Extracting Models From Data Sets: An Experiment*

Guillaume R. Fréchette
NYU

Emanuel Vespa
UCSD

Sevgi Yuksel
NYU

October 28, 2025

Abstract

We experimentally study how individuals form mental models by learning from observational data. While participants are, on average, highly effective at extracting statistical regularities from a limited set of observations, there is substantial heterogeneity at the individual level. Participants presented with identical data sets consistently disagree on which variables are predictive of the outcome and how those variables relate to one another. Some systematically ignore relevant variables; others use irrelevant ones. We discuss the implications of this heterogeneity and provide evidence that these differences reflect distinct approaches to learning from data and can be interpreted as arising from cognitive frictions.

*We would like to thank numerous seminar participants for helpful comments. We thank Jay Cizeski, Jimena Galindo, Yiyuan Hu, Jie Liao, and Dingyue Liu for outstanding research assistance. This research was supported by the National Science Foundation under Grants 2315663, 2315664, 2315665, 2437062. We are responsible for all errors. Fréchette: gf35@nyu.edu. Vespa: evespa@ucsd.edu. Yuksel: sevgiy@gmail.com.

1 Introduction

People must adopt and rely on *models* (sometimes referred to as *narratives*) to make sense of the world, forecast future events, and assess the optimality of potential actions. Although a growing literature in economic theory studies the implications of adopting incorrect mental models, little is known about what types of mental models people form. In this paper, we experimentally study the mental models formed by learning from patterns in observational data. Specifically, we study the kinds of inferences individuals make regarding the statistical relationships among variables from examining data.[1]

Examples of decision makers relying on past observations to learn about the relationships between different variables abound. For instance, in the academic job market, our encounters with previous candidates inform our understanding of what observable attributes (e.g., job market paper and reference letters) predict future performance. Similarly, consumers draw on their experiences to understand how a product's reviews, packaging, and other visible attributes correlate with its quality. Non-professional investors look for patterns in past data to predict a company's value based on its observable characteristics. What is common to all these examples is that the decision maker learns from a limited set of observations without using sophisticated data-analysis techniques.

These examples raise fundamental questions: How effectively do people learn from data—that is, can they identify statistical patterns from a limited set of observations? To what extent are mistakes shaped by features of the data? Do mistakes reflect individual learning styles, with different people drawing conflicting conclusions from the same data? Do errors arise mainly from inferring patterns that aren't there, overlooking those that are, or both?

Our experiment is designed to isolate the core element that is common in the examples provided above: a decision maker forms inferences about the statistical relationships between different variables from a limited set of observations. Participants are presented with several different *data sets*. Each data set corresponds to a set of observations about the functioning of a hypothetical machine and consists of data describing the status of three binary variables: two lights and one sound.[2] Participants are given an opportunity to study these data sets, and take notes about them, with the understanding that at a later stage, they will be asked to make predictions about the same

[1] Throughout the paper, we use the term *mental model* to denote a description of how different variables statistically relate to each other. In discussing agents adopting these models, we do not imply they are consciously aware of these models; instead, we posit that their decision-making is consistent with the principles underlying these models.

[2] Our design builds on an experimental paradigm in psychology called *blicket machines* that has been widely used to study causal inference in children (Weisberg, Choi and Sobel, 2020).

machine (predict the sound using information on the status of the lights), by solely relying on the notes they have taken earlier, without any access to the original data set. Thus, our experiment produces two types of data: participants' notes and their predictions for each machine. The main object of interest is the participants' predictions, which reveal their understanding of the statistical relationships among the variables. We use these relationships to classify participants with respect to the different mental models that rationalize their predictions. We study the degree to which these models align or systematically deviate from optimal behavior. The notes serve a dual purpose: First, as an experimental tool, they compel participants to engage with the data and form an understanding of it. Second, they offer insights into how participants organize and learn (or fail to learn) from a set of observations.

The following key features of our experimental design are crucial for the interpretation of our results. First, our implementation is free of context. Outside of the laboratory, contextual cues may affect the inferences people make. However, our focus here is more foundational: examining how an agent's mental model is shaped solely by the patterns in a data set, free from the influence of confounding factors. Second, we give participants no information on the data-generating process; that is, we deliberately do not present participants with a set of possible models. Hence, there are no restrictions on the models that participants might consider. Our aim is to shed light on how people learn from a set of observations when they are not directed or suggested to *look* at the data in a specific way. Third, each participant in our experiment encounters multiple data sets, which cover a range of statistical correlations between the variables. This framework enables us to assess the extent to which mistakes are a function of the environment (aspects of the correlation structure) or an inherent characteristic of the participants (a person-specific trait).

A baseline finding is that, despite the abstract framing of the experiment and the minimal information provided about the data-generating process (DGP), participants, on average, use information from the data sets effectively to guide their predictions. In particular, the way participants condition their predictions on the status of the lights varies systematically across data sets, reflecting the underlying correlation structure. On average, participants achieve a prediction accuracy of 71%, well above the random benchmark of 50%, though still below the optimal benchmark of 81%.

The previous result masks substantial and persistent heterogeneity across participants, which is one of our main findings. Participants who learn (or fail to learn) effectively in one data set are likely to do so in others. For instance, while 26% of participants achieve prediction accuracy within three percentage points of the optimal benchmark across all data sets, 24% consistently fall short by 15 or more percentage points. Although some variation across participants is to be expected, the

magnitude and consistency of this heterogeneity—observed across nearly 300 predictions spanning 11 data sets—is striking. It suggests the existence of distinct participant types who systematically interpret and use data in different ways, irrespective of the underlying properties of the data sets.

A key implication of this result is that participants often *disagree*: they make opposing predictions 28% of the time despite being presented with identical information. Furthermore, a consequence of distinct participant types is that such disagreement is predictable. Specifically, we show that individuals who draw opposing conclusions about what constitutes optimal behavior when presented with the same data set tend to do so again when faced with a new data set. For any pair of participants, correlation in their disagreement rates for two randomly selected data sets is 0.34.

To understand the roots of this persistent heterogeneity, we next analyze (i) the types of mistakes participants make, (ii) how these mistakes depend on the statistical structure of the data sets, and (iii) how consistently such mistakes are repeated by the same participant. We then explore how these mistakes relate to the types of notes participants write for themselves.

As highlighted in the literature, decision makers may adopt mental models that are incorrect in different ways. For instance, Hanna et al. (2014) show that experienced Indonesian seaweed farmers fall short of the production frontier because they ignore a crucial variable—pod size—failing to recognize its importance in the production process. A different kind of mistake is illustrated by the informational redundancies problem in Akerlof and Shiller (2010): a realtor reads a newspaper predicting rising housing prices, then hears the same claim repeated by a colleague who read the same article. A naive realtor may treat the repeated claim as independent confirmation. In general, people may fail to identify relevant correlations or rely on ones that are irrelevant or nonexistent.

Both types of mistakes—ignoring a relevant variable and conditioning on an irrelevant one—are commonly observed in our data, occurring 24% and 20% of the time, respectively, in data sets where both mistakes are possible. Conditioning on an irrelevant variable is more likely (28%) when there is a spurious correlation between variables. Conversely, weakening the strength of the true correlation slightly increases the likelihood that relevant variables are missed (goes up to 26%). Analyzing behavior across multiple data sets for each individual, we find a strong tendency to repeat the same type of mistake. This reinforces our earlier findings on persistent heterogeneity across participants and provides a foundation for understanding disagreement—helping to explain why individuals consistently diverge in how they learn from and interpret data.

Next, we show that the mental models participants form are responsive to the strength of the available evidence. Even participants we classify as adopting incorrect mental models are 19

percentage points more likely to make optimal predictions in the cases that matter most—namely, for light configurations where the data most strongly support one prediction over the other. This finding suggests that some of the mistakes we observe may reflect attempts to simplify a complex learning task, with participants optimizing their behavior primarily for the most common cases.

We also demonstrate that the two types of errors discussed above—ignoring a relevant variable and conditioning on an irrelevant one—manifest in qualitatively distinct ways. Conditioning on an irrelevant variable can be interpreted as constrained optimal behavior. Among models that condition on all variables (including the irrelevant variable), subjects' behavior is typically consistent with the one that performs the best. In contrast, ignoring relevant variables is associated with stochastic behavior. These participants fail to recognize useful statistical associations between the observables (the lights) and the outcome (the sound). In many such cases, they overlook even the most salient regularity: the base rate favoring the sound being on. Thus, their behavior cannot be rationalized as constrained optimal and instead points to more fundamental limitations in extracting statistical patterns from data.

Finally, we analyze the notes participants write for themselves to offer direct evidence on their approach to learning from data. We document substantial variation in note-taking strategies and show that these strategies are closely linked to whether participants make optimal predictions or exhibit systematic errors. Specifically, we find that only about half of the participants study the data sets with the explicit goal of identifying statistical patterns they will later use in the prediction task by summarizing key statistical information. Failing to summarize key information at the note-taking stage makes it more difficult for participants to rely on such patterns at the prediction stage. This is true even of subjects who have the information in their notes, but have failed to summarize it along the dimensions that reveal these patterns.

2 Connections to the Literature

A growing literature in economic theory studies how incorrect mental models can influence an agent's beliefs and actions.[3] A central premise of this research is that the degree to which an agent adjusts her beliefs or actions in response to experience, or is influenced by others' arguments, is constrained by her initial, potentially incorrect mental model. As we discuss later, an emerging body

[3]For recent papers on learning under misspecifications, see Esponda and Pouzo (2016), Fudenberg et al. (2017), Bohren and Hauser (2021), Heidhues et al. (2018) and Samuelson and Steiner (2025). Spiegler (2020) provides a review of the approach to causal misconception, utilizing tools from computer science (e.g. Pearl (2009)).

of empirical research has begun to explore these insights. Our approach in this paper differs from the existing literature in that, rather than studying the consequences of specific misconceptions, we focus on understanding when and how these misconceptions arise in the first place. And, in particular, our focus is on documenting sources of possible heterogeneities in misconceptions (i.e., across different DGPs or idiosyncratic to individuals) with associated consequences (i.e., disagreement).

Several recent experimental studies document settings in which individuals may form an incorrect understandings of their environment. For example, participants often ignore described correlations when making decisions (e.g., Eyster and Weizsäcker (2010), Enke and Zimmermann (2019)), behave as if their vote matters even when it is not pivotal (e.g., Esponda and Vespa (2014, 2024)), or act in a second-price auction as if it were a first-price auction (e.g., Cason and Plott (2014); see also Martin and Muñoz-Rodriguez (2019)).[4] In these studies, participants form an incorrect understanding of their environment due to challenges in reasoning through the described structure.[5] By contrast, this paper focuses on misconceptions that stem solely from analyzing data, when no description of the DGP has been provided. In other words, our experiment contributes to this literature by providing insights on what types of mental representations arise when participants rely exclusively on data, without knowledge of the underlying rules or institutional details.

In light of these differences, we next provide an overview of the experimental literature on the topic. A common approach in recent work is to present participants with data sets together with an interpretation (e.g., a causal narrative) that rationalizes the data. This interpretation may either correspond to the true DGP or be incorrect and therefore misleading. Regardless of their accuracy, these interpretations may be effective in terms of influencing the agent's understanding of the data. Charles and Kendall (2024) demonstrate how giving the same data set to two groups, but providing each with a different narrative, can lead to different choices, in line with the predictions of Eliaz and Spiegler (2020). Their work also illustrates how some narratives can arise endogenously when participants are asked to provide advice to others and that there can be substantial heterogeneity in such home-grown narratives. Similarly, Barron and Fries (2024) develop a test of model

[4]Additional examples include participants not understanding equilibrium effects (Dal Bó et al. (2018)), having difficulties in learning from prices (Ngangoué and Weizsäcker (2021)), not understanding adverse selection (Charness and Levin (2009), Martínez-Marquina et al. (2019), Ali et al. (2021)), or difficulties in understanding sample selection (Esponda and Vespa (2018), Araujo et al. (2021), Enke (2020)).

[5]Evidence also suggests providing participants with corrective data or experience may not always be effective in correcting such misconceptions. For instance, Esponda et al. (2024) demonstrate that many individuals who misunderstand aspects of Bayesian updating fail to learn from informative data, even after substantial exposure. Similar challenges are observed in correcting for sample selection biases (Esponda and Vespa, 2018) and in addressing incorrect beliefs (Fudenberg and Vespa, 2019).

persuasion within the framework of Schwartzstein and Sunderam (2021), showing that narratives are particularly effective in changing beliefs when they align more closely with the data. Finally, Ambuehl and Thysen (2024) investigate how individuals choose between different causal models. Specifically, participants are *simultaneously* presented with a data set and at least two competing models.[6] Their findings reveal behavioral heterogeneity: some participants are drawn to models promising favorable outcomes, others take a more cautious approach, and some evaluate models based on their fit with the data.

A few recent papers study how individuals update their beliefs in the presence of competing data-generating processes, or models, that could explain observed data. Aina and Schneider (2025) show that while some participants account for model uncertainty by putting positive weights on more than one model in updating their beliefs, many ignore model uncertainty altogether and use an updating rule that fully focuses on the model that best fits the data. A similar finding is reported in Musolff and Zimmermann (2025), where participants form estimates ignoring model uncertainty, particularly when the competing models are complex. They also find, however, that when directly asked, participants beliefs do account for model uncertainty. Augenblick et al. (2025) provides a theory in which participants simplify a complex prediction problem by *focusing* on a single model, neglecting uncertainty across alternative models. The theory connects neglect of model uncertainty to overprecision and disagreement. Data from a new laboratory experiment and data from the Survey of Professional Forecasters provide evidence for how these measures are linked, in line with their theoretical predictions.

The previous papers primarily focus on how an exogenously provided model or a set of alternative models can influence a person's interpretation of data and consequently their decision making. In contrast, our goal is to study how participants learn from a data set when no information about the DGP is provided, namely they have not been presented with a set of models. Two main aspects of our experiment are key to achieving this goal.

First, we investigate paradigmatic correlation structures. As Pearl and Mackenzie (2018) notes, three-node networks (i.e., involving three variables) with two links are the building blocks of Bayesian networks. We study the DGPs that arise from these structures; details appear in the next section.[7] Hence, our design lets us study difficulties that arise in any setting with more than

[6]Charles and Kendall (2024) also have participants face multiple narratives, but a second narrative is presented after participants have made decisions knowing the first narrative.

[7]Specifically, Pearl and Mackenzie (2018) (page 113) claims "The next step after a two-node network with one link is, of course, a three-node network with two links, which I will call a 'junction.' These are the building blocks of all Bayesian networks (and causal networks as well)." Subsequently, they define three types of three-node, two-edge

two variables, namely by opening up to the possibility that participants form models that neglect an existing correlation and/or that rely on a nonexistent one. More broadly, the rich set of DGPs will let us study the extent to which heterogeneity in behavior is driven by the structure of the DGP. In addition, because participants make many predictions (at least 54 in each) across multiple distinct structures (five), we can also assess how much variation in behavior is attributable to structural versus person-driven components, clarifying when disagreement is likely to arise and how predictable it is. Second, our design decouples learning from prediction: participants first learn from data, then make predictions for specific instances. This separation encourages them to articulate an internal (mental) model of the data before engaging in predictions.

Three papers are more closely connected in that they do provide data but do not present subjects with models. Kendall and Oprea (2024) examines how participants make inferences about DGPs that are generated using finite automata with two underlying binary variables (an input and output) that can be correlated in time. Specifically, participants observe an input and output one at a time for 12 periods and in each of the following 12 periods they guess the output after being shown the input. They document that many participants have difficulties inferring the underlying DGP, suggesting that serial correlations can be challenging even in an environment with two binary variables. A simpler setting with one binary variable that may be correlated in time is studied in Payzan-LeNestour et al. (2025), where participants forecast the return for an asset over 200 periods. Their aim is to assess the extent to which participants can distinguish features of the underlying DGP; specifically, permanent from transitory shocks. They find that a majority of their participants successfully achieved the task. Relatedly, Grass et al. (2025) design an experiment in which data is also gradually revealed and find that people's models tend to be sticky in the sense that final models remain strongly influenced by earlier models formed using a subset of variables.[8]

Beyond economics, a substantial literature in cognitive psychology addresses related questions, which we review in Appendix A. The appendix organizes this literature in three areas, though, naturally, there are overlaps. A first area focuses on providing participants with summary tables of data (typically a two-by-two table with frequencies of two binary variables), and assessing whether participants infer a correlation. One crucial finding in this literature is that in environments with

networks: chain, collider, and fork. The DGPs in our Design Section instantiate Pearl and MacKenzies examples: the one-link structure appears in their Fig. 6.2; the common-consequence (Berksons paradox) in Fig. 6.3; and the full structure appears in multiple places (e.g., Simpsons paradox and the Berkeley admissions paradox).

[8]There is also recent work eliciting narratives—causal models—from people about real-world phenomena. This covers, for instance, eliciting narratives that regular people and experts use to explain macroeconomic phenomena (e.g., Andre et al. (2021)) or narratives used by managers to understand their industry (e.g., Han et al. (2025).)

context, priors play a major role. This work partly informs our choice of focusing on an abstract setting.[9] A second area in psychology focuses on cue-effect environments, where, for instance, a classic reference is Pavlov's dogs (Pavlov, 1927). This literature on associative learning studies perceptions of correlations, typically by providing participants information on trials one at a time. One key finding is that the closeness (in terms of time) between cue and effect is crucial for making associations. To abstract from time playing a role, in our setting participants see all the data at once. Finally, a third group of papers focuses on causality. A prominent example is Steyvers et al. (2003), where participants are provided with a data set and two possible DGPs, and their task is to indicate which of the two DGPs is most likely to have generated the data. This portion of the psychology literature is closer to papers in economics that provide narratives in addition to data. Instead, as we highlighted earlier, our focus is on assessing how people deal with correlations when we do not provide them with any model as a guideline.

3 Experimental Design

We first provide an overview of the experimental design. Then, we describe the experiment in further detail from the perspective of what was presented to the participants. Other aspects of the experimental design are described later in this section.

3.1 Task

Each session consists of two parts. In Part 1, participants are presented with 11 different *data sets*, one at a time. After seeing all data sets, participants proceed to Part 2, where they are asked to make predictions for each data set they observed in Part 1. Two important features link Part 1 to Part 2: (i) The Part 1 data set is informative about predictions in Part 2 but is *not available* to the participants during this second part; and (ii) when participants are presented with data sets in Part 1, they can type *notes-to-self* on the computer terminal and these notes *will be available* to them during the relevant prediction tasks in Part 2.

