases the fraction of subjects who view at least one data chart by around 10 percentage points. As columns 3 and 4 show, there is no effect on the choice of the correct interpretation.

Table D.14: Effect of availability of advice in Experiments 1 and 2

VARIABLES	(1) View a data chart	(2)	(3) Correct interpretation chosen	(4)
Menu	6	10	6	10
Advice shown	-0.094***	-0.102***	0.024	0.056
	(0.033)	(0.032)	(0.036)	(0.036)
Constant	0.760***	0.789***	0.624***	0.616***
	(0.026)	(0.024)	(0.029)	(0.029)
Observations	764	764	764	764

Notes: Standard errors in parentheses, clustered by subject. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

E Experiment instructions

E.1 Experiments 1 and 2

Welcome!

This is a research study run by the Department of Economics at the University of Zurich, Switzerland, and the Norwegian School of Economics.

This study will take between one and one-and-a-half hours to complete. The average participant will earn about Fr.40 for completing this study. This consists of the base payment of Fr.20 that you will get just for completing the study, and an additional payment of, on average, Fr.20. The additional payment depends on your decisions and on luck.

Anonymized data from this study may be shared publicly.

This study has been approved by the Ethics Boards of the Faculty of Economics at the University of Zurich and at the Norwegian School of Economics.

Do you consent to participating in this study?

| No, I do not consent. Please end my participation. ○ | Yes, I consent. Please continue. ○ |

››

Instructions

Please read these instructions carefully. There will be a comprehension check. You will only be able to continue with the study once you have correctly answered all comprehension check questions.

(You will not be able to pass these questions by guessing or by trial and error.)

The Main part of this study has 29 rounds.
In an additional part, you will make decisions in 2 more rounds and we will ask several questions about you.

Your study payment equals your earnings from one randomly chosen round. Each round is equally likely to be chosen. Your earnings depend on your choice in that round and on a bit of luck.

Here's how. In each round you choose how much to spend on an action. That spending may influence a bonus. For the randomly selected round:

Your study payment = bonus – spending on the action

Sometimes, your spending increases the bonus, sometimes it doesn't. Whether it does, and by how much, depends on a mechanism, which you'll learn about next.

The mechanisms

In each round, some mechanism determines whether and how your spending on the action (shown as 👆) affects the bonus (shown as 💵). Next to your action and the bonus, each mechanism involves a third quantity that we call "dots" (shown as ⬤) and a fourth quantity we call "squares" (shown as ⬜).

Any two quantities may or may not be directly related:

- 👆 may or may not directly affect 💵
- 👆 may or may not directly affect ⬤
- ⬤ may or may not directly affect 💵
- 💵 may or may not directly affect ⬤
- Etc.

A quantity directly affects another if changing the first quantity directly *causes* the second quantity to change (even if other quantities are held fixed).

A mechanism specifies which quantities are related and how.

Pictures like the following will show what can affect what

For instance, in this mechanism

the action can affect the dots (as you can see by the arrow from 👆 to ⬤), and the dots can affect the bonus (as you can see by the arrow from ⬤ to 💵). The action cannot affect the bonus directly (there is no arrow from 👆 to 💵), but it can affect it *indirectly*, by changing the dots.

The direction of the arrows is important. This mechanism, for instance

is different from the previous mechanism. Here, the arrow is from 💵 to ⬤, not from ⬤ to 💵. Hence, the bonus can affect the dots, but the dots cannot affect the bonus. As in the previous mechanism, 👆 can affect ⬤. But because the dots can no longer affect the bonus, the action no longer has an indirect effect on the bonus. Neither can the bonus affect the action, so the two quantities are unrelated in this mechanism.

There will be mechanisms other than these two. They differ from each other by what can affect what. You can always see this by whether there is an arrow between two quantities and by the direction of these arrows.

To repeat:
Your action may affect the bonus. Whether and how it does so depends on the mechanism in the current round.

Note:
In each round of the experiment, a different mechanism is at play.

How do I decide how much to spend on the action?

In each round, two or three advisors recommend how much you should spend on the action. You select a spending level by choosing which advisor's recommendation to follow.

In each round, only one advisor will be correct. We will not tell you which one that is.

You select an advisor by clicking on their icon, like this:
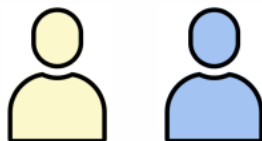
(This choice is just an example. It has no consequences for anything in this study. The Next button will appear once you have clicked on an advisor.)

(next page)

Where do advisor recommendations come from?

We have data from running the mechanism in the current round many times. It shows the action, the number of dots, the number of squares, and the bonus that obtained each time.

Each advisor has a theory about what the mechanism in the current round is. Each advisor derived their recommendation by analyzing the data of that round from the viewpoint of their own theory. Some advisors have the wrong theory about the mechanism in the current round. These advisors' recommendations and their prediction of your study payment are wrong, too.

(next page)

Advisors' favorite theories

Each advisor has a favorite theory about the mechanism. They will tell you what their theory is, along with their recommendation, and predictions about your bonus, as you will see on the next page. (You will see only one advisor, for illustration. You are not yet making any choices.)

We now explain each part of an advisor's theory and recommendation in detail.

Ok

My recommendation
Here's the best you can do:

Spend **Fr.8.00 on the Action.**
Expect **Payment Fr. 22.64**
(=bonus Fr.30.64
– action cost Fr.8.00)

Show explanation

My theory
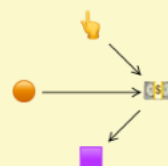
Show in words

How do I choose between advisors?

You may choose between advisors in whatever way you like.

In case you consider it useful, you can see data charts from the mechanism of the current round. You will see a dashboard like this. A data chart will show up as soon as you click on any link.