The framing used in the instructions is that each of the 11 data sets is generated by a different

[9] In addition, some papers let participants control the production of the data sets and it is well documented that this feature can affect the participant's understanding of correlations. For example, there is a connection to the literature on illusion of control. For a recent paper in economics that relates illusion of control and mental models, see Fan (2024).

machine. A machine consists of lights of different colors and can make a sound.[10] Each data set consists of 27 trials, where each trial records an occurrence of the machine's operation. For each trial, the record shows the status of two lights—red and blue—whether they were on or off, and whether the machine made a sound.[11] In the paper, we use variables R, B, and S to refer to the red light, the blue light, and the sound, respectively, and denote their status by assigning values 0 or 1 to these variables. The instructions (reproduced in Online Appendix H) are carefully worded to avoid a direct suggestion of a causal relationship between the lights and the sound.[12]

An example of a data set from Part 1 is shown in Figure 1. The 27 trials are presented on the right side of the screen, all at once, where each trial is a row.[13] On the left side of the screen, the participant can take notes. These notes are made available to them in Part 2 when they make predictions. In Part 1, the participant sees 11 screens such as this one, one for each data set.[14]

When participants face Part 1, they also know the prediction task that they will face in Part 2. In general, subjects know they will be asked to make a prediction on whether the machine makes a sound conditional on the status of some of the other variables, namely, the lights. As a reference, we present an example in Figure 2. The participants can see their notes and the status of the red and blue lights. They are asked to guess whether the machine makes a sound. The order of presentation for the machines, the trials within each data set, as well as the prediction tasks are randomized across subjects.[15] At the end of the session, one prediction task is selected at random, and if their prediction is correct, they receive $25, in addition to the $10 show-up fee.

3.2 Within-Subjects Treatments: 11 data sets

In this section, we describe in three stages how we generated the 11 data sets presented to the participants. First, we present a set of directed acyclic graphs (DAGs), which formed the basis

[10] As discussed in the literature review, this design is an adaptation of the so called *blicket machines*, a device used in experiments with children (Gopnik et al., 2004) to study causal inference.

[11] Although the status of only two lights (red and blue) is reported, as a way to motivate the potential probabilistic nature of how different variables (the lights and the sound) are related, participants are told the machine may include other lights, the status of which they will not be informed about.

[12] The instructions include the line "The lights and the sounds may or may not be related to each other."

[13] We used 27 trials because we could fit 27 rows at most on the computer screens used in the laboratory with no need to scroll up or down. See Section 3.3 for more on this.

[14] Once participants have taken notes for a data set and they move on to the next one, they cannot return to the previous ones.

[15] The last nine prediction tasks in treatment Unspecified, discussed later, are an exception.

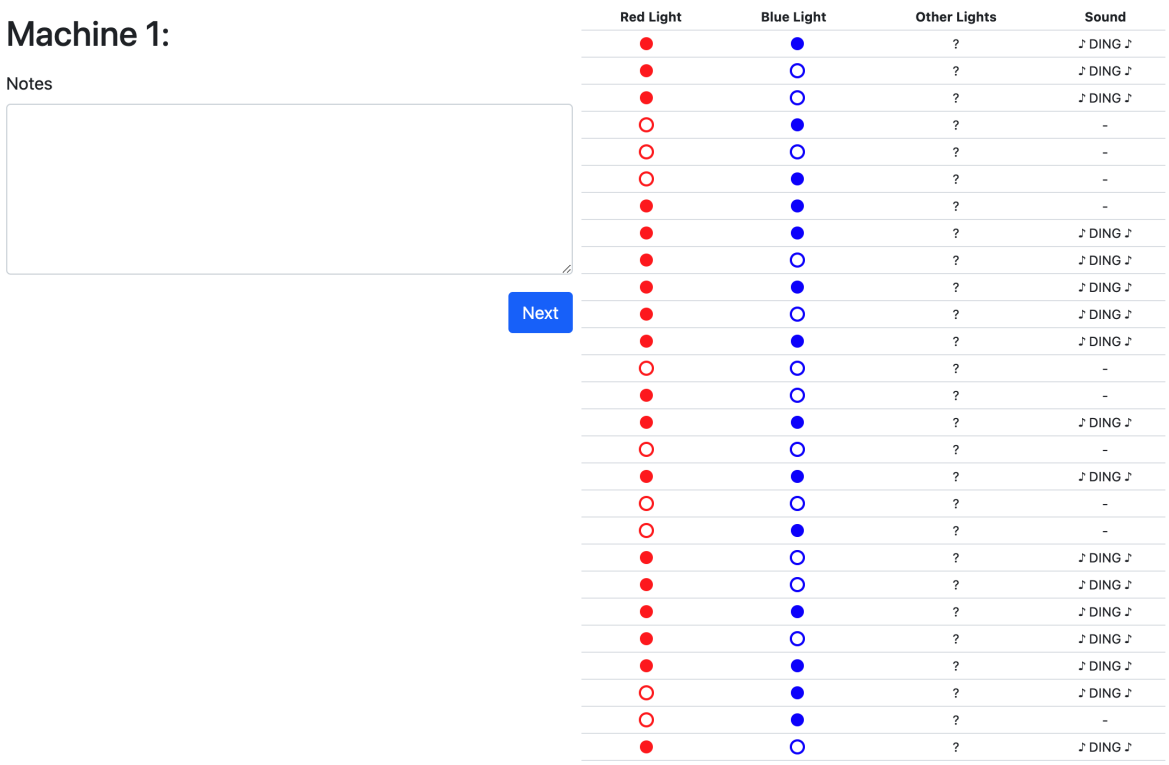


Figure 1: Screenshot of Part 1

Notes: The participants can leave *notes-to-self* after inspecting the data set presented on the right side of the screen. Each trial is a row in the right-hand-side table. A light that is on (off) is represented with a full (hollow) circle. The sound is captured with "Ding," and no sound, with a dash (-).

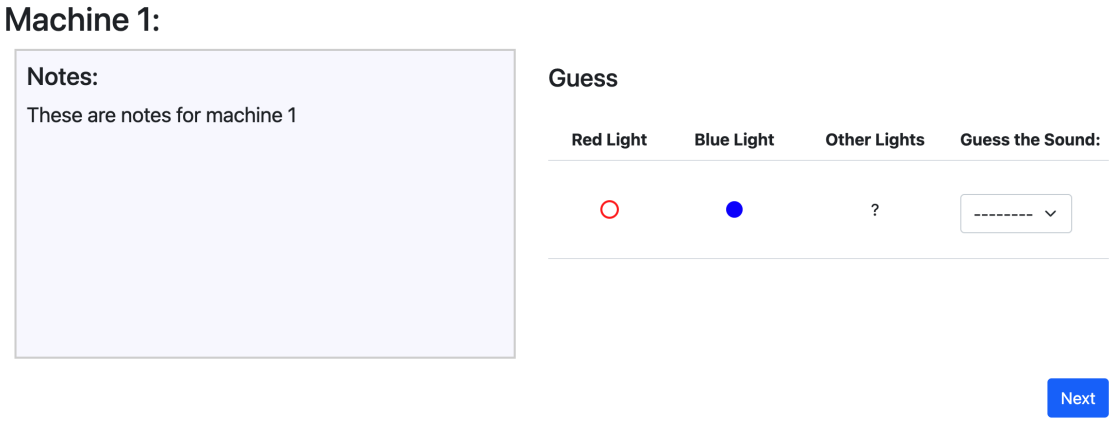
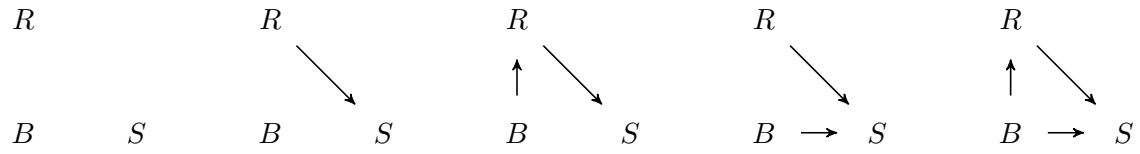


Figure 2: Screenshot of Part 2

Notes: The participant can read the *notes-to-self* they wrote in Part 1. In this example, they are provided with the status of the red and blue lights—hollow (full) means light off (on)—and a drop-down menu lets them guess whether the machine makes a sound.



No Correlation	**One Link**	**Chain**	**Common Conseq.**	**Full**
$(B \perp S, R \perp S, B \perp R)$	$(B \perp S, B \perp R)$	$(B \perp S \mid R)$	$(B \perp R)$	
Cor. w/ S: none	Cor. w/ S: R only	Cor. w/ S: R and B	Cor. w/ S: R and B	Cor. w/ S: R and B
To predict S: none	To predict S: R only	To predict S: R only	To predict S: R and B	To predict S: R and B

Figure 3: Directed Acyclic Graphs (DAGs)

Notes: R, B, and S are binary variables that capture the status of the red light, the blue light, and the sound, respectively. The independence conditions implied by the DAG are reported in parentheses. For instance, in No Correlation, all variables are independent from others, and in Chain, B is independent of S conditional on R. Below the independence relations, all variables correlated with S are reported. Finally, the last row reports the variable(s) one would optimally condition on to predict S when both R and B are observed.

for how we created the data sets.[16] Second, we outline the process by which we chose 11 distinct sets of parameters—each referred to as a *parametrization*—associated with these DAGs. Third, we describe how these parametrizations were transformed into data sets. This process generates variation in data sets with the aim of understanding how participants' behavior—namely, their ability to extract information from a set of observations—changes with the statistical relationship between variables, the amount of noise, and the complexity of the data set.[17]

Selected DAGs

We focus on five DAGs involving R, B, and S. These DAGs and their key characteristics are represented in Figure 3.[18]

In *One Link*, R and S are the only variables that are correlated. As a consequence, the optimal prediction for S conditions on R.[19] *Chain* is closely related to *One Link*, with the difference being the additional link between B and R. Here, not only R but also B is correlated with S. Notice, however, that conditional on R, B and S are independent. For this reason, when the status of both

[16] DAGs are often used to describe causal relationships between variables. Because participants simply face a prediction task, causality does not play a role in our experiment. We use the DAGs as an easy way to summarize possible correlation structures.

[17] More detailed descriptions of each data set are provided in Online Appendix B.

[18] To ease presentation, we fix colors so that, for instance, B is independent from S in One Link. In practice, however, the design varied the meaning of B and R.

[19] To ensure this, we choose parametrizations where $p(S = 1 \mid R = 1) > 0.5$ and $p(S = 1 \mid R = 0) < 0.5$.

lights are known, an optimal prediction for S only conditions on R. However, if the status of the red light is unknown, one should exploit the correlation between the blue light and the sound by conditioning on the status of the blue light to predict S for *Chain* (but not for *One Link* where there is no such correlation). *One Link* and *Chain* therefore provide an interesting comparison, which we come back to in later sections of the paper.

In *Common Consequence*, the outcome S can be interpreted as the common consequence of two variables, R and B. Both R and B are correlated with S, and the optimal prediction rule for S conditions on both variables. In fact, we consider two versions of common consequence, depending on whether R and B jointly generate the effect on S (AND condition), or independently generate the effect (OR condition). *Full* changes the statistical relationship between the variables relative to *Common Consequence (OR)* in a similar manner to how *Chain* compares with *One Link*, that is, by adding a link between B and R. Here, the optimal prediction for S conditions on both lights, the same as with *Common Consequence*.

As a benchmark, we also include a DAG, *No Correlation*, in which all variables are independent, where the optimal prediction for S does not condition on R or B. Note that as we move from *No Correlation* to *Full*, the statistical relationship between the variables becomes increasingly complex. Focusing on the number of variables that are correlated with S, variation from none (*No Correlation*) to one (*One Link*) and two (*Chain*, *Common Consequence* and *Full*) exists. Focusing on the number of variables that is optimal to condition on to predict S, variation from none (*No Correlation*) to one (*One Link* and *Chain*) and two (*Common Consequence* and *Full*) exists.

From a DAG to *Parametrization*

There are many different joint probability distributions over the three variables (R, B, and S) consistent with each DAG.[20] We pick 11 *parametrizations* based on the DAGs discussed above: one parameterization for *No Correlation*, and two parameterizations for *One Link*, *Chain*, each version of *Common Consequence* (AND and OR), and *Full*. In DAGs with two parametrizations, we manipulate the strength of the statistical relation between the different variables. In the *Low noise* parameterization, if an arrow goes from one variable to another, the probability that the realization of the latter one matches the former is equal to 90 percent. In the *High noise* parameterization,

[20]For instance, consider *One Link* in Figure 3. The DAG does not specify the likelihood of $B = 1$ or $S = 1$; it simply imposes that these events are independent. Similarly, how exactly R is correlated with S is also not pinned down beyond the requirement that these variables are not independent.

the corresponding probability is only 80 percent.[21]

Online Appendix B covers the specific way in which we chose the 11 parametrizations. Here, we describe the criteria that underlie these choices. First, the probability that $S = 1$ is set to approximately 0.62 in all 11 cases. This implies that a prediction rule for S that does not condition on the lights can achieve an accuracy of approximately only 62 percent in all cases, which is better than random, while still generating sufficient incentives to use the observables (the status of the lights) to predict the sound. Our second criterion aims to improve identifiability of whether a participant is using the optimal prediction rule for a given parametrization. As a reference, consider *One Link*: in both the low- and high-noise conditions, if the participant is provided with the status of the lights and is asked to predict the sound, the optimal strategy is to guess the machine will make a sound only when the red light is on—a *deterministic prediction rule* we refer to as *G w/ R*, for "Guess when $R = 1$." Such a rule achieves a prediction accuracy of 90 or 80 percent, depending on the noise condition (as shown in Table 6 of Online Appendix C). Using an alternative prediction rule such as guessing the sound is on whenever the red light or the blue lights are on (referred to as *G w/ R or B*) lowers prediction accuracy to 76 and 70 percent in the low- and high-noise conditions, respectively. We chose the parameters for each case (subject to the constraints stated above) to increase as much as possible the prediction accuracy cost associated with using a suboptimal prediction rule.

To summarize, we chose the 11 parameterizations to (i) keep the likelihood of $S = 1$ around 62 percent; (ii) satisfy the different noise conditions; and (iii) increase the cost of using a suboptimal prediction rule.

From Parametrization to data set

Parameterizing the DAGs moves us a step closer to creating the data sets that were presented to the participants. Specifically, for each of the 11 parameterizations, we can compute the joint distribution over realizations of the lights and the sound. If participants were to see a large data set, the frequency of observing different trials in this data set (corresponding to different light and sound configurations) would closely match this distribution. However, to make the task of learning from data manageable for subjects, we limited the size of the data set shown to the participants.[22]

[21] When two arrows point toward the same variable (as in Common Consequence and Full DAGs), we assume errors are independent in the OR condition, but correlated in the AND conditions.

[22] The design deliberately avoids providing subjects with "data summaries" that report relative frequency of different trials, because such an approach would already suggest a specific way of learning from data. See Esponda et al. (2024) for how providing summarized data can significantly alter learning outcomes.

For each of the 11 parameterizations, we created a 27-trial data set where empirical frequencies over different trials are closest to the true probability distribution.[23] All participants are presented with the same 11 data sets. Thus, our experimental design generates two main sources of variation: (i) for a fixed subject, variation across different data sets, and (ii) for a fixed data set, variation across subjects.[24]

Between-Subjects Treatments

In addition to the within-subject variation in terms of the data sets that participants face, we conducted two manipulations that we implemented across subjects. In the *Specified Prediction* treatment, we informed participants that Part 2 predictions consist of guessing whether the machine will make a sound after they are shown the status of the red and blue lights. We have two implementations of this treatment. In the first implementation, we constrained notes to 280 characters. In the second implementation, we constrained notes to 75 characters.[25]

In the *Unspecified Prediction* treatment, participants were told that Part 2 predictions may consist of (1) predicting the sound after learning the status of both, one, or none of the lights; or (2) predicting one light after observing either the other light, the sound, both, or none. Notes were constrained to 75 characters. This treatment allows us to add additional prediction questions where they have either partial or even no information on the status of the light.

Part 2 Predictions

In the Specified Prediction treatment, participants faced 27 prediction rounds associated with each machine in Part 2.[26] For each machine, prediction rounds were presented one by one and in random order. Observing many predictions for each machine is useful to identify the prediction rules being

[23] Randomly drawing 27 trials independently for each subject using the true probability distribution would have created significant variation across subjects in the data sets presented to them. Although this approach would have generated rich data in other ways, it would have made it difficult (i) to control incentives and (ii) to identify whether subjects are using the optimal prediction rule.

[24] We also observe for a fixed subject and data set, variation in predictions across different light configurations.

[25] With the initial character limit of 280 (which used to be the limit on Twitter), a non-negligible proportion of subjects were transcribing the entire data set directly in the notes. To identify the extent to which results depended on such behavior, we conducted a second implementation that constrains note length for the median participant from the first implementation.

[26] Participants were asked to predict the 27 trials that they observed in the Part 1 data set. However, this was not revealed to the participants. In the instructions, they were only told that in Part 2 they would make several predictions for each machine.

used, because we can observe participants making predictions for the same light configurations a number of times.

In the Unspecified Prediction treatment, just as in the Specified Prediction treatment, participants made the same 27 predictions. In addition, they made nine more predictions, which were presented after the initial 27. As in earlier rounds, these predictions were also about the sound, but participants were provided with partial or no information on the status of the lights.[27]

Implementation Details

We conducted four sessions of each treatment at the EconLab in UC San Diego, where participants are students at the university. The instructions, which are available in Online Appendix H, describe the Part 2 prediction task but do not specify the number of predictions that participants will make per machine. After reading instructions, participants could move at their own pace but were informed they would not be able to leave early. A total of 88, 72, and 70 participants are in Specified Prediction with the 280-character note limit, Specified Prediction with the 75-character note limit, and Unspecified Prediction with the 75-character note limit, respectively.

3.3 Discussion of the Design

We designed the experiment with the following goals in mind.

First, we adopt a context-free implementation by framing datasets as corresponding to abstract machines. Although this approach abstracts away from other important forces that could influence an agent's understanding of their environment, it allows us to study how an agent's mental model is shaped by the patterns they identify in observations in the absence of any confounds. Driven by similar considerations, we do not focus on the dynamics of learning. For this purpose, we provide data sets in their entirety all at once to our participants. Many interesting questions could be asked about how context or past experiences within the same environment impact learning, which we hope can be tackled in future work.

Second, we provide no information about the data-generating process. Specifically, participants are not given a set of models or a prior over these models. This unstructured design approach places no restrictions on the range of possible models participants may consider. For example, a

[27] Specifically, there were four rounds where only the status of the red lights was revealed (two with $R = 1$ and two with $R = 0$), four rounds where only the status of the blue lights was revealed (two with $B = 1$ and two with $B = 0$) and one round where no information on the lights was provided.

participant might speculate about whether serial correlation exists between different observations or whether one data set informs another. Additionally, because each data set is presented all at once, no information is available on the timing of events (e.g., the sequence of lights and sound), which is known to affect causal inference (see, e.g., Bramley et al. (2018)). Our aim is to shed light on how people learn from a set of observations when they are not directed or suggested to *look* at the data in a specific way. Motivated by similar reasons, we never ask participants directly about their "models," as this line of questioning might influence how they learn from their observations. Instead, we study how people make predictions, an exercise that has a natural counterpart in many real-world applications. These predictions "reveal" participants' mental models, that is, how predictions condition on different observable variables reflect the inferences they made regarding the statistical relationship between variables.

Finally, each participant in our experiment encounters 11 data sets, which cover a range of possibilities with respect to the statistical correlations between the variables. This approach enables us to examine whether a participant's ability (or inability) to adopt optimal models remains consistent across different environments, or if it is shaped by the features of the environment.

4 Preliminary Findings

In this section, we present initial findings that provide background for the main results discussed in the next section. We show that, on average, participants learn from the data sets provided to them: their predictions systematically vary with both the observables (i.e., the light configurations) and the underlying statistical structure of the data set. However, there is substantial heterogeneity, and participants exhibit systematic deviations from optimal behavior in all data sets.

Predictions: Aggregate Response to the Presence of Correlations and Noise

As described in Section 3, participants in our experiment are not given any information about the data-generating process associated with each machine but can learn about it by studying the data sets provided to them. Since this is the only information participants can rely on, we use the correlation structure represented in each data set as a benchmark and define the accuracy and optimality of predictions with respect to this benchmark.[28] More precisely, prediction accuracy is defined as the probability that a participant's guess is correct for a randomly drawn observation

[28] As explained earlier, these data sets were deliberately selected such that the correlation structures correspond to those presented as DAGs in Figure 3.

from the corresponding data set. For example, consider a participant who guessed that the sound is on when both lights are on. Suppose that, in the relevant data set, there are ten such observations, and in eight of them the sound is indeed on. The participant's prediction accuracy for that guess is defined to be 80%. Optimality, by contrast, is defined as a binary measure indicating whether the prediction maximizes accuracy. In the example above, the participant's guess would be coded as optimal (value 1), since the most frequent outcome for that light configuration is that the sound is on.[29]

Figure 4 plots the cumulative distribution of these measures computed at the participant level for different data sets separated into three categories: the No Correlation data set, the High Noise data sets, and the Low Noise data sets. The vertical lines in the left panel depict the level of prediction accuracy that could have been achieved by an agent who always makes optimal predictions in the three corresponding categories. As expected, this benchmark depends on how informative R and B are about S, and hence is lower in the High Noise data sets than in the Low Noise data sets, but lowest in the No Correlation data set where S is independent of R and B. In the right panel focusing on prediction optimality, by definition, the best achievable value is 1 for all three categories.

A first observation when focusing on the left panel is that in all three categories, the vast majority of participants make predictions that are more accurate than random, which corresponds to a prediction accuracy of 0.5. This finding shows participants are able to leverage information in the data set to improve their prediction accuracy. In No Correlation, less than 20% of participants do worse than random. Because in this data set the lights carry no information, performing better than random indicates that most participants do use information on the frequency of $S = 1$, the only feature of the data set with value for predictions. In fact, approximately 20% of participants are exactly at the optimal prediction accuracy in this data set.

If participants ignore the lights and focus solely on the frequency of $S = 1$, they would achieve similar levels of accuracy across all datasets, as this frequency is held roughly constant at around 62% by design. Consequently, changes in prediction accuracy when comparing the Low or High Noise data sets with No Correlation indicate that participants are also using information on how the lights correlate with the sound to make their predictions in these data sets. Indeed, we observe that the distributions of prediction accuracy sharply improve relative to No Correlation. In fact, we see a clear ordering depending on the amount of information in the data set. Prediction accuracy

[29] Guessing that the sound is off would be coded as suboptimal (value 0). If the frequency with which the sound is on is exactly 50% for a given light configuration, any guess is considered optimal. In Section 5.3 we relax these definitions and examine the extent to which each participant's behavior can be rationalized by a flexible prior.

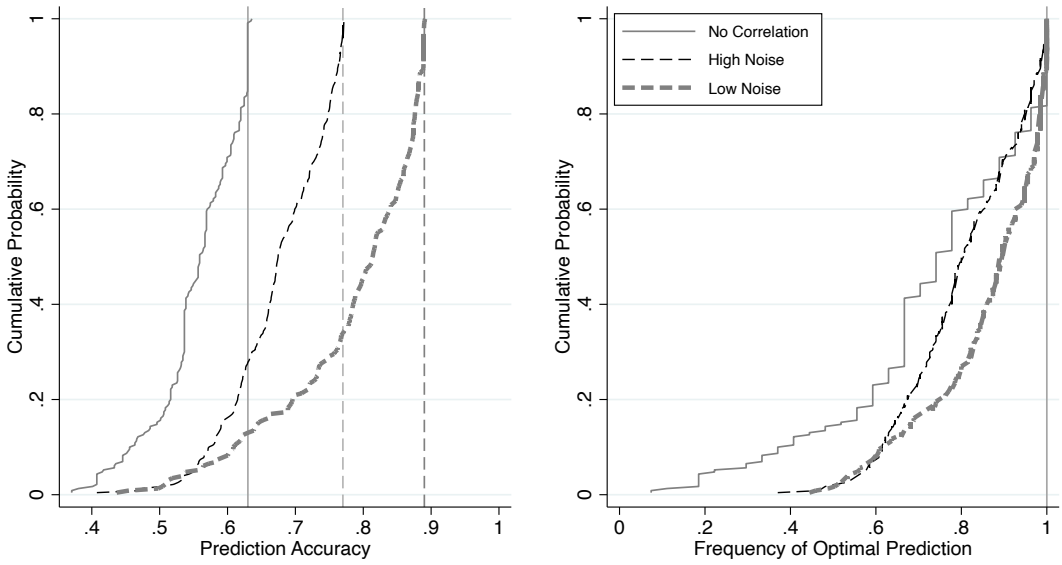


Figure 4: Distributions of Prediction Accuracy and Optimality

Notes: In the left panel, each observation corresponds to the expected prediction accuracy of a subject in the data sets corresponding to the specific category, and the vertical lines denote the prediction accuracy for an agent who guesses optimally in the three respective categories. In the right panel, each observation corresponds to the frequency with which a subject's guesses are optimal given the available information in the data sets corresponding to the specific category, and hence, the best achievable value is 1 for all three categories.

is lowest in the No Correlation dataset, higher in the High Noise dataset, and highest in the Low Noise dataset. Interestingly, as seen in the right panel of Figure 4, this ordering is preserved with respect to optimality of predictions.[30]

Overall, these patterns suggest that most subjects are engaged, are extracting payoff-relevant information from data sets, and perform better when statistical patterns are easier to discern.

Limited Between-Subject Variation

Figures 13 and 14 in Online Appendix D display cumulative distributions of prediction accuracy and optimality across the three between-subjects treatments. These treatments—restricting the number of characters in the note-to-self and broadening the prediction task—do not lead to significant differences in participant behavior.[31] We find no evidence that either manipulation affected

[30]Equality of the distributions can be rejected by a Kolmogorov-Smirnov test (p value < 0.001 in all pairwise comparisons). Note that the share of participants whose predictions are optimal more than 80% of the time increases across these categories from 40% in No Correlation to 52% in High Noise, to 74% in Low Noise.

[31]Tables 10 and 11 in Online Appendix F further demonstrate limited treatment effects with respect to the different type of mistakes that will be discussed in future sections.

prediction accuracy or optimality. Accordingly, we pool data from all treatments in the main analysis.

The primary motivation for the *Unspecified Prediction* treatment was to observe how predictions change when only partial information about the lights is available in additional prediction tasks. We make use of this later in Section 5.2.2 to examine participants' understanding of *conditional independence*.

How Do Predictions Change as a Function of Observables?

Here, we examine how participants' predictions vary across light configurations and whether these patterns align with the underlying statistical relationships in each dataset. To do so, we begin by aggregating over the Low and High Noise conditions and describe how predictions differ across light configurations for each DAG.[32]

Each graph of Figure 5 presents the aggregate frequency of predicting $S = 1$ for each possible light configuration (R, B).[33] Black dots denote optimal behavior for each case. As can be seen in all cases, a clear pattern in the direction of optimality always exists. For example, in One Link and Chain, the likelihood of predicting the sound to be on changes by 50 percentage points depending on the status of the red light, but the status of the blue light has a much smaller impact. In Common Consequence (both OR and AND) and Full, as predicted, the status of the blue light also has a large impact on predictions.

So far, these findings provide aggregate-level evidence that many participants can extract information from the data sets and condition their predictions on the variables they observe (i.e., the status of the lights). In fact, the conditioning varies with the structure of the data set in the direction of optimal behavior.

Nonetheless, we also see that, at the DAG level, predictions deviate from optimality in revealing ways. For example, in Chain, when only the blue light is on, 28% of the participants suboptimally predict the sound to be on, suggesting predictions vary (suboptimally) with the status of the blue light. Such mistakes are not unique to Chain. Regardless of the structure of data set, in all cases, in at least one light configuration, predictions are suboptimal more than 20% of the time.

[32]Technically, Common Consequence [AND] and [OR] are not distinct DAGs, but they differ in terms of the optimal prediction rule for S. Thus, we separate out these cases when presenting results at the DAG level. Results are further separated by noise levels in Figure 15 of Online Appendix D.

[33]The p-values in the top-left corner of each graph report results from F-tests on whether the frequency of guessing the sound changes with the status of the blue and red lights (see notes to Figure 5 for details).

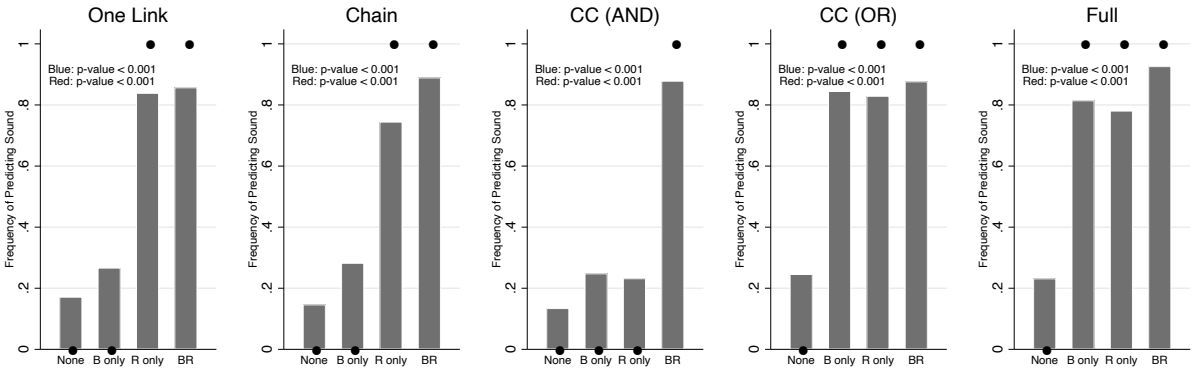


Figure 5: Predictions by Light Configuration

Notes: *None* refers to trials where both the red and blue lights are off; *B only* (*R only*) refers to trials where only the blue (red) light is on; *BR* refers to trials where both lights are on. Black dots denote optimal behavior. Reported p-values refer to F-tests on whether prediction frequency changes with the status of the blue and red lights. For blue (red), we test jointly whether the prediction frequency changes from *None* to *B only* and from *R only* to *BR* (from *None* to *R only* and from *B only* to *BR*).

Result 0. *Participants, on average, condition their guesses on the observables. Their predictions for each data set vary in the direction of optimal behavior. However, in all data sets, there are deviations from optimal behavior.*

Hence, in this section we have shown that despite the abstract framing of the experiment, participants—on average—utilize information extracted from the data sets to inform their predictions. Specifically, the way participants condition their predictions on the status of the lights varies substantially across data sets, in a manner that reflects the underlying correlation structure of each data set. In addition, the evidence presented in Figure 4 indicates that there is substantial heterogeneity. In the next section we will better characterize the source of such heterogeneity and aim to understand its main drivers.

5 Main Results

5.1 Heterogeneity and Disagreement

Our section of preliminary findings indicates that, in aggregate, participants respond to the main manipulations, but there is substantial heterogeneity in behavior. The analysis so far has not identified the source of this variation: in our experiment, such heterogeneity could stem from systematic differences in behavior across datasets or from systematic differences across participants (i.e., types). In this section, we show that prediction optimality, in aggregate, varies very little

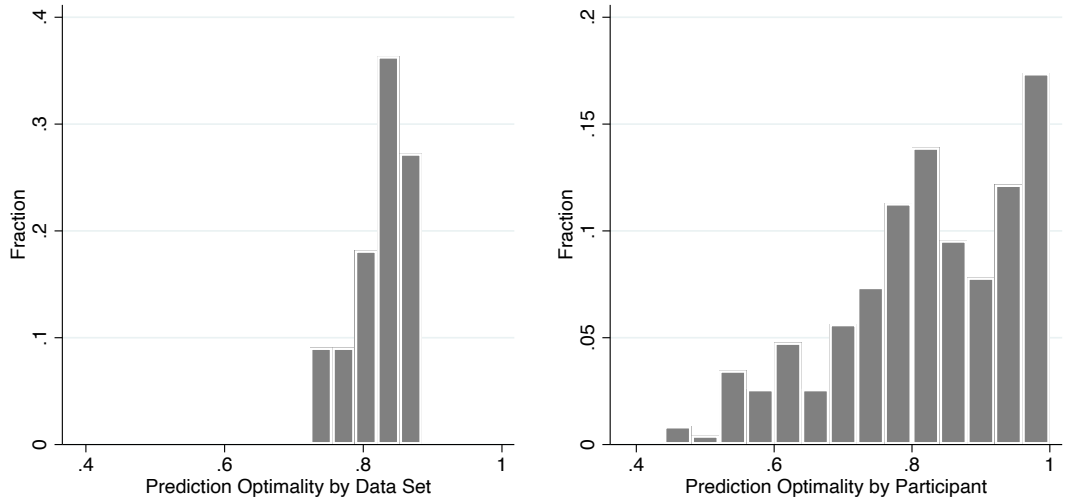


Figure 6: Distribution of Average Prediction Optimality by Data Set and by Participant

Notes: Each observation corresponds to the frequency with which guesses are optimal given the available information in the data sets corresponding to the specific category, and hence, the best achievable value is always 1.

across datasets. Namely, we find that the majority of the observed variation is attributable to differences across participants rather than across datasets. Next, we examine the consequences of this heterogeneity, highlighting how it gives rise to systematic and predictable disagreement among participants.

What Drives Variation in Prediction Optimality: Data Sets or Individual Differences?

By design, our experiment allows us to distinguish between two sources of heterogeneity: (i) variation attributable to differences in the properties of the data sets, and (ii) variation arising from differences across participants.

Figure 6 presents two distributions of prediction optimality: one aggregated at the data set level and the other at the participant level. For comparison, Figure 12 in Appendix D displays analogous distributions, focusing instead on losses in prediction accuracy relative to the best achievable benchmark. As the two figures exhibit highly similar patterns, we restrict our discussion to the main figure.

A first observation from the left panel is that, although prediction optimality varies across data sets, the extent of this variation is relatively modest: optimality rates range from 73 to 88% across the 11 data sets. By contrast, prediction optimality varies substantially across participants,

with individual rates ranging from 46 to 100%. Notably, 21% of participants achieve prediction optimality above 95%, while 26% remain below 75%.

Although some heterogeneity across participants is to be expected, the magnitude and consistency of this variation when aggregated over almost 300 predictions in 11 data sets is striking. Indeed, when we decompose the total variance in prediction optimality, only 4% is attributable to differences across data sets, whereas 96% is due to variation across participants.[34]

Result 1. *A participant's ability (or inability) to make optimal predictions is highly consistent across data sets.*

Is Disagreement Between Participants Predictable?

We conclude this section by highlighting the implications of systematic participant-level heterogeneity. Result 1 shows that, when presented with the same information, participants can arrive at systematically divergent inferences about which action is optimal. To quantify this divergence, we define the *disagreement rate* between any pair of participants at the data set level as the probability that they make opposing predictions when faced with the same input. By definition, the disagreement rate is zero for any pair of participants who both adhere perfectly to the optimal prediction rule. However, heterogeneity in how participants learn from and interpret the data can give rise to substantial disagreement.

We summarize our analysis of disagreement rates here and report detailed results in Online Appendix F. Disagreement between pairs of participants appears to be highly systematic. Figure 16 illustrates this by plotting the average disagreement rate among participant pairs in the last two data sets against their disagreement rates in the first two data sets.[35] The strong correlation between these measures indicates a high degree of consistency across data sets in how much individual participants disagree with each other.[36]

Finally, we show that these patterns persist even when restricting attention to participant pairs whose predictions align in the absence of information about the lights (which we investigate in treatment *Unspecified Prediction*. This finding supports the interpretation of disagreement rates

[34] Using the panel structure of our data, we use the *xtsum* command in Stata to decompose how much of the variance in prediction optimality (computed at the participant-data set level) is due to variation between panels (i.e., across data sets) and how much is due to variation within panels (i.e., across participants for a fixed data set).

[35] Since data sets are shown in random order, we compare each participants disagreement rate with others in their first two data sets to the disagreement rate with the same participants in their last two.

[36] Additional analyses in Online Appendix F, including Figure 23, rule out the possibility that observed disagreements are solely driven by stochastic prediction behavior.

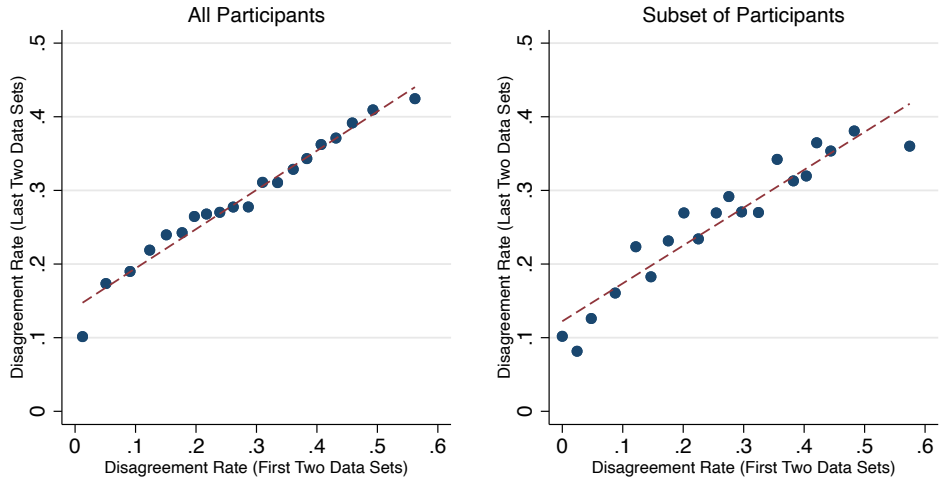


Figure 7: Disagreement Rates in Early and Late Data Sets

Notes: Figure shows binned scatter plots. Disagreement Rate is the frequency with which a pair of participants make opposing predictions about S after observing the same information on (R, B). The left panel includes all participants; the right panel focuses on a subset of participants (60% of the overall population) whose predictions about S are aligned in the absence of any information on (R, B).

as meaningful measures of *polarization*—that is, divergent responses to new information. In this context, polarization refers to cases where initially aligned participants draw systematically different inferences once additional information (i.e., the status of the lights) is introduced.