Data on this round's mechanism

Relation between pairs of quantities:	Relations fixing third quantity:
🔴 by 👇	🔴 by 👇 fixing 💴
🟩 by 👇	🟩 by 👇 fixing 💴
Show remaining charts	Show remaining charts

If you click on some link above, the corresponding data chart will appear.

[Show me!]

Relations between two quantities

The left-hand column of the data dashboard shows whether higher values of one quantity tend to coincide with higher values of the other quantity.

Data on this round's mechanism

Relation between pairs of quantities:	Relations fixing third quantity:
🔴 by 👇	🔴 by 👇 fixing 💴
🟧 by 👇	🟧 by 👇 fixing 💴
Show remaining charts	Show remaining charts

For instance, you will be able to see the relation between the number of 🟧 and the money spent on the action. In some round, this could look like this:

🟧 by 👇

Average of 🟧 (y-axis: 0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56)

Money spent on 👇 (x-axis: 1 to 3, 3 to 6, 6 to 10, 10 to 12, 12 to 15)

This graph means that when the amount of money spent on the action was between 1 and 3 francs, the average number of 🟧 was about 28. And when amount of money spent on the action was between 12 and 15, the average number of 🟧 was about 50.

Relations between two quantities holding the third quantity fixed

The right-hand column shows whether higher values of one quantity tend to coincide with higher values of the other quantity *if you hold the third quantity fixed.*
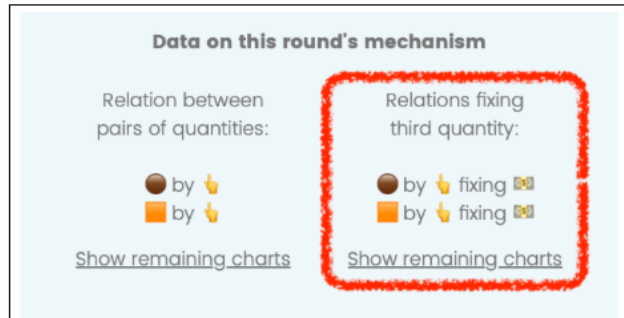
For instance, you will be able to see the relation between the number of 🟧 and the money spent on the action *holding the number of* 🔵 *fixed.* In some round, this could look like this:
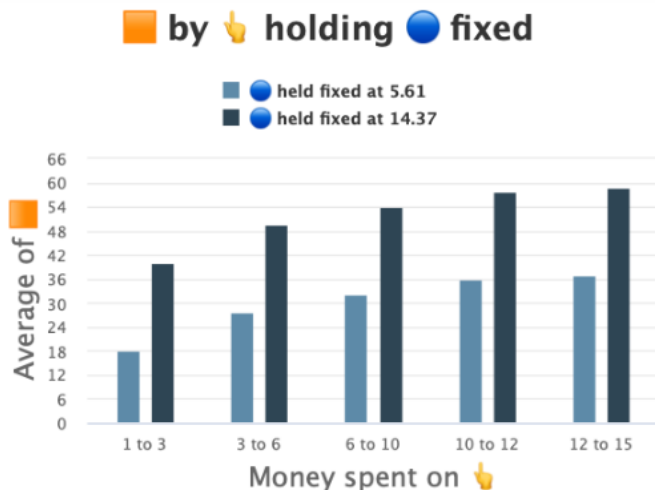
The lighter bars in this graph show how spending and the number of 🟧 related *when the number of* 🔵 *was 5.61.* In this case, when spending on the action was between 1 and 3 , the average number of 🟧 was about 18. When spending on the action was between 12 and 15, the average number of 🟧 was about 36. The darker bars in this graph show how spending and the number of 🟧 related *when the number of* 🔵 *was 14.37.* For action spending between 1 and 3, an average number of 🟧 of about 40 obtained. If action spending was between 12 and 15, that average was nearly 60.

Important

1. Stubborn advisors

- Data cannot change an advisor's theory about the mechanism. But each advisor relies on data to find what their theory implies is the best action for you to take. Each advisor does this correctly. In other words: the data do not affect the advisors' opinions about **whether** their mechanism is right. They only affect advisors' recommendations. Advisors always recommend the action which—if their theory is correct—is genuinely best for you.
- While data cannot change any advisor's theory about what can affect what, it might change **your** view on which theory is right, and hence, on which advisor you wish to follow.

2. Data from running the mechanism

- In all the data you see in any of the rounds, 👆 is always chosen independently and randomly. It is never influenced by any other quantity.
- If a mechanism has two quantities and neither of them is influenced by any other quantity, then these two quantities will be unrelated.
- The relation between any two quantities always involves variability. For instance, consider a mechanism in which a higher 👆 tends to raise 🎰. This does not mean that a higher 👆 certainly raises 🎰, it just means a higher 🎰 is more likely.
- The ranges of the quantities plotted in the data charts are chosen arbitrarily.

You don't need to inspect the data charts, but we want you to know that you can.

Correlation and causation

Here are some other things you may want to (but do not have to) consider if you choose to inspect the data:

Two quantities may appear related because the first causes the second.

Real world example
Rain and wet clothes occur together because rain causes wet clothes.

Experiment example
More ⚫ and a higher bonus may occur together because ⚫ cause the bonus
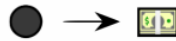
Two quantities may appear related because the second causes the first.

Real world example
Fire engines and forest fires occur together even though fire engines do not cause forest fires.

Experiment example
More ⚫ and a higher bonus may occur together because the bonus causes ⚫.

Even though neither of two quantities may cause the other, they appear related if a third quantity affects them both simultaneously.