Result 2. *Consistency of behavior across different data sets makes predicting who will disagree with whom possible.*

Summary and Outline

Our analysis so far demonstrates that participants are actively learning from the data sets—that is, they form mental models that enable them to condition their predictions on observable variables, and these models adapt to the statistical structure of each data set.

More importantly, the analysis reveals that, although individual participants may behave consistently across different data sets, there is substantial heterogeneity across participants. This cross-participant variation suggests the existence of distinct participant types, who systematically interpret and utilize data in different ways, irrespective of the underlying properties of the data sets.

The following sections investigate the sources of this heterogeneity in greater detail. To this

end, we examine the types of mistakes participants make, focusing on two key dimensions: (i) how these mistakes depend on the statistical structure of the data sets, and (ii) how consistently such mistakes are repeated by the same participant. Then, we explore how these mistakes relate to the types of notes participants write for themselves. Finally, we argue that some of these errors may reflect attempts to simplify a complex inference task.

5.2 Systematic Patterns that Organize Heterogeneity

Aiming to understand systematic differences across subjects, we start with a more granular analysis. Specifically, the unit of observation that we now focus on are the choices that a given participant makes in a given data set and we will refer to each observation as a *subject-dataset*. This allows us to (i) examine how mistakes relate to the statistical structure of the data set, and (ii) assess whether individuals tend to repeat the same types of mistakes across data sets. We find that while the prevalence of different types of mistakes vary with the statistical structure, participants generally fall into one of three broad groups: those who are predominantly optimal, those who consistently ignore all relevant variables, and those who condition on irrelevant variables or use relevant ones in suboptimal ways.[37]

5.2.1 Can Participants Be Rationalized as Using Deterministic Prediction Rules?

As a first step, we match each subject-dataset observation to a prediction rule. This approach allows us to identify the kinds of mistakes participants are prone to make in each environment. A prediction rule is a mapping from R and B to S; namely, it specifies how the prediction of the sound depends on the status of the lights.[38] As described in Section 3 (and listed in Table 5 in Online Appendix C), there are 16 deterministic prediction rules.[39] In addition to these rules, we also consider a stochastic prediction rule that predicts $S = 1$ with some probability p (to be estimated) that is independent of the status of the lights.

We match subject-dataset observations to prediction rules using a two-step procedure: (1) We

[37]This last category involves very few observations, making it easier for presentation purposes to merge it with the preceding one.

[38]In the analysis presented in this section, we do not consider prediction rules that condition on the order of events in the prediction task or the data set. The computer interface was designed to minimize the salience or attractiveness of such rules. We also do not find compelling evidence supporting the use of such rules (see Online Appendix F).

[39]These rules differ on whether and how predictions change with R and B. For instance, *G All* is a rule that predicts $S = 1$ for all light configurations and *G w/ R* is a rule that predicts $S = 1$ only when $R = 1$.

estimate the distribution of prediction rules that are used at the population level; and (2) for a subject-dataset observation, given the guesses, we compute the posterior likelihood of following each prediction rule, using population-level estimates as a prior. We classify the subject-dataset observation as using a specific prediction rule by identifying the highest posterior.

Here, we provide an outline of this methodology and refer the reader to Online Appendix E for further details.[40] We use a finite mixture model to estimate the distribution of prediction rules used. The method specifies a set of candidate prediction rules (which corresponds to the 17 rules discussed above) and then estimates their prevalence in the population allowing for the possibility of implementation errors. Formally, for each data set, we use the $27 \times 230 = 610$ predictions to estimate 18 parameters: the probability distribution over the set of rules, implementation error for the rules, and the probability of predicting the sound for the stochastic rule. Then, we use the mixture-model estimates as a prior and compute the Bayesian posterior that a subject-dataset observation is using each of the candidate rules given the set of predictions they make. Each subject-dataset observation is associated with the rule that has the highest likelihood according to this posterior.

This simple method to classify subject-dataset observations successfully reflects the differences in predictions in our experiment. 78% of subject-dataset observations are typed as matching a deterministic rule, although it is worth reiterating that we allow such rules to be implemented with some implementation errors. In 46% of these cases, behavior *perfectly* matches the rule; that is, predictions coincide with the rule for *all* 27 rounds. Overall, among *all* observations typed to correspond to a deterministic rule, predictions match the rule 94% of the time. Focusing on the 22% of subject-dataset observations that are typed as corresponding to a stochastic rule, the overall frequency with which the sound is predicted is 60%. Such behavior could possibly be driven by probability matching behavior (i.e., when the predicted frequency of sound matches the observed frequency of sound), which we discuss further in Section 5.3.[41]

[40] This technique was recently used in Aoyagi et al. (2024) to classify subjects into different repeated-game strategies. Simulations (reported in Online Appendix E) demonstrate that type shares can be reliably recovered for all data sets used in our experiment by following this estimation procedure. For 98% of the participants classified as following a deterministic rule, our classification would be unchanged if instead we identified the rule with which their predictions align most closely.

[41] Indeed, modal behavior perfectly aligns with probability matching. Overall, in 42% of the cases, the frequency with which participants predict the sound to be on is within five percentage points of the true frequency in the data set. See Figure 19 in Online Appendix F for more information on the distribution.

How Does Noise Impact The Prevalence of Mistakes?

Note that, from an ex-ante perspective, how the statistical features of the data set might influence the prevalence of these mistakes is not clear. For instance, noise can affect behavior in various ways. The absence of clear patterns in the data might increase the likelihood that participants erroneously identify nonexistent correlations. Alternatively, in an effort to simplify the environment, participants may ignore observable variables but make optimal predictions subject to this constraint.

A more disaggregated summary of the results is presented in Figure 8, which separates data sets by the noise condition (Low vs. High).[42] Aggregating across Low Noise data sets, we find that only 18% of observations are typed as not corresponding to a deterministic rule. Meanwhile, of those that are typed as consistent with a deterministic rule, the vast majority are typed as using the optimal rule. In fact, close to two-thirds of the observations correspond to the optimal rule in low-noise data sets. Of those, about half perfectly match the optimal rule (in all 27 rounds). Qualitatively, these findings are replicated with high noise: two-thirds of observations are classified as consistent with a deterministic rule and the optimal rule is the modal rule that is most consistent with predictions. However, the comparison of low-noise to high-noise data sets reveals a sharp decline in (i) the share classified as corresponding to a deterministic rule, (ii) the share classified as the optimal rule, and (iii) the share perfectly consistent with the optimal rule.

Why is the change in the share corresponding to deterministic rules noteworthy? The use of a deterministic prediction rule, optimal or not, indicates that a participant can extract statistical patterns—whether correct or incorrect—from the data set, which they then utilize for making predictions. As seen in Figure 8, a significant share of observations (18 and 27%, respectively, with low and high noise) are classified as corresponding to stochastic behavior, which does not condition predictions on light configurations. This observation suggests limitations on the extent to which these people form mental models by studying data sets, and indicates that their ability to do so varies with the strength of the statistical relationships observed in the data sets.

One implication of these results is that participants are more likely to display variation in their predictions in high-noise data sets, which can be captured in a few ways. First, the variance (across subjects) in the overall frequency of predicting the sound increases by 42 percentage points in high-noise relative to low-noise data sets. Second, given a light configuration, the likelihood that any

[42]Table 8 in Online Appendix D reports for each data set the share of observations typed as corresponding to the most popular deterministic rules. In addition, the table also includes information on the share of observations that are not classified as consistent with a deterministic rule.

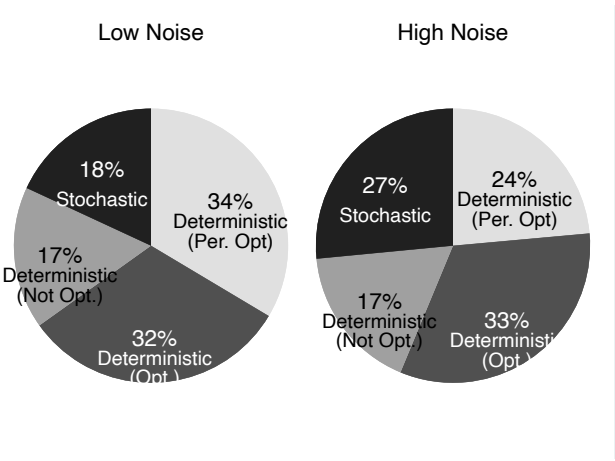


Figure 8: Fraction of Subjects Using Each Type of Prediction Rule

Notes: *Deterministic* (*Stochastic*) refers to whether the subject is typed as (not) using a deterministic rule; *Opt.* refers to whether the rule the subject is typed as using is the optimal one. *Per. Opt.*, for *perfectly optimal* separates out those observations where *all* predictions of the participant are consistent with the optimal rule.

two participants disagree about their prediction (make opposing guesses about the sound) increases by 28 percentage points in high-noise relative to low-noise data sets. Furthermore, disagreement increases not only across participants but also across different predictions by the same participant. That is, the likelihood that two predictions by the same participant for the same light configuration is in disagreement increases by 50% in high-noise relative to low-noise data sets.[43] Taken together, these results show that behavior is more variable (both within and across subjects) in high-noise data sets.

Result 3. *The majority of guesses are consistent with the use of a deterministic rule. The optimal rule is the modal prediction rule in low and high-noise data sets. High noise decreases the likelihood that predictions are consistent with the optimal rule and increases variation and disagreement in predictions.*

A sizeable share of subject-dataset observations are classified as not using optimal prediction rules. This finding presents new questions: What types of deviations from optimal behavior do we observe? Are these mistakes systematic, and how do they depend on the statistical structure of the data set?

[43]Figure 16 in Online Appendix D provides further evidence on these observations.

What Types of Mistakes Do People Make?

In principle, participants can make multiple types of mistakes that result in suboptimal behavior. First, they may fail to incorporate a relevant variable into their predictions. We refer to this as *ignoring a relevant variable*. Second, participants may condition their predictions on variables that are statistically irrelevant to the outcome. We label this error as *conditioning on an irrelevant variable*. Finally, there are other, less frequent, mistakes—such as conditioning on the correct variables but in a suboptimal manner—which we group under the category of *other* mistakes.

While the optimal rule describes the set of variables one should optimally condition on, we can identify the variables the participant actually conditions on from the prediction rule they are classified as using. For instance, the rule *G All* (for "Guess $S = 1$ for all light configurations") does not condition on any of the lights, whereas the rule *G w/ R* only conditions on the red light, and the rule *G w/ R or B* conditions on both lights. If the subject-dataset observation is classified as using *G All* when *G w/ R* is optimal, the observation will be classified as ignoring a relevant variable. By contrast, if a subject-dataset observation is classified as using *G w/ R or B* when *G w/ R* is optimal, the observation will be classified as conditioning on an irrelevant variable.

We present results using two figures. Figure 9 includes data sets where the optimal prediction rule is a function of only one of the observable variables. Figure 10 captures data sets where the optimal rule is a function of both observable variables.[44] In these cases, we distinguish between instances where the subject-dataset observation ignores only one relevant variable and those where both are ignored.

Finally, among the observations that are classified as corresponding to the optimal rule, the figures separate the subset of observations that match the optimal rule perfectly, denoted as Optimal (no errors) versus Optimal (w/ errors). We also develop a measure of "loss" associated with each category: for each observation, we compute its prediction accuracy and the decline it represents relative to the optimal prediction rule. Note that, because participants in the experiment were paid based on the accuracy of their guesses, the estimated loss for each category is directly linked to expected loss in payments (the latter is \$25 times the former). Each pie chart shows the breakdown into these categories and the average loss within that category.

Before we describe the prevalence of different types of mistakes, we note that with respect to optimal behavior, both figures convey a consistent pattern. In all data sets, the majority (or close to it) of observations belong to one of two categories associated with optimal behavior. This finding

[44] See Online Appendix E.7 for information on the No Correlation dataset.

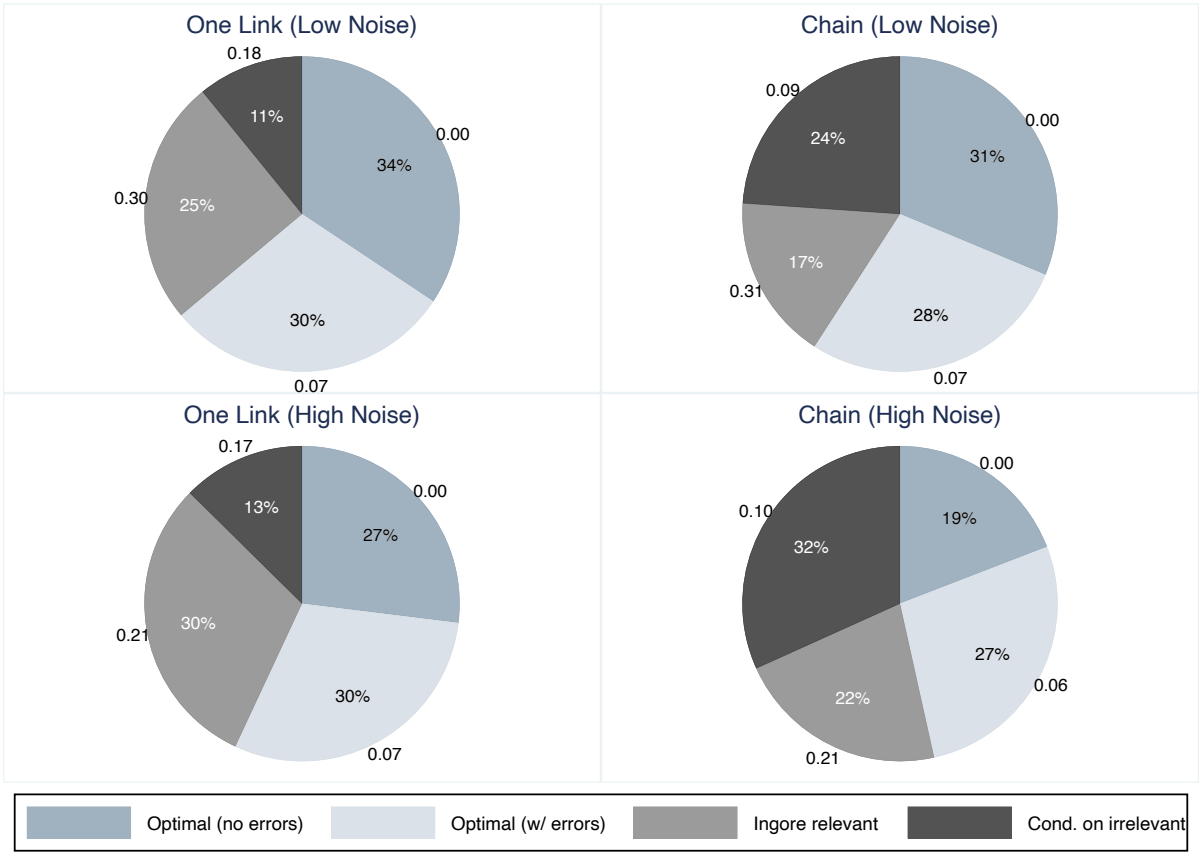


Figure 9: Mistakes and Associated Costs in DAGs with One Relevant Variable

Notes: See Section 3 for detailed descriptions of each DAG. Percentages in each slice of the pie charts denote the relative share of that category. For each slice, the loss associated with such behavior, defined as the decline in guessing accuracy relative to optimal behavior, is also reported outside of the pie chart.

shows that many participants are able to identify distinct patterns in a variety of different data sets, which they then use to make predictions.

To describe mistakes, we start with Figure 9, which depicts DAGs (One Link and Chain), where conditioning on only one of the variables is optimal. First, in both One Link and Chain, we observe that both types of mistakes capture a non-negligible share of observations. However, in both cases, ignoring a relevant variable is the more damaging mistake, because it leads to a larger loss. Second, a common pattern emerges in One Link and Chain with respect to the effect of noise: both types of mistakes become more likely in high-noise environments. Third, the contrast between One Link and Chain informs us about how the underlying correlation structure in the data set can impact the prevalence of different types of mistakes. Recall that the optimal prediction rule is G *w/* R in both One Link and Chain, but B and S are correlated only in the latter. The share conditioning on the irrelevant variable B more than doubles in Chain relative to One Link. This finding suggests that a non-negligible share of participants struggle with understanding conditional independence and display a tendency to condition on *any* variable that correlates with the outcome, an issue we study further in the next section.

Next, we focus on Figure 10, which depicts DAGs (Common-consequence and Full), where conditioning on both variables is optimal. We first notice that there is evidence of all types of mistakes. However, the share corresponding to ignoring both relevant variables is larger in all cases. It captures between 21% of observations (CC(AND)L and Full L) and 43% (CC(OR)H) of observations. Perhaps more importantly, the relative share of this mistake increases substantially in all cases, when we move from low noise to high noise: by nine percentage points in CC(AND) and Full, and 20 percentage points in CC(OR).

Our results point toward heterogeneity in the types of mistakes people are prone to make. The results from this section inform us about how the environment (underlying statistical structure of the data set) impacts the prevalence of these mistakes. We summarize these results below.

Result 4. *Both mistakes—ignoring a relevant variable and conditioning on an irrelevant one—are commonly observed. Conditioning on an irrelevant variable is more likely when the optimal prediction rule ignores variables correlated with the outcome. High noise increases the likelihood that relevant variables are missed.*

In the next section, we leverage our experimental design—which allows us to observe predictions from the same participant across multiple data sets—to examine the consistency of these mistakes in order to understand the systematic disagreement documented earlier.