Real world example
Bald people are more often men than women, and people who wear bow-ties are also more often men rather than women. But neither does baldness cause bow-ties, nor do bow-ties cause baldness.
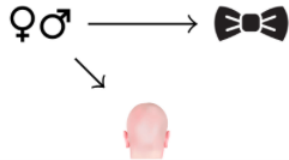
Experiment example
A higher bonus might go along with more ⚫. This might be because the action raises both the number of ⚫ and the bonus, even though ⚫ do not affect the bonus, and the bonus does not affect the number of ⚫.

If a quantity affects another only indirectly (i.e. through an intermediary variable), then the two quantities will appear unrelated once you hold the intermediary variable fixed.

Real world example
Excessive salt consumption raises blood pressure, which, in turn, increases the frequency of heart attacks. Hence, excessive salt consumption and heart attacks are related. Yet, if we only consider people with normal blood pressure levels, we do not see a relation between salt consumption and heart attack frequency. Neither do we see such a relation amongst people with high blood pressure.
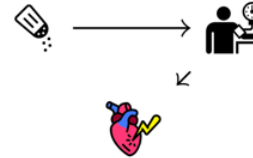
Experiment example
A higher action may increase the number of ⚫, which, in turn, increase the bonus. Hence, action and bonus are related. Yet, if we hold the number of ⚫ fixed, we do not see a relation between the action and the bonus.

Even though neither of two quantities may cause the other, if they both affect a third quantity, they may appear related once you hold the third quantity fixed.

Real world example
Yamaha corporation sells both pianos and motorcycles. Neither do piano sales affect motorcycle sales, nor do motorcycle sales affect piano sales. Yet, Yamaha's total revenue (the third quantity) consists of revenue from pianos and revenue from motorcycles (as well as revenue from a few other products). If you only consider years in which Yamaha's revenue say, 100 million, it will look as if high piano sales coincide with low motorcycle sales, and vice versa. But that's just because you are looking only at years in which revenue from pianos and revenue from motorcycles (and that of a few other products) sum to 100 million, so if one is higher, the other *must* be lower.
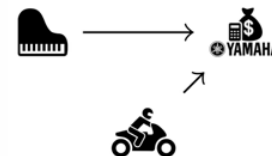
Experiment example
A higher action may increase the number of ⚫, and the bonus may increase the number of ⚫. Neither does the action affect the bonus nor does the bonus affect the action. Yet, they both affect the number of ⚫. If you consider the relation between the action and the bonus when you hold the number of ⚫ fixed (say, at 5), it will look as if the action decreases the bonus. But that's just because you are looking only cases in which the action and the bonus together cause a total number of five ⚫.

Spending on the action

You can increase the action by spending more on it. However, the more you are already spending on the action, the less effective additional spending becomes. Each advisor's recommendation already takes this into account.

But we let you know about how spending on the action affects the action, so you know what to expect if you happen to pick an advisor whose theory is wrong.

Recall,

Your study payment = bonus (💵) – spending on the action.

The way in which your spending affects the action is the same in each round. The figure below show details.

The more a higher action raises the bonus, the more you optimally want to spend on the action.

Here's why. You want to increase your spending as long as doing so increases your bonus by more than your additional spending. If raising 👆 raises 💵 a lot, then you can spend a lot until spending becomes so ineffective that it raises 💵 by less than your spending -- even a small increase in the action will lead to a relatively large increase in 💵. However, if raising 👆 raises 💵 only a little, then it only makes sense to raise spending as long as the effect on the action is quite large. Hence, in this case, you want to spend relatively little on the action. And if raising 👆 lowers 💵 or does not change it at all, then you do not want to spend money on 👆 at all.

Your spending on the action will be discounted from your study payment, no matter whether the advisor you follow is correct or not.
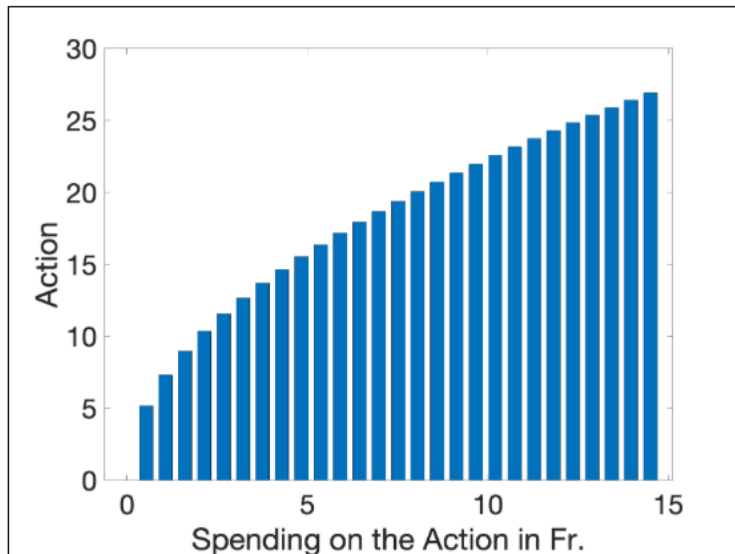
Unavailable information

In some rounds, you will not have access to the data dashboard.

In other rounds, you will not see what action the advisors recommend or the bonus the advisor expects if you follow the recommended action.

Importantly, even in these rounds, the advisor you choose determines what action you take, and therefore how much you will earn from this experiment, if that round is selected for payment.

Hence, think carefully about your choice in all rounds, even if some information is unavailable in a given round.

Summary

This study has 29 rounds

- You will be paid for one single decision from this entire study, drawn at random.
- In each round, you choose an action by picking one advisor whose recommendation you will follow. The action may or may not affect the bonus. For the round that is randomly chosen to be carried out, your study payment will equal the bonus minus the amount you spent on the action in that round.
- In each round, a different mechanism determines the relation between the action, the numbers of circles and squares, and the bonus. We do not tell you which mechanism is in effect in any given round.
- In each round, two or three advisors have inspected data from the mechanism in the current round through the lens of their theory. They each recommend what action to take, and the study payment you can expect if you take the recommended action and if their theory about the mechanism is correct.
- Advisors may be mistaken about the mechanism in effect in the current round. But they never make mistakes in calculating the implications of their theory. Hence, an advisor's recommendation may be wrong only because he has the wrong theory, but not because he makes mistakes in applying his own theory.
- You will be able to inspect data charts. We have made them by repeatedly running the mechanism that actually determines how the four quantities are related to each other in the current round.

Comprehension check

To make sure you understand these instructions, please select all the correct statements below (and leave the incorrect statements unselected). If you do not pass the first or second time, please go back to review the instructions. If you still cannot pass, please contact the study personnel.

(If you cannot pass in spite of having re-checked the instructins, raise your hand.)

- ☐ Advisors believe in their theory so strongly that data does not change their opinion of whether their theory is right. They only use the data to derive 👆 that is best *if* their theory is right.

- ☐ All advisors are always right about the mechanism in the current round.

- ☐ Which 👆 is best to take does *not* depend on the relation between 👆 and 🎰.

- ☐ Advisors might be mistaken about the mechanism, but if they have the mechanism right, they always recommend the Action that is truly best.

- ☐ Regardless of whether my chosen advisor is right or wrong, I will pay the cost of the action recommended by the advisor I have chosen.

- ☐ It's best to take a high action if the action has no effect on 🎰, and to take a low action if the action has a strong effect on 🎰.

- ☐ I will be paid for exactly one round, randomly selected from all the rounds of this experiment. That single round will entirely determine my payment for this study.

- ☐ In each round, I will earn something. My total payment for the experiment will be the sum of all these earnings across the rounds.

- ☐ The direction of the arrows in a mechanism (for instance this one 👆→● 🎰) is irrelevant.

- ☐ In each round, the relation between 👆, ●, ▨, and 🎰 is determined by a different mechanism.

A new round starts

A **new mechanism** relates
your Action, the number of circles, the number of squares, and the
Bonus.

››

Pick your advisor and action

Make a choice by clicking on them. Then click Next.

Data on this round's mechanism

Relation between
pairs of quantities:

Relations fixing
third quantity:

Show remaining charts

💵 by 👆 fixing 🔵
💵 by 🔵 fixing 🟧
💵 by 🟧 fixing 🔵
🔵 by 👆 fixing 🟧
💵 by 👆 fixing 🟧

Show remaining charts

How to read these charts   Explanation correlation and causation   Explanation costs

My recommendation

Here's the best you can do:

Spend **Fr.3.45 on the Action.**
Expect **Payment Fr. 22.50**
(=bonus Fr.25.96
- action cost Fr.3.45)

Show explanation

My theory

👆
🟧 → 🔵
💵

Show in words

My recommendation

Here's the best you can do:

Spend **Fr.12.48 on the Action.**
Expect **Payment Fr. 17.50**
(=bonus Fr.29.98
- action cost Fr.12.48)

Show explanation

My theory

👆
🔵 → 🟧
💵

Show in words

Click on an advisor to select the Action they recommend.

Pick your advisor and action

Make a choice by clicking on them. Then click Next.

Data on this round's mechanism

Relation between
pairs of quantities:

Relations fixing
third quantity:

*This information
is hidden in this round.*

*This information
is hidden in this round.*

How to read these charts | Explanation correlation and causation | Explanation costs

My recommendation

*Information hidden
(but the recommended action
still determines your payoff)*

My theory

👆

Show in words

My recommendation

*Information hidden
(but the recommended action
still determines your payoff)*

My theory

👆
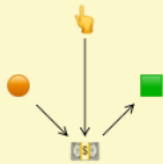
Show in words

Click on an advisor to select the Action they recommend.

In this round, the advisors explain their theories in words.

Instructions

In this round, the advisors explain their theories in words. They no longer show graphical representations. Their theories involve the usual four quantities:

- Your action
- Circles
- Squares
- Bonus

All other aspects are unchanged from the previous rounds of this study.

<< >>

Past data

Relation between pairs of quantities:

Relations fixing third quantity:

Show remaining charts

Circles by Action fixing Bonus
Bonus by Action fixing Circles
Squares by Action fixing Bonus
Squares by Action fixing Circles
Circles by Action fixing Squares
Bonus by Action fixing Squares

Show remaining charts

How to read these charts | Explanation correlation and causation | Explanation costs

---

My theory

Squares are the mechanism through which the action affects the bonus.

Circles are only a symptom, but not a cause of the bonus. That is, a higher action increases the number of squares, which then increases the bonus. And a higher bonus increases the number of circles, but that does not matter for your decision.

My recommendation

Here's the best you can do:

Spend **Fr.5.13 on Action.**
Expect **Payment Fr. 17.50**
(=Bonus Fr.22.63
- Action cost Fr.5.13)

My theory

Circles are the mechanism through which the action affects the bonus.

Squares are only a symptom, but not a cause of the bonus. That is, a higher action increases the number of circles, which then increases the bonus. And a higher bonus increases the number of squares, but that does not matter for your decision.

My recommendation

Here's the best you can do:

Spend **Fr.12.50 on Action.**
Expect **Payment Fr. 22.50**
(=Bonus Fr.35.00
- Action cost Fr.12.50)

In this round, you are a rice farmer.

Instructions

In this round, you are a rice farmer. Your action is to select the amount of fertilizer to use. Your objective is to maximize rice growth minus the cost of fertilizer.