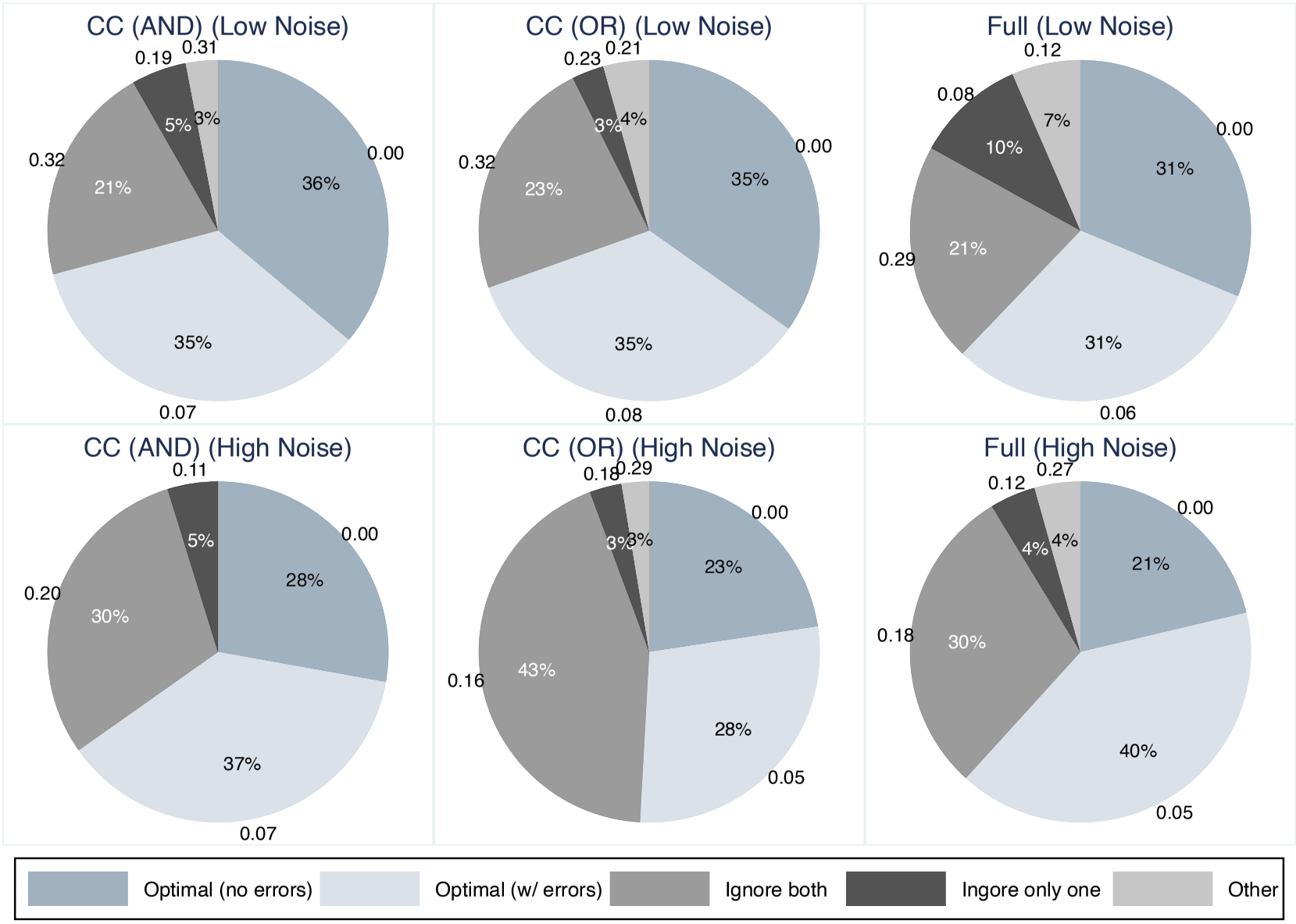


Figure 10: DAGs with Two Relevant Variables: Mistakes and Associated Loss

Notes: See Section 3 for detailed descriptions of each DAG. Percentages in each slice of the pie charts denote the relative share of that category. For each slice, the loss associated with such behavior, defined as the decline in guessing accuracy relative to optimal behavior, is also reported outside of the pie chart.

Table 1: Behavior in One Link and Chain

		High Noise			
		Optimal (no errors)	Optimal (w/ errors)	Ignores relevant variable	Condition on irrelevant variable
	Optimal (no errors)	**.55**	.26	.07	.11
Low	Optimal (w/ errors)	.10	**.48**	.20	.21
Noise	Ignore relevant	.03	.15	**.67**	.14
	Cond. on irrelevant	.09	.16	.21	**.54**

Notes: The table reports the likelihood of different categories of behavior in the high-noise data set of each DAG as a function of the category of behavior in the low-noise data set of the same DAG. For example, 67% of subjects who ignored a relevant variable in One Link L or Chain L also ignore a relevant variable in One Link H or Chain H.

5.2.2 Is Behavior at the Individual-Level Consistent across Data Sets?

The analysis presented in the previous section relied on observations at the subject-data set level. The interpretation of these findings varies based on whether the patterns are consistent for individuals across different data sets. A lack of consistency would suggest that the information participants extract—and how they use it—depends largely on the environment, i.e., the statistical structure of the data. On the other hand, evidence of consistency would indicate the heterogeneities identified at the subject-data set level are actually capturing distinct participant types. Our earlier findings (Results 1 and 2) point in this direction but do not establish whether participants prone to making suboptimal predictions systematically repeat the same types of mistakes across different environments, i.e., different data sets. The results presented in this section confirm that this is indeed the case.

Table 1 shows that participants are likely to repeat the same mistakes across multiple data sets. The table focuses on the One Link and Chain data sets, where two types of mistakes—ignoring a relevant variable and conditioning on an irrelevant variable—are commonly observed.[45] The table reports the likelihood of different categories of behavior in the high-noise data set of each DAG as a function of category of behavior in the low-noise data set of the same DAG. The diagonal values, highlighted in bold, correspond to full consistency. For instance, we see that participants whose predictions perfectly match the optimal rule in the low-noise data set achieve the same outcome with a 55% likelihood in the high-noise data set. Alternatively, participants who ignore a relevant

[45]Corresponding values for the Common Consequence and Full data sets (with different categories) reported in Table 15 of Online Appendix F reveal similar results.

variable (condition on an irrelevant one) in the low-noise data set repeat the same mistake with a 67 (54)% likelihood in the high-noise data set.

Finally, mistakes in One Link and Chain are predictive of behavior in other DAGs. To illustrate this, we divide participants into three groups based on their behavior in One Link and Chain: subjects who are more likely to ignore relevant variables, subjects who are more likely to condition on irrelevant variables, and those who consistently follow the optimal rule. This results in three similarly sized groups (see Table F.1 for shares). Those who are more likely to ignore relevant variables in One Link and Chain have a 61% chance of ignoring both R and B in Common Consequence and Full, while those who are more likely to condition on irrelevant variables do that only 17% of the time. The frequency of mistakes is reversed when it comes to conditioning on irrelevant variables in No Correlation. Subjects who are more likely to ignore relevant variables in One link and Chain make that mistake only 8% of the time in No correlation, while those who are more likely to condition on irrelevant variables do this 20% of the time. Finally, participants who consistently follow the optimal rule in One Link and Chain are less likely than the other two groups to make either mistake, at 6% in both cases (ignoring both variables in Common Consequence and Full, or conditioning on an irrelevant variable in No Correlation).[46,47]

How Sophisticated are the Participants Classified as using the Optimal Rule?

Results so far show systematic differences across participants in their ability to learn from data. We strengthen this finding using data from the *Unspecified Prediction* treatment to demonstrate that some participants classified as consistently applying the optimal prediction rule also show further sophistication and exhibit an understanding of *conditional independence*.

Consider the One Link and Chain DAGs. For a fixed noise level, the conditional distribution $p(S \mid l)$ is identical in both, yielding the same optimal prediction rule: *G w/ R*. However, while B is correlated with S in Chain, it is not in One Link. This distinction becomes meaningful in the additional tasks (rounds 28-36 of Part 2) of the *Unspecified Prediction* treatment, where

[46]Other measures also show consistency of behavior across data sets at the individual level: focusing on the ten data sets where it optimal to condition on lights, (i) 41% of participants are *always* classified as using a deterministic prediction rule, whereas 16% are classified as using one less than half the time; and (ii) 34% of participants *always* condition on at least one of the observable variables, whereas 19% do so less than half the time.

[47]It is also worth nothing that we observe limited learning effects across different data sets. This should not be surprising given that the experimental design does not provide feedback to participants. Namely, we don't find much evidence that experience with an earlier data set impacts learning from a later data set. See Appendix F.2 for more detailed analysis of learning effects.

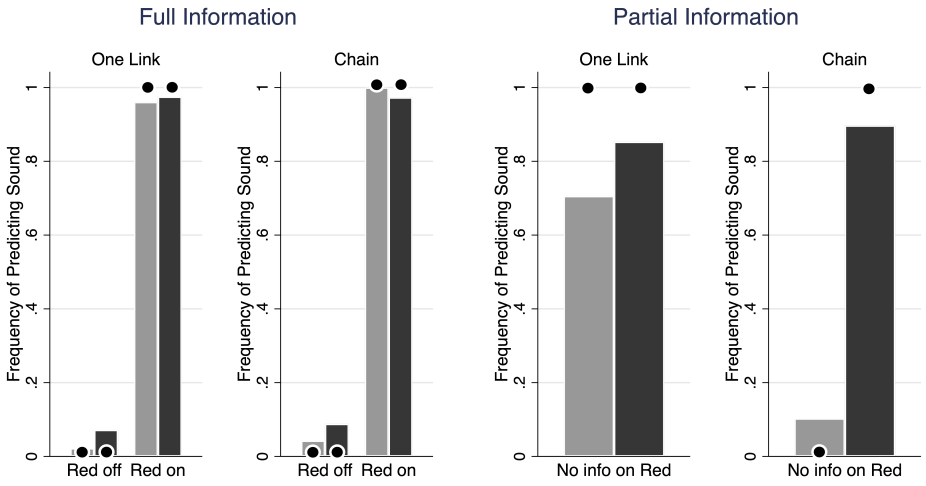


Figure 11: Conditioning on the Blue Light in the Presence or Absence of Information on the Red Light

Notes: Bar colors differentiate between when the blue light is on (black) vs. off (gray). Black dots denote optimal behavior. Full information refers to rounds 1-27, where information on both lights were provided. Partial information refers to rounds 28-36, where information on only one or neither light was provided. We focus on cases where only B is observed. The figure restricts attention to participants who are classified as using the optimal prediction rule across all four One Link and Chain conditions (when focusing on guesses with full information). Graphs pool across noise levels.

participants are shown only partial information about the lights. The optimal prediction when only B is observed differs between One Link and Chain: it is optimal to condition on B only in the latter (i.e. one should exploit the correlation between B and S when that is the only information available, but not otherwise).[48]

Figure 11 compares how predictions condition on B with and without information about R. We restrict the analysis to the 24% of participants who followed the optimal rule under full information in all four relevant data sets. By construction, their full-information behavior is near-optimal and nearly identical between One Link and Chain. Yet their behavior diverges sharply under partial information: in Chain, the $B = 0$ to $B = 1$ increase in frequency of guessing $S = 1$ is 79 percentage points; in One Link, it is only 15 percentage points (p-values < 0.001 and $= 0.156$, respectively).

These results suggest that participants who behave optimally under full information form mental models that are more sophisticated than what their full-information predictions alone reveal. Their behavior under partial information reflects an understanding of conditional independence: they use

[48]Using data from the *Unspecified Prediction* treatment, Fréchette et al. (2025) documents that prediction optimality declines and disagreement among participants increases when predictions are made with partial rather than full information.

the correlation between B and S when R is not observed, but correctly disregard it when R is known.[49]

Result 5. *Participant behavior is consistent across data sets: whether a participant behaves optimally or makes a specific type of mistake in one data set is highly predictive of similar behavior in others.*

We note that Result 5 provides a foundation for the earlier findings on disagreement, helping to explain why individuals consistently diverge in how they learn from and interpret data.

5.3 Why Do People Make Mistakes?

Our results indicate that most participants form mental models, as revealed by the prediction rules identified in the typing exercise. These rules, however, do not always align with the optimal ones. Can these deviations from optimal behavior be rationalized as arising from frictions or constraints in the learning process, as opposed to implementation differences from similar notes? Understanding these connections could help us determine when certain mistakes are more likely to occur, which is the focus of our next analysis.

Table 2 contrasts prediction optimality across light configurations, offering further insight into when different types of mistakes are likely to arise. In each data set, we isolate the light configuration $l := (R, B) \in \{(0,0), (1,0), (0,1), (1,1)\}$ that offers the "strongest evidence" for optimal behavior. We define this as the light configuration $l^\star$ at which $max\{p(S|l), 1 - p(S|l)\}p(l)$ takes the highest value. Formally, $l^\star$ corresponds to the light configuration for which deviating from optimal behavior (by guessing randomly or in the opposite direction) would be most costly for the agent in terms of a decline in prediction accuracy (and payoffs). Participants classified as using the optimal rule achieve 95% prediction optimality for the $l^\star$ configuration, with a similarly high rate for the other configurations. This pattern follows directly from the classification criteria but serves as a useful benchmark nonetheless. Optimality rates corresponding to the different types of errors, reported in the second and third rows of Table 2, are more informative.

Two important observations emerge from Table 2. First, regardless of error type, optimality rates are higher for the $l^\star$ configuration than for other light configurations. Second, participants who condition on an irrelevant variable still achieve high optimality rates with $l^\star$.[50]

[49] As seen in Figure 17 in Appendix D this pattern is not as pronounced for the remainder of the participants who are not classified as using the optimal rule.

[50] Note that regardless of the error type, it is still possible to achieve 100% optimality at $l^\star$. See discussion on

The first finding shows that even participants prone to mistakes tend to learn from data in ways that help them perform better in the cases that matter most. The second highlights a sharp qualitative distinction between the two types of mistakes: those who condition on an irrelevant variable perform substantially better—at least in the cases that matter most—than those who ignore relevant variables. The following analysis offers additional support for this distinction. In the One Link and Chain data sets, 81% of participants who condition on an irrelevant variable are classified as using the *G w/ R or B* or *G w/ R & B* prediction rules. These rules are *constrained optimal* in the sense that they perform best among the class of rules that condition on both lights. By contrast, only 11% of participants who ignore relevant variables in these DAGs are classified as using the rule *G All*, which corresponds to the optimal way to make predictions subject to the constraint that predictions do not condition on the lights. An overwhelming majority of such participants (87%), instead, are classified as displaying stochastic behavior.[51]

In summary, these results suggest that the mental models participants form are responsive to the strength of the available evidence. Thus, the mistakes we observe may reflect attempts to simplify a complex learning task, with behavior optimized primarily for the cases that matter most. Furthermore, ignoring *all* relevant variables appears to be a qualitatively distinct type of error, as it is often associated with stochastic behavior, cannot be rationalized as constrained optimal, and consistently results in lower levels of prediction optimality than other types of mistake.

Overinference and Underinference

The findings above also inform how the two types of mistakes observed in our experiment can be interpreted as instances of *overinference* and *underinference*. Conditioning on an irrelevant variable reflects overinference: participants extract statistical associations from the data but misattribute informativeness to variables that do not improve prediction accuracy. Their behavior is often constrained optimal within the class of rules that condition on both lights, suggesting a failure to consider simpler models or to understand conditional independence. In contrast, ignoring relevant variables reflects underinference: participants fail to recognize useful statistical associations between the observables (the lights) and the outcome of interest (the sound). In the majority of these cases, participants fail to exploit even the most salient statistical regularity: the base rate favoring the

constrained optimal behavior below.

[51] Corresponding results for the Common Consequence and Full data sets are presented in Table 9 in Online Appendix F, where similar patterns emerge. Participants perform better when facing $l^\star$ regardless of the error type. Moreover, those who condition on only one light or on both lights—albeit in a suboptimal way—outperform participants who ignore both lights entirely at this light configuration.

Table 2: Prediction Optimality by Error Type and Light Configuration in One Link and Chain Data Sets

	Share	($l^\star$) Lights with Strongest Evidence	Others
Optimal	0.57	0.95	0.94
Ignore relevant variable	0.24	0.67	0.54
Cond. on irrelevant variable	0.20	0.90	0.69

Notes: "Lights with Strongest Evidence" denotes the light configuration(s) in each data set for which deviating from optimal behavior (by guessing randomly or in the opposite direction) is most costly; "Others" denotes all other light configurations.

sound being on. Thus, their behavior cannot be rationalized as constrained optimal; instead, it suggests fundamental limitations in extracting statistical regularities from data.

Can Mistakes Be Rationalized with Subjective Priors?

At the beginning of the results section, we defined optimality and accuracy purely with respect to the data sets presented to participants. The previous sections showed that even participants classified as using a suboptimal rule (at least in aggregate) are responsive to the strength of the evidence available to them.

This suggests that, for at least some participants, behavior may be rationalized by strong prior beliefs about what constitutes optimal behavior—beliefs that are revised only when sufficient evidence points to the contrary. Although the abstract framing of the experimental design was intended to minimize such *home-grown* priors, we cannot conceptually rule them out as a possible explanation. However, in what follows, we assess whether each participants behavior can be rationalized by *some* prior within a reasonable class—and ultimately rule it out as a primary explanation.

To clarify, consider a Bayesian agent who begins with a prior belief that the likelihood of $S = 1$ is nearly 1, irrespective of the light configuration. A data set consisting of 27 observations might not provide sufficiently strong evidence to dissuade the agent from predicting $S = 1$ for any light configuration (which requires the posterior likelihood of $S = 1$ to fall below 0.5). Such an agent would be classified as ignoring all relevant variables. Starting with other priors could result in other mistakes such as conditioning on an irrelevant variable.

Specifically, we compute the minimal number of predictions a participant would need to change

Table 3: Behavior in One Link and Chain and Notes

	Share	Codes All Data	Summarizes Freq.	Identifies Cor.	Other
Always Optimal	0.30	0.24	0.52	0.15	0.08
More likely to ignore relevant	0.31	0.30	0.12	0.14	0.44
More likely to cond. on irrelevant	0.33	0.11	0.15	0.43	0.31

Notes: "Codes All Data" consists of notes that would allow full replication of the data set (including order of the trials). "Summarizes Frequencies" consists of notes that summarize empirical frequencies of different kinds of trials (R, B, S). All other cases are classified under "Other", which is further separated based on whether the participant identifies correlations and causal relationships between different variables.

such that their behavior can be fully rationalized by a prior.[52] The key idea that this analysis builds on is that regardless of an agent's prior, the optimal response to information should be monotonic in the strength of the evidence provided in the data set for a large set of priors.[53]

This analysis is presented in Online Appendix F. Here, we summarize two main takeaways. First, although there is substantial heterogeneity, most participants classified as using a suboptimal prediction rule also deviate significantly from Bayesian behavior—even when we allow for highly flexible subjective priors. Second, participants who consistently ignore relevant variables are even farther from Bayesian behavior than those displaying other types of errors.[54]

Result 6. *Even participants classified as following suboptimal rules tend to make more optimal predictions in the cases that matter most. This tendency is much weaker among those who ignore all relevant variables.*

[52]This analysis is similar in spirit to the critical cost efficiency index (CCEI) of Afriat (1973), that measures the minimal degree by which budget constraints need to be adjusted to remove all GARP violations from a participant's collection of choices (thereby rendering their behavior rationalizable).