If this round is selected for payment, you will be paid according to the rice growth that your choice of fertilizer generates (minus the costs of the fertilizer).

There are two experts advising you on your use of fertilizer. Their theories involve four quantities:

- Rice growth
- Fertilizer use
- Soil quality
- Presence of ladybugs (which feed on other insects)

All other aspects are unchanged from the previous rounds of this study.

<< >>

My theory

Fertilizer directly increases rice growth. So does soil quality.

Both rice growth and fertilizer directly affect the presence of ladybugs; growth attracts them, and fertilizer repels them. But their presence is a symptom rather than a cause of rice growth.

Soil quality in the present year does not depend on fertilizer use, ladybug presence, or rice growth.

My recommendation

Here's the best you can do:

Spend **Fr.0.55 on fertilizer.**
Expect **Payment Fr. 22.51**
(=sales Fr.23.06
- fertilizer cost Fr.0.55)

My theory

Fertilizer directly increases rice growth. So does the presence of ladybugs.

Both fertilizer and rice growth affect soil quality, growth depletes it, and fertilizer regenerates it. But soil quality is a symptom rather than a cause of rice growth in the current year.

The presence of ladybugs does not depend on fertilizer use, soil quality, or rice growth.

My recommendation

Here's the best you can do:

Spend **Fr.8.00 on fertilizer.**
Expect **Payment Fr. 17.50**
(=sales Fr.25.50
- fertilizer cost Fr.8.00)

In this round, you are a maths teacher.

Instructions

In this round, you are a maths teacher. Your action is to choose how much to invest in BlendEd, a blended education technology that combines traditional classroom teaching with online learning. Your wish to increase students' maths comprehension, as measured by their grades.

If this round is selected for payment, you will be paid according to the maths comprehension that your choice of BlendEd generates (minus the costs of the investment).

There are two experts advising you on your use of blended learning. Their theories involve four quantities:

- Amount of BlendEd.
- Student maths comprehension.
- Students' self-directed learning.
- Students' computer skills.

Self-regulated learning involves students setting their own speed of instruction and learning, and regulating their learning strategies themselves.

All other aspects are unchanged from the previous rounds of this study.

<< >>

Past data

Relation between pairs of quantities:

Relations fixing third quantity:

Show remaining charts

Math comprehension by Self-dir. learning fixing BlendEd
Self-dir. learning by Computer skills fixing BlendEd
Math comprehension by BlendEd fixing Computer skills
Self-dir. learning by BlendEd fixing Math comprehension
Math comprehension by Self-dir. learning fixing Computer skills
Self-dir. learning by Computer skills fixing Math comprehension

Show remaining charts

How to read these charts | Explanation correlation and causation | Explanation costs

---

My theory

BlendEd directly increases mathematics comprehension. In addition, it enhances self-regulated learning, which also increases mathematics comprehension.

Computer skills directly depend on mathematics comprehension, but BlendEd affects it only through its effect on mathematics comprehension.

My recommendation
Here's the best you can do:

Spend **Fr.4.96 on BlendEd.**
Expect **Payment Fr. 22.50**
(=Maths comprehension Fr.27.46
- BlendEd cost Fr.4.96)

My theory

BlendEd enables better self-regulated learning, which, in turn, increases mathematics comprehension.

BlendEd also affects computer skills, but computer skills neither affect nor depend on mathematics comprehension or on self-regulated learning.

My recommendation
Here's the best you can do:

Spend **Fr.11.50 on BlendEd.**
Expect **Payment Fr. 17.52**
(=Maths comprehension Fr.29.02
- BlendEd cost Fr.11.50)

Click on an advisor to select the Action they recommend.

Part 2

You have completed Part 1 of this study.

Part 2 consists of decisions in two rounds. Additionally, we will ask several questions about yourself, spread over approximately 10 pages.

Your risk attitudes

On each of the next two pages, you will choose a lottery from a menu, such as this:

Choose the alternative you prefer most					
Get **$35 or $1** with 50/50 chance	Get **$30 or $6** with 50/50 chance	Get **$26 or $8** with 50/50 chance	Get **$22 or $10** with 50/50 chance	Get **$18 or $12** with 50/50 chance	Get **$14** for sure
○	○	○	○	○	○

We will randomly select 5% (1 in 20) participants such as yourself. If you are one of the selected participants, we will randomly choose one of the two pages, and carry out the lottery you have chosen on that page. You will then get the corresponding money amount *instead of* your other earnings from this study.

Hence, please choose on each page as if you were choosing a real lottery. You might be!

Choose the alternative you prefer most from amongst the six lotteries below

Get **Fr.14** for sure	Get **Fr.18 or Fr.12** with 50/50 chance	Get **Fr.22 or Fr.10** with 50/50 chance	Get **Fr.26 or Fr.8** with 50/50 chance	Get **Fr.30 or Fr.6** with 50/50 chance	Get **Fr.35 or Fr.1** with 50/50 chance
○	○	○	○	○	○

‹‹   ››

(next page)

Choose the alternative you prefer most from amongst the six lotteries below

Get **Fr.1.50 or Fr.52.50** with 50/50 chance	Get **Fr.9 or Fr.45** with 50/50 chance	Get **Fr.12 or Fr.39** with 50/50 chance	Get **Fr.15 or Fr.33** with 50/50 chance	Get **Fr.18 or Fr.27** with 50/50 chance	Get **Fr.21** for sure
○	○	○	○	○	○

‹‹   ››

On each of the next three pages, you'll consider a situation in which you need to decide how to act, but you are unsure whether and how your action will affect the outcome you're interested in. We are interested in how you typically approach such problems. There are no right or wrong answers. Please answer sincerely.

Suppose you're considering increasing your intake of foods that contain lycopene, whose health benefits have recently gained some attention (which you are not expected to know). Some experts suggest that lycopene can help reduce the risk of cardiovascular disease. However, other experts claim that the correlation between consuming lycopene and reduced cardiovascular risk is spurious, as individuals who consume more lycopene also tend to have healthier lifestyles overall. As a result, they claim, we see a relation between the consumption of lycopene-rich foods and reduced disease risk, even though lycopene does not have any health benefits. Suppose you have a mild dislike for foods containing lycopene.

	Strongly disagree	Weakly disagree	Weakly agree	Strongly agree
In situations like this, I tend to follow **whatever advice seems easiest**.	○	○	○	○
In this situation, I would check the facts to determine whether there really is a **correlation between lycopene consumption and cardiovascular risk**.	○	○	○	○
Unfortunately, in situations like this, unfortunately, advice that gives me **less hope** is more often the right advice to take, so I tend to follow that.	○	○	○	○
In situations like this, the **reality is rarely simple**. I am more likely to follow advice based on more comprehensive and elaborate theories.	○	○	○	○
In situations like this, I tend to follow advice that gives me **more hope** about my health outcomes.	○	○	○	○
In situations like this, I tend to follow the advice that **lets me do something** to improve things as opposed to keeping with the status quo.	○	○	○	○
In this situation, I would check the facts focusing on **determining which analyses are most methodologically sound** to ensure any correlations are not spurious.	○	○	○	○

In situations like this, **simpler explanations are usually better**. I am more likely to follow advice based on simpler and more straightforward theories.

○   ○   ○   ○

In this situation, I would look into the facts. I would **only consider people with an otherwise healthy lifestyle**, to determine whether a relation between the consumption of lycopene and disease risk is present for this population.

○   ○   ○   ○

Subjects also answer parallel questions for the following scenarios:

- *Suppose you're considering eating less food with additives E250 and E252, whose health implications have recently found some publicity (which you are not expected to know). Some experts suggest these additives can increase your risk of developing cancer. However, other experts claim that the correlation between consuming these additives and cancer is spurious, as some individuals both lead unhealthy lifestyles and frequently consume foods that contain these additives. As a result, they claim, we see a relation between the consumption of these additives and cancer risk, even though the additives themselves do not causally affect cancer risk. Suppose you generally like foods containing these additives and no good alternatives are available.*

- *Suppose you are managing a company that is experiencing a decline in sales. The sales team sees the reason for the decline in a lack of advertising efforts. It recommends spending more on advertising to increase brand awareness and attract new customers. However, the finance team argues that the decline in sales is due to both a regional economic downturn and demographic changes in the consumer base. By contrast, the finance team advises that investing more in advertising will not address the root cause of the problem and may worsen the financial situation of the company. Despite the conflicting opinions, the company must decide on the level of advertisement efforts.*

Questions about yourself and about your decisions

In this final part of the study, we will ask you some questions about yourself and about your decisions in this study. There will be 6 pages of questions.

Please answer the questions on this page as well as you can.

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (Enter a dollar amount with up to two digits after the comma, *without* the $-sign)

[   ]

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (Enter a whole number without any text or spaces)

[   ]

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how may days would it take for the patch to cover half of the lake? (Enter a whole number without any text or spaces)

[   ]

If you're running a race and you pass the person in second place, what place are you in?

○ First

○ Second

○ Third

A farmer had 15 sheep and all but 8 died. How many are left?

[▲▼]

Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?

[   ]

How many cubic meters of dirt are there in a hole that is 3m deep x 3m wide x 3m long? (Enter a whole number without any text or spaces)

[   ]

To what degree do you agree with the following statements?

	Strongly disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Strongly agree
Radiation derived from the use of a mobile phone increases the risk of a brain tumor	○	○	○	○	○
A positive and optimistic attitude towards life helps prevent cancer.	○	○	○	○	○
We can learn languages listening to audios while we are asleep.	○	○	○	○	○
Osteopathy is capable of causing the body to heal itself through the manipulation of muscles and bones.	○	○	○	○	○
The manipulation of energies bringing hands close to the patient can cure physical and psychological maladies.	○	○	○	○	○
Homeopathic remedies are effective as complements in the treatment of some diseases.	○	○	○	○	○
Stress is the principal cause of stomach ulcers.	○	○	○	○	○

	Strongly disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Strongly agree
Natural remedies, such as Bach flower remedies, help overcome emotional imbalances.	○	○	○	○	○
By means of superficial insertion of needles in specific parts of the body one can treat problems with pain.	○	○	○	○	○
Nutritional supplements like vitamins or minerals can improve the state of one's health and prevent diseases.	○	○	○	○	○

	Strongly disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Strongly agree
By means of hypnosis, it is possible to discover hidden childhood traumas.	○	○	○	○	○
One's personality can be evaluated by studying the form of their handwriting.	○	○	○	○	○
The application of magnetic fields on the body can be used to treat physical and emotional alterations.	○	○	○	○	○
Listening to classical music, such as Mozart, makes children more intelligent.	○	○	○	○	○
Our dreams can reflect unconscious desires.	○	○	○	○	○
Exposure to Wi-Fi signals can cause symptoms such as frequent headaches, problems sleeping, or tiredness.	○	○	○	○	○
The polygraph or lie detector is a valid method for detecting if someone is lying.	○	○	○	○	○
Diets or detox therapies are effective at eliminating toxic substances from the organism.	○	○	○	○	○
It is possible to control others' behavior by means of subliminal messages.	○	○	○	○	○

What is the following mathematical object usually called: $P(A \mid B)$? If you know it, please write it out below. Otherwise, enter "No"

Complete the following aphorism, if you know it: "Correlation does not ..."