[53]Fixing the configuration of the lights, we posit that if it is optimal for an agent to predict $S = 1$ for a light configuration under a given data set, then it should remain optimal to do so when the data set reflects a stronger signal—whether through more rounds, a higher percentage of $S = 1$ outcomes, or both. For example, if it is optimal to predict $S = 1$ when the data set includes 10 rounds of this configuration in which $S = 1$ in 70% of cases, it should also be optimal to predict $S = 1$ when the data set includes 20 rounds of this configuration in which $S = 1$ in 80% of cases. We show that this monotonicity property holds for any prior characterized by a beta distribution.

[54]This is consistent with earlier findings linking these errors to stochastic behavior, which cannot be rationalized as optimal under any prior.

5.4 Different Approaches to Learning from Data Sets as Seen in Notes

Our results so far suggest that participants learn from data sets in markedly different ways, leading to distinct mental models. Some display a high level of sophistication, while others consistently repeat the same types of mistakes. These findings have relied exclusively on participants' predictions, but in our experiment, participants take *notes* that they then use during the prediction task. Thus, these *notes-to-self* provide additional perspective on how participants examine and learn from data sets. We now study the types of notes participants take, how they relate to behavior in the prediction task, and discuss how the evidence arising from the notes is consistent with earlier results.[55] Here, we focus on the four data sets associated with the One Link and Chain DAGs, where two types of mistakes—ignoring a relevant variable and conditioning on an irrelevant one—are commonly observed.[56]

We classify notes—at the subject-dataset level—according to their content and structure. The resulting categories reflect increasing levels of abstraction, based on how explicitly the notes capture the statistical structure of the data set. The first category (labeled *Codes All Data*) includes observations where participants' notes attempt to transcribe the entire data set, that is, the status of all variables across all trials, typically in order. The second category (labeled *Summarizes Frequencies*) consists of observations where notes summarize the frequency of different types of trials. By definition, the remaining two groups contain observations with little or no numerical content. In the third category (labeled *Identifies Correlations*), participants describe statistical or causal relationships between variables in qualitative terms. The fourth category (labeled *Other*) captures all remaining observations, where notes contain no discernible numerical or verbal references to patterns in the data set.[57] These four categories each account for approximately 21 to 28% of all observations.

First, we establish that participants tend to maintain a consistent note-taking approach across different data sets. For instance, 77% of participants are classified into the same note-taking group in all four data sets shown in Table F.1.[58] This consistency in note-taking provides further evidence that participants adopt stable strategies in how they learn from and engage with data.

Next, we examine the connection between prediction behavior and note-taking styles. In Table

[55] Details on how the notes were coded are provided in Online Appendix G, which includes the protocol that two research assistants were provided to code the text in the notes.

[56] Online Appendix F provides additional analysis for the other data sets.

[57] Examples of note classifications are provided in Online Appendix F.

[58] Sixty percent of participants are classified into the same group across *all* data sets. Most of the variation arises between the last two categories, which are more difficult to distinguish.

F.1, participants are assigned to four groups based on their behavior in the One Link and Chain data sets. The first group consists of those who consistently use the optimal prediction rule. The second includes participants who are more likely to ignore a relevant variable than to condition on an irrelevant one; the third, those for whom the reverse is true.[59]

First, we observe that while optimal behavior can emerge from various note-taking styles, it is most strongly associated with notes that summarize frequencies (52% of such cases). Second, different types of mistakes are linked to distinct note-taking approaches. Ignoring a relevant variable is associated with two contrasting styles: either attempting to code the entire data set (30%) or providing no statistical information at all (44%). Conditioning on an irrelevant variable, by contrast, is most often associated with notes that either verbally describe correlations (43%) or omit statistical content entirely (31%).

Furthermore, Table 12 in Appendix F.1 reveals that using a deterministic prediction rule is associated with specific styles of note-taking, in particular, reporting key frequencies or directly identifying correlations in the data. This finding suggests that some participants study data sets directly with the goal of identifying patterns they will later use in the prediction task. Failing to do this at the note-taking stage translates to participants not using such patterns at the prediction stage, even if notes contain sufficient information to reveal these patterns (e.g., in those cases where notes code all the data).

Analysis in Appendix F.1 further demonstrates that coding all the data is a highly ineffective note-taking strategy, and it is costly in two key ways. First, participants in this group use nearly twice as many characters in their notes compared to others. Second, this strategy is also costly in terms of time. Neither the length of the notes nor the time spent taking notes is associated with higher prediction optimality (see Tables 16-20 in Appendix F). Together, these findings suggest that the suboptimal behavior associated with this note taking strategy is not due to a lack of effort or willingness to learn from the data, but to an inability to make use of the available information.

These patterns also point to distinct mechanisms underlying stochastic behavior in our experiment, particularly given its association with ignoring relevant variables. Some participants who code all the data might sample from these observations to make predictions. For example, when asked whether the sound is on when both lights are on, they may be finding a similar trial in their notes (where both lights were on) and base their prediction on whether the sound was on in that specific trial. Alternatively, stochastic behavior could be a response to a lack of information,

[59]We do not report on the remaining fourth group, which consists of only the 6% of participants for whom both types of mistakes are observed with equal probability.

as with the last category of notes, and could also result from difficulties interpreting the complex information in the first category of notes. It is worth noting that we find limited evidence for *willful* probability matching behavior: only 5% of participants in the second category—who specifically code the probability of the sound—display stochastic behavior.

Taken together, these findings show heterogeneity in how participants learn from data sets, reinforcing our earlier results. Some note-taking styles clearly suggest participants are deliberately looking for patterns in the data. They do so by either summarizing key statistical properties or by qualitatively identifying and verbally reporting correlations among variables. Others are unable or choose not to engage with the data in the same way, which proves costly at the prediction stage. These differences in how participants approach learning from data also have implications for anticipating disagreement. As shown in Table 18 Appendix F, a participant's note-taking style predicts the extent to which others disagree with their predictions. Furthermore, Table 23 in Online Appendix F provides a more detailed analysis by examining the connection between different note-taking styles and the specific prediction rules used at the DAG level.

Result 7. *Participants differ considerably in how they approach learning from data, as reflected in their note-taking styles. These styles are closely linked to whether participants make optimal predictions or exhibit systematic errors.*

6 Discussion

This paper reports results from an experiment designed to study how individuals form mental models by learning from observational data. Despite the abstract nature of the experiment—which deliberately strips away contextual cues—participants are, on average, highly effective at extracting statistical regularities from the data and using them to guide their predictions.

However, at the individual level, our results reveal substantial and persistent heterogeneity across participants in their ability to learn from observational data. Participants who are presented with identical data sets consistently disagree on which variables are predictive of the outcome and how those variables relate to one another. That is, participants disagree on what constitutes optimal behavior after seeing the same information because they organize and interpret the data using distinct mental models.

In addition, our findings provide further insights on when and why participants systematically use conflicting mental models to organize observations.

First, we find that mistakes are not uniformly distributed across environments. Instead, different types of mistakes are associated with distinct features of the data. For example, when spurious correlations are present, participants tend to condition on variables that should be ignored. In contrast, in noisy environments—where one might expect participants to perceive patterns that do not exist—they instead tend to overlook variables that are in fact relevant.

Second, despite features of the data sets being predictive of mistakes, we find that participants display a strong tendency to repeat the same type of mistake across different data sets. These patterns suggest the presence of persistent individual-level differences in how people interpret and learn from data, and provide a foundation for understanding disagreement.

Third, we provide evidence that these differences are linked to distinct approaches to learning from data and can be interpreted as arising from cognitive frictions—namely, challenges individuals face in discerning statistical structure from observational data, as reflected in the following patterns.

(1) Even among participants who are classified as adopting incorrect mental models, prediction optimality increases with the strength of the evidence in the data. This suggests an attempt to simplify the prediction task by focusing only on the most consequential cases. This pattern is particularly pronounced among participants who condition on irrelevant variables. By contrast, those who ignore relevant variables naturally learn very little from the data. This mistake, therefore, is reflective of more fundamental limitations in the ability to read data.

(2) How individuals organize their observations at the note-taking stage plays a critical role in determining whether and how those observations are used later on. Even when notes contain the information needed for optimal prediction, participants who fail to represent the underlying statistical relationships explicitly are more likely to make systematic errors. These findings suggest a more basic challenge than simply forming the correct model, namely that, for some subjects, the difficulty lies in the process of abstraction itself—extracting statistical patterns from raw data—which proves to be a significant source of friction.

By tracing disagreement among equally informed individuals to variation in their underlying mental models, our results illuminate an important mechanism that may shed light on diverse phenomena such as ideological polarization or high trading volumes in financial markets.[60] Thus, our

[60]Specifically, our result indicate directional disagreement on how to bet on certain events among subjects who were given identical information. This connects to the large literature in finance, going back to Miller (1977) and Harrison and Kreps (1978), studying sustained disagreement between optimists ("bulls") and pessimists ("bears") in financial markets. Recently, Bastianello et al. (2025) trace disagreements among professional financial forecasters to differences in the mental models they use when valuing firms. Echoing our results, they show that heterogeneity

results show that disagreement can persist even under identical information, reflecting differences in how individuals interpret and learn from data. Nonetheless, our findings also point to possible avenues for improving learning outcomes and reducing disagreement. Participants who simplify the task by focusing on the most salient cases might adopt more nuanced models when exposed to richer data or stronger incentives. Similarly, participants who study the data carefully—as reflected in their notes—but fail to aggregate observations to uncover statistical relationships may learn more effectively when guided toward specific patterns. That is, even individuals who struggle to infer the true structure on their own may still recognize and adopt alternative models that better align with the data when such models are exogenously presented to them. Given the varied nature of mistakes, our results highlight the challenges one-size-fits-all interventions are likely to face in helping people adopt more accurate models. At the same time, by organizing this heterogeneity, our findings suggest avenues for designing a few interventions that could help many people.

References

Afriat, Sydney N, "On a system of inequalities in demand analysis: an extension of the classical method," *International economic review*, 1973, pp. 460–472.

Aina, Chiara and Florian H. Schneider, "Weighting Competing Models," *Working Paper*, 2025.

Akerlof, George A and Robert J Shiller, *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*, Princeton university press, 2010.

Ali, S Nageeb, Maximilian Mihm, Lucas Siga, and Chloe Tergiman, "Adverse and advantageous selection in the laboratory," *American Economic Review*, 2021, *111* (7), 2152–78.

Ambuehl, Sandro and Heidi C. Thysen, "Choosing Between Causal Interpretations: An Experimental Study," *Working Paper*, 2024.

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart, "Narratives about the Macroeconomy," *Working Paper*, 2021.

Aoyagi, Masaki, Guillaume R Fréchette, and Sevgi Yuksel, "Beliefs in Repeated Games: An Experiment," *American Economic Review*, 2024.

reflects both firm-specific factors and analyst characteristics, with the latter playing a more important role.

Araujo, Felipe A, Stephanie W Wang, and Alistair J Wilson, *American Economic Journal: Microeconomics*, 2021, *13* (4), 1–22.

Augenblick, Ned, Matthew Backus, Andrew T. Little, and Don Moore, "Assumptions, Disagreement, and Overprecision: Theory and Evidence," *Working Paper*, 2025.

Barron, Kai and Tilman Fries, "Narrative Persuasion," *Working Paper*, 2024.

Bastianello, Francesca, Paul Décaire, and Marius Guenzel, "Mental Models and Financial Forecasts," *Working Paper*, 2025.

Bohren, J Aislinn and Daniel N Hauser, "Learning with heterogeneous misspecified models: Characterization and robustness," *Econometrica*, 2021, *89* (6), 3025–3077.

Bramley, Neil R, Tobias Gerstenberg, Ralf Mayrhofer, and David A Lagnado, "Time in causal structure learning.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2018, *44* (12), 1880.

Cason, Timothy N and Charles R Plott, "Misconceptions and game form recognition: Challenges to theories of revealed preference and framing," *Journal of Political Economy*, 2014, *122* (6), 1235–1270.

Charles, Constantin and Chad Kendall, "Causal Narratives," *Working Paper*, 2024.

Charness, G. and D. Levin, "The origin of the winner's curse: a laboratory study," *American Economic Journal: Microeconomics*, 2009, *1* (1), 207–236.

Dal Bó, Ernesto, Pedro Dal Bó, and Erik Eyster, "The demand for bad policy when voters underappreciate equilibrium effects," *The Review of Economic Studies*, 2018, *85* (2), 964–998.

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

Enke, Benjamin, "What you see is all there is," *The Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

_ **and Florian Zimmermann**, "Correlation neglect in belief formation," *The Review of Economic Studies*, 2019, *86* (1), 313–332.

Esponda, I. and E. Vespa, "Hypothetical Thinking and Information Extraction in the Laboratory," *American Economic Journal: Microeconomics*, 2014, *6* (4), 180–202.

Esponda, Ignacio and Demian Pouzo, "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 2016, *84* (3), 1093–1130.

_ **and Emanuel Vespa**, "Endogenous sample selection: A laboratory study," *Quantitative Economics*, 2018, *9* (1), 183–216.

_ **and** _ , "Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory," *Review of Economic Studies*, 2024, *91* (5), 2806–2831.

_ **,** _ **, and Sevgi Yuksel**, "Mental models and learning: The case of base-rate neglect," *American Economic Review*, 2024, *114* (3), 752–782.

Eyster, Erik and Georg Weizsäcker, "Correlation neglect in financial decision-making," *Working Paper*, 2010.

Fan, Tony Q, "Choice-induced Misspecified Mental Models," *Working Paper*, 2024.

Fréchette, Guillaume R, Emanuel Vespa, and Sevgi Yuksel, "Extracting Statistical Relationships from Observational Data: Predicting with Full or Partial Information," in "AEA Papers and Proceedings," Vol. 115 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2025, pp. 637–642.

Fudenberg, Drew and Emanuel Vespa, "Learning Theory and Heterogeneous Play in a Signaling-Game Experiment," *American Economic Journal: Microeconomics*, 2019, *11* (4), 186–215.

_ **, Gleb Romanyuk, and Philipp Strack**, "Active learning with a misspecified prior," *Theoretical Economics*, 2017, *12* (3), 1155–1189.

Gopnik, Alison, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks, "A theory of causal learning in children: causal maps and Bayes nets.," *Psychological review*, 2004, *111* (1), 3.

Grass, Paul, Philipp Schirmer, and Malin Siemers, "Sticky Models," *Working Paper*, 2025.

Han, Yi, David Huffman, and Yiming Liu, "Minds, Models, and Markets: How Managerial Cognition Shapes Pricing Strategies," *Working Paper*, 2025.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning through noticing: Theory and evidence from a field experiment," *The Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Harrison, J Michael and David M Kreps, "Speculative investor behavior in a stock market with heterogeneous expectations," *The Quarterly Journal of Economics*, 1978, *92* (2), 323–336.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic expectations and misguided learning," *Econometrica*, 2018, *86* (4), 1159–1214.

Kendall, Chad W and Ryan Oprea, "On the Complexity of Forming Mental Models," *Quantitative Economics*, 2024, *15*, 175–211.

Martin, Daniel and Edwin Muñoz-Rodriguez, "Misperceiving Mechanisms: Imperfect Perception and the Failure to Recognize Dominant Strategies," *Working Paper*, 2019.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review*, 2019, *109* (10), 3437–74.

Miller, Edward M, "Risk, uncertainty, and divergence of opinion," *The Journal of finance*, 1977, *32* (4), 1151–1168.

Musolff, Robin and Florian Zimmermann, "Model Uncertainty," *Working Paper*, 2025.

Ngangoué, M Kathleen and Georg Weizsäcker, "Learning from unrealized versus realized prices," *American Economic Journal: Microeconomics*, 2021, *13* (2), 174–201.

Pavlov, Ivan Petrovich, "Conditioned reflexes.," *London: Oxford University Press*, 1927.

Payzan-LeNestour, Elise, Yunshen Yang, and Qihe Tang, "Distinguishing Permanent from Transitory Changes: Experimental Evidence on Human Forecasting Ability," *Working Paper*, 2025.

Pearl, Judea, *Causality*, Cambridge university press, 2009.

_ **and Dana Mackenzie**, *The book of why: the new science of cause and effect*, Basic books, 2018.

Samuelson, Larry and Jakub Steiner, "Robust Latent Data Representations," *Working Paper*, 2025.

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

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

Steyvers, Mark, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum, "Inferring causal networks from observations and interventions," *Cognitive science*, 2003, *27* (3), 453–489.

Weisberg, Deena Skolnick, Elysia Choi, and David M Sobel, "Of blickets, butterflies, and baby dinosaurs: childrens diagnostic reasoning across domains," *Frontiers in Psychology*, 2020, *11*, 2210.

Online Appendix for

Extracting Models From Data Sets: An Experiment Using Notes-to-Self

Guillaume Fréchette Emanuel Vespa Sevgi Yuksel

CONTENTS:

A A Brief Overview of Related Work in Cognitive Psychology

In this appendix, we provide a brief overview of the literature in cognitive psychology that addresses related questions to our paper. This literature is extensive and our focus is on summarizing the main connections between our paper and that literature. We also map links to less directly related papers, and provide references for readers who seek further detail.

One way to organize this literature is to divide it into three main areas, each discussed in a separate section of the appendix. This division, guided by the way the literature has developed, is imperfect, as there are overlaps at several points. A further caveat is that we do not intend our overview to be exhaustive. Our aim is to highlight the main connections rather than provide a full survey of this vast literature.

The first section focuses on studies in which participants were given summary frequencies of data to assess possible correlation between two variables. As noted there, much of this work examines two binary variables, where summary statistics can be presented in a two-by-two table.

The second section relates to *associative learning*, which investigates the extent to which subjects form connections between cues and outcomes, typically when experiencing trials one by one. A classic reference here is Pavlovs dogs (Pavlov (1927)).

The third section turns to work that explicitly addresses causal structures, drawing on the framework of Pearl (2009). While the emphasis in this literature is on causality, correlations remain in the background, providing a natural connection to our work.

In each section, we briefly discuss how the psychology literature informs our design choices, other relevant connections, and the ways in which our paper complements the work we review. The main message is that while there are clear connections, to our knowledge no paper in this literature systematically manipulates the data-generating process (DGP) to assess possible sources of heterogeneity across individuals or across DGPs.

A.1 Learning Covariation from Statistical Summaries

Exogenous data sets. Inhelder and Piaget (1958) report one of the earliest experiments on covariation. Focusing on the cognition of children, they used a format that influenced many subsequent studies. Participants were shown cards with faces varying by eye color (brown or blue) and hair color (brown or blond). Each subject was asked about the relationship between eye and hair

color across different sets of cards. In this study, the variables were contextual in the sense that participants could have preconceptions about the connection between eye and hair color. The main finding was broadly positive: beginning at around age 15, participants were able to identify and compute correlations.

Another early contribution, but within a largely context-free framework, is Smedslund (1963). Here, the subjects were nurses who judged the relationship between symptoms and diagnoses. The environment was stripped of context in that the two symptoms and two diagnoses were denoted simply by letters. In contrast to the first study, the conclusion was largely negative, suggesting that adults have a poor understanding of correlation.

While early evidence on whether people identify correlations was mixed, Ward and Jenkins (1965) established conditions under which covariation can be recognized. They provided one group of participants with trial-by-trial information on two variables, a second group with only a summary table, and a third group with both trial-by-trial information followed by the summary table. The group that received only the summary table was most likely to recognize the covariation between events.

Building on these insights, subsequent studies focused on presenting participants with two-by-two tables summarizing outcomes of two binary variables. In these settings, subjects are given summary statistics on two binary variables, and their main task is to judge whether a correlation exists.

Endogenous data. A natural step within this framework is to let subjects generate the data themselves. Jenkins and Ward (1965), for instance, employed an abstract environment in which a participant pressed one of two buttons and then observed one of two outcomes. The subject repeated this process for sixty trials, with the resulting data automatically organized into a two-by-two table of summary outcomes. A second participant (a spectator) observed the data but did not participate otherwise. The experiment included two conditions: in the first, participants were instructed to obtain one of the two outcomes as often as possible; in the second, they were asked to learn how to produce each outcome. Afterward, participants rated, on a 0-100 scale, the degree of control they believed their choices had over the outcomes. Jenkins and Ward (1965) found that subjects overstated the degree of control: their ratings correlated with the proportion of successful trials, but not with the true degree of control. This finding held for both the active participant and the spectator.[61]

[61] A related literature, with Langer (1975) as a seminal contribution, explores the tendency for individuals to value having control over outcomes even when randomness is involved.

Subsequent studies provided a more nuanced view of these results. Alloy and Abramson (1979) showed that under certain conditions, people could more accurately detect the extent to which their choices influence outcomes. They documented that depressed subjects were more likely than non-depressed subjects to correctly assess the correlation between actions and outcomes. Peterson (1980) further showed that part of the difficulty may be driven by demand effects. In one manipulation, participants were explicitly introduced to the hypothesis that the outcome could be random and were shown a sequence of outcomes consistent with randomness. These participants were more likely to recognize cases in which their choices were irrelevant, relative to those who had not been presented with this hypothesis. These two papers are part of a broader line of work that largely qualified the negative findings of Smedslund (1963) and Jenkins and Ward (1965) (see, e.g., Crocker (1982); Alloy and Abramson (1982); Dickinson et al. (1984)).

A large subsequent literature developed theories and experiments to test the procedures people may follow to learn a correlation from two-by-two tables of summary statistics (e.g., Crocker (1981)). A detailed description of this literature is beyond the scope of this paper, but see Mata (2016) for a recent survey that integrates more recent contributions with earlier work, and Perales et al. (2017) for connections to causal reasoning.

Continuous variables.

While most work in this area focuses on environments with two binary variables, some experiments used a continuous framework. For instance, Erlick (1966) presented participants with information on two continuous variables in an abstract setting. Participants observed a list of trials and were asked to provide a subjective estimate of the correlation. In a given session, subjects faced several lists corresponding to different underlying correlation coefficients. The main finding was that participants subjective reports tended to vary with the underlying correlation, though errors were larger for negative than for positive coefficients.[62]

Subsequent studies with continuous and abstract frameworks found similar results. For example, Jennings et al. (1982) presented 64 undergraduates with 10 pairs of realizations from two continuous variables, one pair at a time. Participants were later asked to provide a subjective assessment of the covariation between the two variables. Across several problems, the experimenters manipulated the degree of correlation. The results showed that participants struggled to detect

[62]Erlick (1966) was one of 160 experiments reviewed by Peterson and Beach (1967), who argued that the cognitive system works as an intuitive statistician inferring the environment. The heuristics-and-biases program (Tversky and Kahneman (1974)) can be seen in part as a reaction to this intuitive statistician view. For a more detailed discussion, see Lejarraga and Hertwig (2021).

weaker correlations but performed better when correlations were stronger. For further examples, see Alloy and Tabachnik (1984).

The crucial role of context. A clear message from earlier studies is that results depend on whether the experiment uses an environment with context or not. One additional manipulation of the Jennings et al. (1982) study discussed earlier is that participants are also asked to judge correlations of two continuous variables but now in an environment with context and without data (e.g., the pair of variables could be students' height and students' weight). In this case, they found that people tend to report meaningful correlations even when the actual evidence of association (not show to participants) is rather small. In other words, this type of evidence suggests that in environments with context priors can play a meaningful role.[63]

In fact, a separate literature that started with Chapman (1967) shows that in environments with context, participants reporting correlations that are not present in the data (illusory correlations) are commonplace. For instance, Chapman and Chapman (1982) document that 91 percent of clinicians respond that a patient who was reported to say that he was 'suspicious of other people' would draw a person with large or atypical eyes in the draw-a-person test. However, actual evidence for associations of this type is very weak.

So far, all studies we have described involve assessing two variables. Attribution theory (e.g., Kelley (1967)), which essentially argues that a person tries to infer causes from observed effects, commonly involves a human assessing more than one possible cause for an observed effect; that is, analysis that involves at least three variables (including possible causes and effect). For example, the fundamental attribution error (Heider (1958)) poses that people have a tendency to overestimate individual dispositions (e.g., she is late because she is selfish) and underestimate situational forces (e.g., traffic). Studies in this literature, however, focus on environments where context is crucial to distinguish between individual dispositions and situational forces.

Graphical presentation formats. The vast majority of this literature focuses on providing participants with a two-by-two table of summary frequencies using two binary variables in the background. As we mentioned earlier, a few studies have used continuous variables providing the information often one trial at a time. More recently, however, there is a literature that provides participants with graphs or data charts and assesses the extent to which people identify correlations. For a recent example on this line of work, see Sher et al. (2017).

[63]See Alloy and Tabachnik (1984) for further evidence that papers using abstract environments are more likely to document participants who detect a correlation relative to environments with context in which behavior can be influenced by strong priors.

Connections to our paper. A first observation is that this strand of the psychology literature largely focuses on environments with two variables. By studying a setting with three variables we can study within the same framework a large number of possible correlations, which are captured by the different directed-acyclic graphs in our design. Purposefully, we did not provide subjects with summaries of statistics, as part of our interest lies in understanding how participants summarize the data using their own notes. Clearly, using an environment with three variables and not providing summaries creates a more demanding setting than the standard one in this branch of the literature. It seems reasonable given all the evidence on context interference to take this step in an environment that is essentially abstract, where it seems natural to expect that priors would not play a relevant role. This means, however, that it is challenging to connect findings from our paper to parts of the psychology literature that rely on context (e.g., attribution theory). This type of connection may be possible for future work that allows for context to play a role.

A.2 Associative Learning

At least since Thorndike (1898), a literature on associative learning has documented how humans and animals learn about the connection between cues and effects. A classic example corresponds to Pavlov (1927)'s dogs that learned that a bell is associated with the delivery of food. This literature has, among other things, documented how associations may be developed, and what may impede learning. A detailed account, particularly in its connection with causal learning, is provided by Le Pelley et al. (2017) and Perales et al. (2017). Here, we briefly summarize some aspects that inform our setting.

First, contiguity of events matters for learning associations. Shanks and Dickinson (1991) asked participants to press on a button at a cost of 1 point. With 90% chance pressing the button leads to an outcome that gave the participant 3 points, though they were not informed of this primitive. Participants were simply asked to maximize the number of points. The evidence suggested that participants pressed the button as frequently as possible only in sessions where the outcome immediately followed the action. Delays of 2 to 4 seconds were enough to increase suboptimality of behavior.[64]

Second, the timing in which more than one association is experienced may impact what is learned. Suppose that subjects learn in a first stage that a certain variable (say, $x = 1$) causes an outcome, and that in a subsequent second stage they experience two variables (e.g., $x = 1$ and

[64]There is substantial research on contiguity, in particular with respect to its connection to causality. See, for instance Lagnado and Sloman (2006), Rottman and Keil (2012), and McCormack et al. (2015).

$y = 1$) causing the outcome. Compare that situation to one in which learning about both variables causing the outcome takes place in a first stage. Research has shown that the association between outcome and the second variable (y) is more likely to form in the latter relative to the former (see, e.g., Aitken et al. (2000)).

Connections to our paper. This literature suggests that how participants learn about correlations may depend on the order in which they experience trials. In addition, it may depend on memory, since trials are typically experienced one by one, and participants are not provided with a history table. In our experiment, for each data set we present participants with all 27 trials at once, and trials are randomly ordered across participants. This meant, for example, that participants could not draw any conclusion with respect to the association between lights due to different timing. In other words, we removed contiguity as a possible tool to identify correlations. In addition, and perhaps more importantly, we do not know of a paper that systematically studies a set of different DGPs to assess the extent to which the same participant's inferences about correlations change as a function of the DGP.

A.3 Causal Structure Learning

This branch of the work in psychology has a DGP, associated with Directed-acylic graphs (DAGs) at its center. There are two main applications of this framework that we describe next.

Learning a Causal Structure from Observations and possible DAGs. A seminal reference in this literature is Steyvers et al. (2003). In one of their experiments, they present participants with a data set and two possible Directed acyclic graphs (DAG) that may have generated it. Specifically, their environment uses three aliens (A, B and C) and aliens have a fixed vocabulary of 10 words. In the common-consequence DAG, random words are independently selected for aliens A and B and alien C has an $\alpha \in [0, 1]$ independent chance of reading the mind of other aliens. Reading the mind means that C would be recorded as repeating the word that they read in the other alien's mind.[65] If C does not succeed in reading any mind, then it randomly selects a word from the 10 available. The experimenters described two DAGs as possible DGPs. In addition to the common-consequence DAG, participants also learn that a common-effect DAG (where alien C learns a random word that may be subsequently repeated by A and B) can generate the data set that they observe. Participants are shown a block of 8 trials generated by one of the two DAGs, where a trial is a word associated with each alien. Subjects are asked to report which of the two

[65] If C succeeds in reading both minds, then it randomly uses one of the two words.

DAGs they believe to be the correct DGP. A session consists of several blocks where participants are presented with different sets of two possible DAGs to assess.

A key result of this experiment is that participants were above chance at detecting the correct DAG, and that they seem to understand that observations cannot distinguish between DAG structures such as $A \to B$ and $A \leftarrow B$. They also document limitations in the comparison of some DAGs.

In a survey of this literature, Rottman and Hastie (2014) mentions that the paper by Steyvers et al. (2003) was the only one in the literature to study how well people can infer a causal structure when only provided data (and a set of possible DAGs).

Learning Causal Structures from Interventions. Causal structures often cannot be inferred from observational data that only offers correlations. To assess causality, the subject must perform experiments (or use the do-operator, Pearl (2009)) and generate data that would be informative. Experiments with this objective in mind are reported in Coenen et al. (2015), Bramley et al. (2015) and the previously mentioned Steyvers et al. (2003). The evidence suggests that many participants can make use of interventions for narrowing down the number of possible structures. However, since interventions are not a focus of our paper, we do not describe these experiments in detail, but for further information the reader is referred to the Rottman (2017) survey.

Connections to our paper. Clearly, manipulating DGPs and assessing participants' abilities to obtain information as the DPGs change is at the center of this literature. The goals relative to our paper are very different. Our interest is not in participants manipulating data to obtain further information. That step can be taken within our framework, but we leave it for future research. Perhaps, more relevant, since our focus is not on understanding whether participants can tell different causal structures apart, we do not provide participants with any specific way in which the data can be interpreted. That is, participants are not offered, for instance, possible DAGs that the data may have been generated by. In fact, we do not provide any information with respect to how the data were generated, precisely because our focus is on understanding how participants interpret a data set and not on whether they can use it to tell given DGPs apart.

B Details of Each Data Sets

In this section, we describe the procedure with which we determined the 11 data sets presented to the subjects.

First, we picked the 5 DAGs (Directed Ayclic Graphs) described in Figure 3 (with two versions of Common Consequence corresponding to AND and OR conditions) as described in Section 3.2. These cover all possible DAGs with three variables (R, B, and S) where one variable (S) is fixed not to cause any of the other variables.[66] For all DAGs where there is some causal relationship between the variables, we also varied the strength of these causal connections as described in Section 3.2.[67]

For each case (Low or High noise) of each DAG (separating AND and OR conditions for Common Consequence), the following parameters remain free:

No Correlation. Three free parameters: probability of $B = 1$, $R = 1$ and $S = 1$.

One Link (Low and High noise). Two free parameters: probability of $B = 1$ and $R = 1$.

Chain (Low and High noise). One free parameter: probability of $B = 1$.

CC AND (Low and High noise). Two free parameters: probability of $B = 1$ and $R = 1$.

CC OR (Low and High noise). Two free parameters: probability of $B = 1$ and $R = 1$.

Full (Low and High noise). One free parameter: probability of $B = 1$.

These free parameters were picked to achieve two goals:

1. Keep the probability of $S = 1$ at(or close to) 62 percent.
2. Increase identification across the 11 different cases: namely, increase cost of using the optimal prediction rule from one case to another while preserving some variation in the data set.

Table 4 lists the parameters chosen for each case. The table also reports the implied probability of each light and sound configuration (R, B, S) given these parameters. Finally, the table also includes the finite sample approximation (over 27 trials) that was shown to the subjects.

[66] In the actual implementation of the experiment, we also vary which lights are labeled as R versus B allowing us to create further variation in the DAGs presented to subjects.

[67] For instance, in the One Link Low Noise condition $p(S = 1 \mid R = 1) = p(S = 0 \mid R = 0) = 0.90$. This value was changed to 0.80 in the high noise condition.

Two characteristics of the parameterization as reflected in the finite sample approximation are worth bringing to attention. (1) In the No Correlation data set, there are no trials where both $R = 1$ and $B = 1$. This is due to the fact that the parameters were chosen to the increase likelihood of observing both $R = 0$ and $B = 0$, a configuration in which optimal prediction about S for this case is uniquely different from all others, helping with identification. (2) In the Common Consequence (OR) High Noise data set, there are no trials where both $R = 0$ and $B = 0$. This is due to the fact that the probability of $R = 1$ and $B = 1$ needed to be high enough to ensure that the probability of the sound remained around 62 percent with high noise.

Table 4: Characteristics of Each Data Set

	Probability of each event			Probability for each configuration of (B, R, S)							
	p(B = 1)	p(R = 1)	p(S = 1)	(0, 0, 0)	(0, 0, 1)	(1, 0, 0)	(1, 0, 1)	(1, 1, 0)	(1, 1, 1)	(0, 1, 0)	(0, 1, 1)
No Correlation	0.10	0.10	0.62	0.31	0.50	0.03	0.06	0.00	0.01	0.03	0.06
	3	2	17	8	14	1	2	0	0	1	1
One Link, Low Noise	0.50	0.65	*0.62*	0.16	0.02	0.16	0.02	0.03	0.29	0.03	0.29
	14	18	17	4	0	4	1	1	8	1	8
One Link, High Noise	0.50	0.70	*0.62*	0.12	0.03	0.12	0.03	0.07	0.28	0.07	0.28
	14	19	17	3	1	3	1	2	8	2	7
Chain, Low Noise	0.70	*0.66*	*0.63*	0.24	0.03	0.06	0.01	0.06	0.57	0.00	0.03
	19	18	17	6	1	2	0	2	15	0	1
Chain, High Noise	0.80	*0.68*	*0.61*	0.13	0.03	0.13	0.03	0.13	0.51	0.01	0.03
	22	19	17	3	1	3	1	4	14	0	1
CC (AND), Low Noise	0.75	0.85	*0.61*	0.03	0.00	0.10	0.01	0.06	0.57	0.19	0.02
	20	23	16	1	0	3	0	2	15	5	1
CC (AND), High Noise	0.80	0.85	*0.61*	0.02	0.01	0.10	0.02	0.14	0.54	0.14	0.03
	22	24	17	0	0	2	1	4	15	4	1
CC (OR), Low Noise	0.54	0.20	*0.62*	0.33	0.08	0.03	0.31	0.00	0.11	0.01	0.13
	12	7	17	9	2	1	8	0	3	0	4
CC (OR), High Noise	0.50	0.25	*0.62*	0.31	0.17	0.05	0.27	0.00	0.08	0.02	0.10
	10	6	17	8	5	1	7	0	2	1	3
Full, Low Noise	0.50	*0.50*	*0.62*	0.36	0.09	0.00	0.05	0.00	0.45	0.00	0.05
	13	13	17	10	3	0	1	0	12	0	1
Full, High Noise	0.35	*0.41*	*0.62*	0.33	0.19	0.01	0.06	0.01	0.27	0.02	0.11
	9	11	17	9	5	0	2	0	7	1	3

Notes: The first three columns (labelled as "Probability of each event") describes the probability with which the blue and red lights as well a the sound is on. If the probability is reported in italics, it is not a free parameter but is derived from other primitives given the DAG structure as described in Section 3.2 and earlier in this Online Appendix. The last 8 columns denote the probability with which the different light and sound configurations occur. Integers underneath each value show the finite sample approximation (27 trials in total) shown to the subjects.

C Description of Prediction Rules

Table 5: Prediction Rules

	Guess By Light Configuration				Conditioning on Lights	
Rule	$R = 1$, $B = 1$	$R = 1$, $B = 0$	$R = 0$, $B = 1$	$R = 10$, $B = 0$	Red?	Blue?
G All	1	1	1	1	0	0
G w/ R	1	1	0	0	1	0
G w/ B	1	0	1	0	0	1
G w/ R & B	1	0	0	0	1	1
G w/ R or B	1	1	1	0	1	1
G Never	0	0	0	0	0	0
G w/ not R	0	0	1	1	1	0
G w/ not B	0	1	0	1	0	1
G w/ not R or not B	0	1	1	1	1	1
G w/ not R & not B	0	0	0	1	1	1
G 0100	0	1	0	0	1	1
G 0010	0	0	1	0	1	1
G 1101	1	1	0	1	1	1
G 1011	1	0	1	1	1	1
G 1001	1	0	0	1	1	1
G 0110	0	1	1	0	1	1

Notes: These 16 rules represent all possible deterministic rules in this setting. A rule is referred to as conditioning on a light, if conditioning on the status of the other light, guesses do change with the status of this light. R and B denote the status of the red and blue lights. To facilitate reading of the table, consider the the following examples: The first rule, abbreviated as *G All*, guesses the sound to be on for all light configurations, and thus does not condition on either the red or the blue light. The fourth rule, abbreviated as *G w/ R & B*, guesses the sound to be on only when both red and blue lights are on. When blue is off, the guess independent of the status of the red light, but when blue is on, the guess depends on the status of the red light. A similar argument shows that this rule conditions on *both* lights.