(next page)

Have you ever taken a class about causal statistical inference?

○ Yes

○ No

In the context of causal statistical inference, what does the acronym "DAG" stand for? If you know it, please write it out below. Otherwise, enter "No."
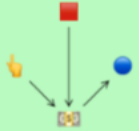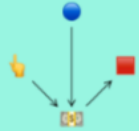
In the main part of the study, you made choices between advisors, such as in the following screenshot:

Please explain in your own words how you typically decided which advisor to choose. What numbers and data did you look at, if any? How did you resolve conflicting advice?

What is your gender?

- ◯ Female
- ◯ Male
- ◯ Transgender
- ◯ None of these

What is your age?

[▲▼]

Where are you from?

- ◯ Switzerland
- ◯ German-speaking Europe other than Switzerland
- ◯ Northern Europe (non-German speaking)
- ◯ Southern Europe
- ◯ East-Asia
- ◯ South, central, or north Asia
- ◯ Middle East
- ◯ Africa
- ◯ North America (USA or Canada)
- ◯ Latin America
- ◯ Australia
- ◯ Other (please indicate)

What is your native language?

- ○ German
- ○ French
- ○ Italian
- ○ English
- ○ Other

At which institution / faculty is your main field of study?

- ○ UZH Theological faculty
- ○ UZH Law
- ○ UZH Business, economics, and informatics
- ○ UZH Medicine
- ○ UZH Vetsuisse
- ○ UZH Philosophical faculty
- ○ UZH Mathematics and sciences
- ○ ETH Architecture and civil engineering
- ○ ETH Engineering sciences
- ○ ETH Natural sciences and mathematics
- ○ ETH Systems-oriented natural sciences
- ○ ETH Management and social sciences
- ○ ZHAW Applied Linguistics
- ○ ZHAW Applied Psychology
- ○ ZHAW Architecture Design, and Civil Engineering;
- ○ ZHAW Health
- ○ ZHAW Life Sciences und Facility Management;
- ○ ZHAW School of Engineering
- ○ ZHAW School of Management and Law
- ○ ZHAW Social Work
- ○ PH Zürich (Zurich University of Teacher Education)
- ○ ZHdK (Zurich University of the Arts)
- ○ Other

What degree level are you currently working towards

- ○ Bachelor
- ○ Master
- ○ Doctorate
- ○ Postdoc
- ○ I am not currently working towards a degree

What was your final grade in your Maturität in Mathematics?

- ○ 6
- ○ 5.5
- ○ 5
- ○ 4.5
- ○ 4
- ○ 3.5
- ○ 3
- ○ 2.5
- ○ 2
- ○ 1.5
- ○ 1
- ○ I do not have a Swiss high school degree Maturitat)

What was your final grade in your Maturität in your main language (German / French / Italian)?

- ○ 6
- ○ 5.5
- ○ 5
- ○ 4.5
- ○ 4
- ○ 3.5
- ○ 3
- ○ 2.5
- ○ 2
- ○ 1.5
- ○ 1
- ○ I do not have a Swiss high school degree Maturitat)

How much money do you spend per month, on average? (Including food, rent, clothing, entertainment.)

- ○ Fr. 500 or less
- ○ Between Fr. 500 and Fr. 1000
- ○ Between Fr. 1000 and Fr. 1500
- ○ Between Fr. 1500 and Fr. 2000
- ○ Between Fr. 2000 and Fr. 2500
- ○ Between Fr. 2500 and Fr. 3000
- ○ Between Fr. 3000 and Fr. 4000
- ○ Between Fr. 4000 and Fr. 5000
- ○ Between Fr. 5000 and Fr. 7500
- ○ Between Fr. 7500 and Fr. 10000
- ○ Fr. 10000 or more

Indicate your agreement with each of the following statements about political issues (such as immigration, unemployment, income inequality, social insurance, healthcare, etc.):

	Completely disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Completely agree
Most political issues are simple in principle. They have straightforward solutions.	◯	◯	◯	◯	◯
Most political issues are inherently complex. They do not have straightforward solutions.	◯	◯	◯	◯	◯

Are you eligible to vote in political elections in Switzerland?

◯ Yes

◯ No

◯ I don't know

Which Swiss political party is closest to your own views?
(Also answer if you are not eligible to vote in Switzerland)

◯ Swiss Party of Labour (PdA/PST-POP) -- left

◯ Social Democratic Party of Switzerland (SP/PS) -- left

◯ Swiss People's Party (SVP/UDC) -- right

◯ Green Liberal Party (GLP/PVL) -- center-left

◯ Christian Democratic People's Party (CVP/PDC) -- center

◯ Green Party of Switzerland (GPS/PES) -- left

◯ Conservative Democratic Party (BDP/PBD) -- center-right

◯ FDP.The Liberals (FDP/PLR) -- center-right

How strongly do you agree with the political party you are closest to?

◯ Agree with them on very few things

◯ Agree with them on some things

◯ Agree with them on most things

◯ Agree with them on nearly everything

How religious are you?

◯ I am not religious at all

◯ I am formally a member of a religion, but I do not practice it

◯ I am a member of a religion, and I practice it, but rarely

◯ I am a member of a religion, and I practice it regularly

◯ I am a member of a religion, and I practice it virtuously

This is the end of this study

Thank you for your participation!

Here is your payment, calculated exactly as stated in the instructions, determined by your choices and by luck:

Your earnings are determined by one of your decisions about which advisor to follow. From this, you earn Fr.23.80. In addition, you receive the completion payment of Fr.20. This yields a total pay of

Fr. 43.80

Please raise your hand. Somebody from the experiment team will come by and help you with your payment.

E.2 Experiment 3

Explanation correlational implications of causal structures in Treatment

Correlation and causation ✖

Here are some other things you may want to (but do not have to) consider if you choose to inspect the data:

Two quantities may appear related because the first causes the second.

Real world example
Rain and wet clothes occur together because rain causes wet clothes.

Experiment example
More ● and a higher bonus may occur together because ● cause the bonus
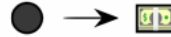
Two quantities may appear related because the second causes the first.

Real world example
Fire engines and forest fires occur together even though fire engines do not cause forest fires.

Experiment example
More ● and a higher bonus may occur together because the bonus causes ●.

Even though neither of two quantities may cause the other, they appear related if a third quantity simultaneously affects them both.

Real world example
In times when people wear sunglasses more often, they also tend to get more sunburns. This is not because sunglasses cause sunburns. Rather, sunnier weather causes more sunburns and also causes people to wear sunglasses. If we only consider periods with little sunshine (that is, if we hold the amount of sunshine constant), we no longer see a relationship between wearing sunglasses and sunburns.
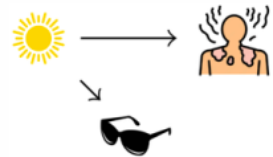
Experiment example
A higher bonus might go along with more ●. This might be because the action raises both the number of ● and the bonus, even though ● do not affect the bonus, and the bonus does not affect the number of ●. If we only consider cases in which the action was low (that is, if we hold the action constant), we no longer see a relationship between the bonus and ●.

If a quantity affects another only indirectly (i.e. through an intermediary variable), then the two quantities will appear unrelated once you hold the intermediary variable fixed.
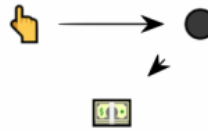
Real world example

More hours spent studying tends to improve understanding which, in turn, tends to improve exam performance. Hence, hours studying and eexam performance are related. But if somebody studies in a way that does not affect understanding, exam performance will be unchanged. That is, if we hold understanding fixed, then we will not see a relation between exam performance and hours spent studying.

Experiment example

A higher action may increase the number of ●, which, in turn, increase the bonus. Hence, action and bonus are related. Yet, if we hold the number of ● fixed, we do not see a relation between the action and the bonus.

When two unrelated quantities jointly affect a third, the two quantities will appear related once you hold the third quantity fixed.
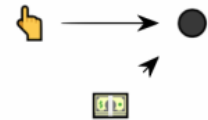
Real world example

A short commute and a nice office both tend to improve people's job satisfaction. Neither does altering somebody's commute affect how nice their office is, nor does changing the office interior affect the commute. Yet, if we do a survey *amongst people with low job satisfaction*, we'll find that those with nice offices have long commutes, and those with subpar offices have short commutes. But that's just because at least one thing must be making them unhappy.
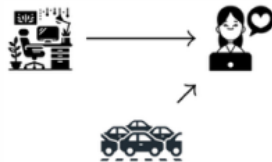
Experiment example

A higher action may increase the number of ●, and the bonus may increase the number of ●. Neither does the action affect the bonus nor does the bonus affect the action. Yet, they both affect the number of ●. If you consider *only cases where the number of ● is low*, you will see that higher actions go together with lower bonuses. But that's just because either the action or the bonus has to be low for the number of ● to be low.
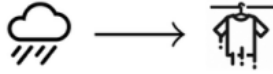
Explanation archetypical causal structures in Control

Examples of mechanisms in this experiment and in the real world

One quantity may directly cause another

Real world example

Rain may cause wet clothes. But wet clothes do not cause rain.

Experiment example

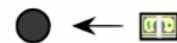● may affect the bonus, even though the bonus does not affect ●.

Real world example

Forest fires may cause the arrival of fire engines. But fire engines do not cause forest fires.

Experiment example

The bonus may affect ●, even though ● does not affect the bonus.

Two quantities that do not affect each other may simultaneously be caused by a third quantity.

Real world example

Sunshine affects people's tendency to wear sunglasses and also whether they get sunburns. But neither do sunglasses cause sunburns, nor do sunburns affect the tendency to wear sunglasses.

Experiment example

A higher action may raise both the number of ● and the bonus. But neither do ● affect the bonus, nor does the bonus affect ●

A quantity may affect another indirectly (i.e. through an intermediary variable).

Real world example

Hours spent studying for an exam tends to affect understanding which, in turn, tends to affect exam performance. But if somebody studies in a way that does not affect understanding (that is, if we hold understanding fixed), hours spent studying does not affect exam performance.
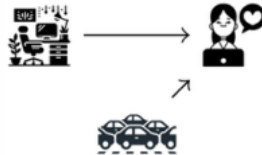
Experiment example

A higher action may change the number of ●, which, in turn, may change the bonus. Yet, if we hold the number of ● fixed, the action cannot affect the bonus.

Two quantities that do not affect each other may jointly affect a third quantity.

Real world example

The length of one's commute and how nice one's office is both affect job satisfaction. But neither does altering the length of commute affect how nice one's office is, nor does changing the office interior affect the commute time.

Experiment example

A higher action may affect the number of ●, and the bonus may also affect the number of ●. Neither does the action affect the bonus, nor does the bonus affect the action. Yet, they both affect the number of ●.

References

Gilboa, Itzhak and David Schmeidler, "Maxmin expected utility with non-unique prior," *Journal of mathematical economics*, 1989, *18* (2), 141–153.

Kruger, Justin and David Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.," *Journal of personality and social psychology*, 1999, *77* (6), 1121.

Pearl, Judea, *Causality*, Cambridge University Press, 2009.

Spiegler, Ran, "Behavioral implications of causal misperceptions," *Annual Review of Economics*, 2020, *12*, 81–106.