D Further Analysis on the Aggregate Level

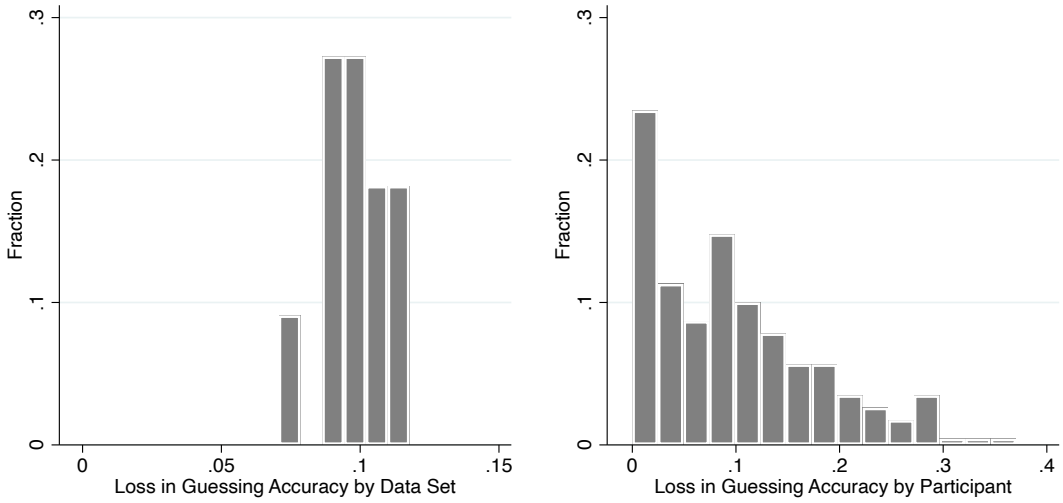


Figure 12: Distribution of Loss in Guessing Accuracy by Data Set and by Participant

Notes: Each observation corresponds to the loss in prediction accuracy relative to the best achievable (assuming optimal predictions) given the available information in the data sets corresponding to the specific category.

Table 6: Prediction Accuracy by Data Set and Deterministic Prediction Rule (%)

Parametrization	Prediction Rule			
	G All	*G w/ R*	*G w/ R & B*	*G w/ R or B*
No Correlation	**62**	40	38	43
One Link High	62	**80**	59	71
One Link Low	62	**90**	64	76
Chain High	61	**80**	78	70
Chain Low	63	**90**	88	84
CC(AND) High	61	70	**80**	63
CC(AND) Low	61	73	**90**	64
CC(OR) High	62	54	45	**75**
CC(OR) Low	62	60	49	**88**
Full High	61	72	63	**77**
Full Low	61	86	82	**90**

Notes: *G All*, guesses that the machine makes a sound for all light configurations. *G w/ R*, guesses the sound only when the red light is on. *G w/ R & B* guesses the sound when both lights are on. *G w/ R or B* guesses the sound when either light is on.

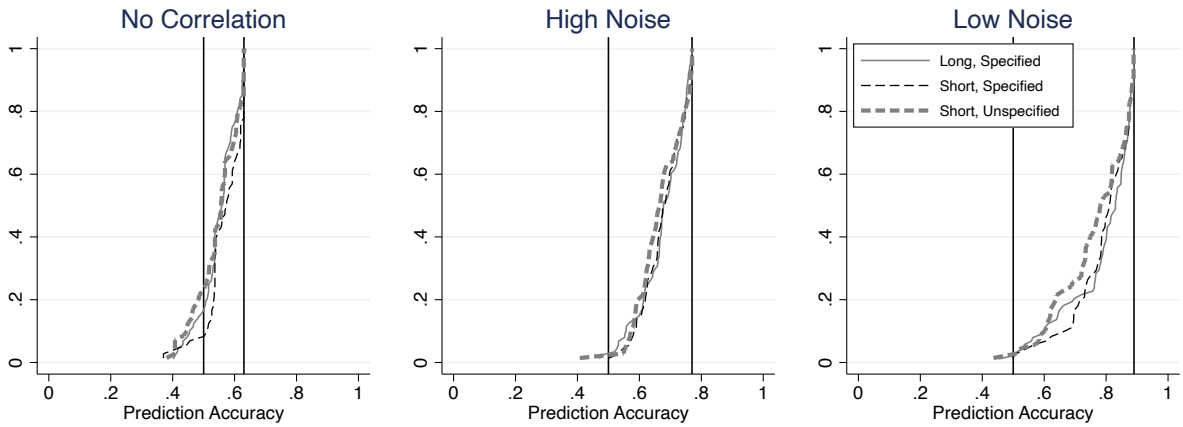


Figure 13: Distribution of Prediction Accuracy (CDF)

Notes: Long vs. Short refers to the length of the notes subjects were allowed to take for each data set. Explicit vs. Unspecified refers to how much information was provided to the subjects about the prediction task. See Section 3.2 for details. The two vertical lines denote the prediction probability for an agent who guesses randomly (on the left) vs. one who guesses optimally (on the right). Each observation takes the average prediction accuracy by subject in the data sets corresponding to the specific category. Equality of the distributions cannot be rejected by a Kolmogorov-Smirnov test (p value > 0.10 in all pairwise comparisons).

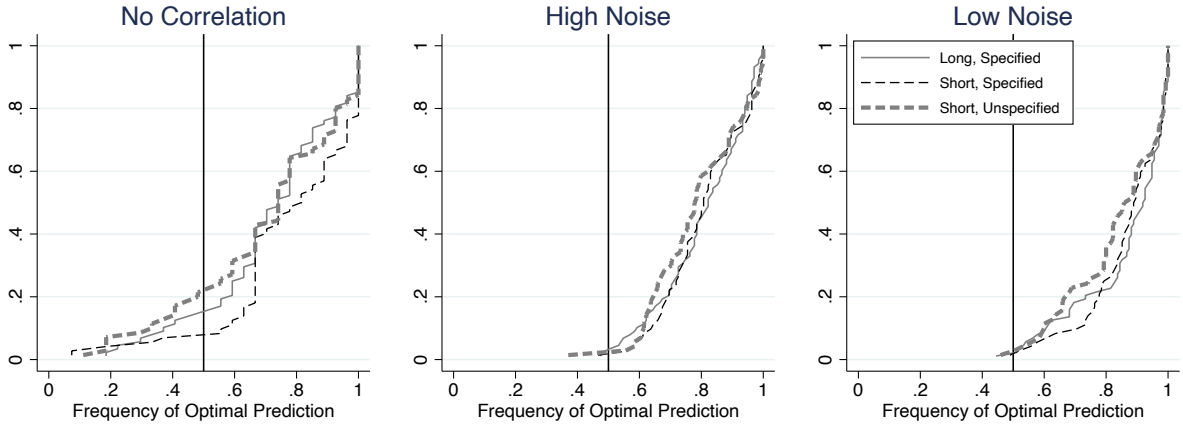


Figure 14: Distribution of Optimality of Guesses (CDF)

Notes: Long vs. Short refers to the length of the notes subjects were allowed to take on each data set. Explicit vs. Unspecified refers to how much information was provided to the subjects on the prediction task. See Section 3.2 for details. The vertical line denotes the frequency of making the optimal guess for an agent who guesses randomly. Equality of the distributions cannot be rejected by a Kolmogorov-Smirnov test (p value > 0.10 in all pairwise comparisons).

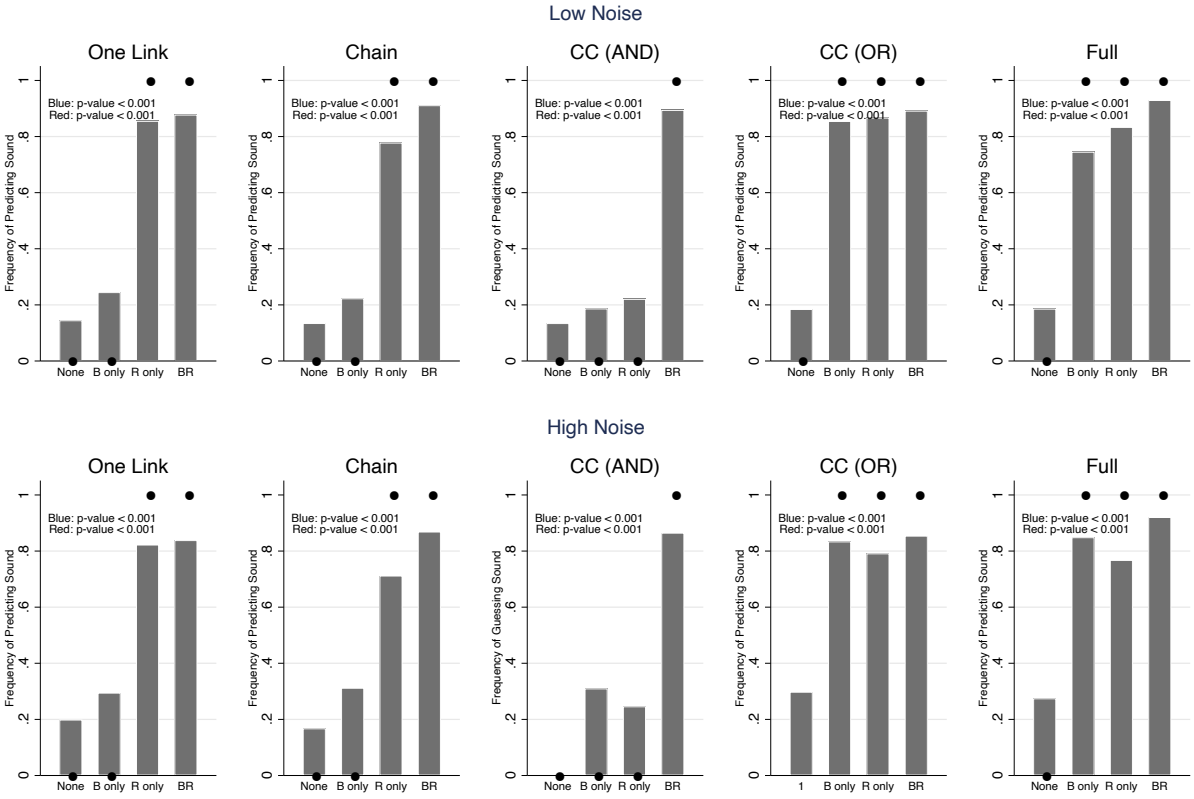


Figure 15: Guesses by Light Configuration

Notes: *None* refers to trials where both red and blue lights are off; *B only* refers to trials where only blue light is on; *R only* refers to trials where only red light is on; *BR* refers to trials where both lights are on. The first category (None) is missing for Common Consequence (AND) with high noise (see Online Appendix B for details on parametrization). Black dots denote optimal behavior.

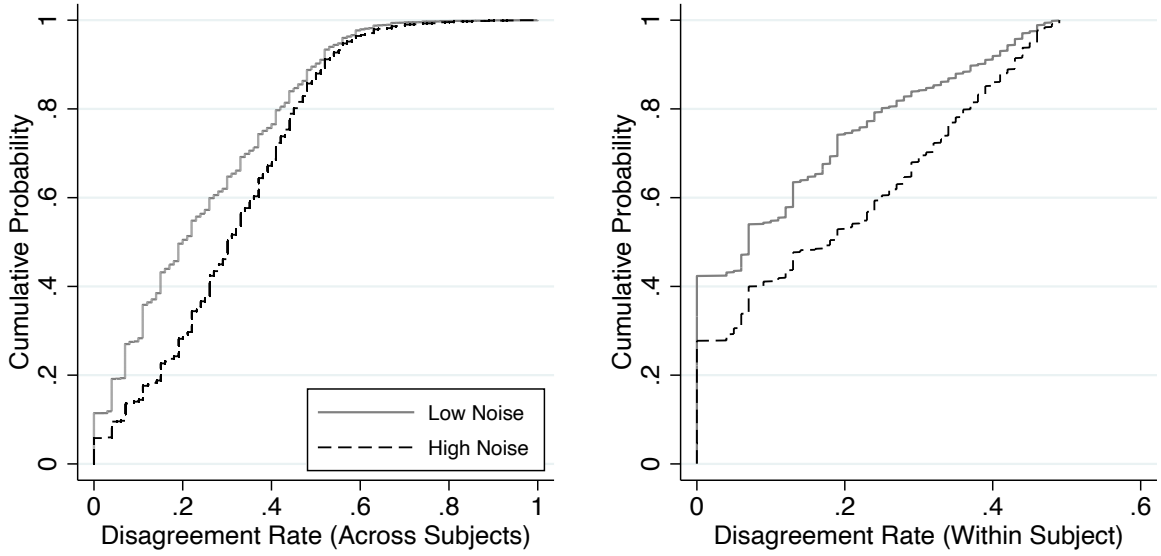


Figure 16: Disagreement Rate by Noise Level

Notes: Disagreement rate across subjects corresponds to the likelihood that any two participants make opposing predictions on the sound after observing the same information on the status of the other variables. This measure is computed at the dataset level for any pair of participants. Disagreement rate within subject corresponds to the likelihood that a participant makes opposing predictions on the sound after observing the same information on the status of the other variables. This measure is computed at the dataset level for each participant.

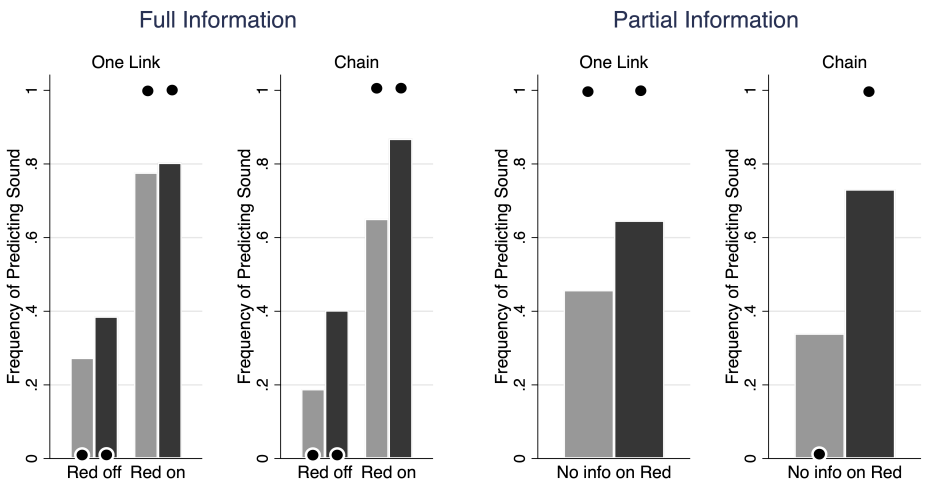


Figure 17: Conditioning on the Blue Light in the Presence or Absence of Information on Red Light

Notes: Bar colors differentiate between when the blue light is on (black) vs. off (gray). Black dots denote optimal behavior. Full information refers to rounds 1-27, where information on both lights were provided. Partial information refers to rounds 28-36, where information on only one or neither light was provided. We focus on cases where only B is observed. The figure restricts attention to participants who are not classified as using the optimal prediction rule for all four One Link and Chain conditions (when focusing on guesses with full information). Graphs pool across noise levels.

E Details and Robustness of Typing Procedure

E.1 Restricting Focus to Stationary Prediction Rules

In our analysis, we focus on a subset of prediction rules that only condition on the light configuration. Namely, we do not allow for prediction rules that condition on the order or sequencing of the dataset or the prediction task. Several measures were taken in the design of the computer interface to minimize the salience and attractiveness of such rules.

- Rows of the data set were deliberately *not* labeled from 1 to 27.
- All rows of the data set were presented simultaneously.
- Order of the rows were randomized at the individual level.
- Prediction tasks were presented one at a time, deliberately *not* labeled from 1 to 27.
- Order of the prediction tasks were also randomized at the individual level and differed from the order associated with the rows observed in the data set.

Despite these measures, in this section we study whether there is any evidence pointing towards the use of such rules. First, the *notes* participants write about each data set are useful as they reveal what aspects of the data set they are attentive to. We do not find any evidence of direct reference to the order or sequencing of the data set (or some other form of serial correlation). But, as reported in more detail in Section 5.4, a non-negligible share of participants code the data set in its entirety in the *notes*. We consider two ways in which participants might condition on the order of events in the data set: (1) Their predictions do not condition on the lights, but match the order with which the sound was observed to be on or off in the data set; (2) Their predictions condition on the lights, and for each light configuration, predictions match order with which the sound was observed to be on or off in the data set. We compute the share of participants whose *notes* code the order of events, and their predictions match behavior as described in (1) or (2) better than *any* of the 16 deterministic prediction rules considered in our analysis. This share is 3.7 percent. 88 percent of such observations are classified as not corresponding to a deterministic rule in our analysis. Among these observations, in only 5.8 percent of cases rules (1) or (2) achieve a fit (match predictions) at a rate higher than 90 percent. Overall, these results suggest that if some participants are drawn to using more complex prediction rules that take into account the order of events in the data set, there is limited scope for such behavior.

E.2 Estimation on the Population Level

We assume that there are 17 possible types. The first 16 are associated with the deterministic prediction rules described in Table 5.[68] Subjects following a deterministic rule are assumed to implement their rule with some noise where the likelihood of making a prediction consistent (inconsistent) with the rule is β $(1-\beta)$ where $\beta \in (0.5, 1]$. In addition, we allow for a stochastic type that predicts the sound to be on with probability $\delta \in [0, 1]$ independent of the light combination.

Let $p \in \Delta^{17}$ denote the share of each type in the population. Let $l \in \{nn, rn, nb, rb\}$ denote all possible light configurations in the first 27 predictions rounds of part 2 of the experiment where the status of both the red and blue lights are revealed to subjects (n denotes either the blue or red light to be off).

Now we describe the behavior of the subjects. Fix the data set. Let $g_{it} \in \{0, 1\}$ denote whether subject i guesses the sound to be on or off in round t of the prediction task. Let $y_{itr} \in \{0, 1\}$ denote whether this guess was consistent with deterministic prediction rule $r \in \{1, .., 16\}$ given the light configuration l observed in that round for the subject. Given p, β and δ, the log likelihood of observing subjects' predictions according to the model can be written as follows:

$$\sum_i \log \left(\sum_{r=1}^{16} p_r \prod_{t=1}^{27} \beta^{y_{itr}} (1-\beta)^{1-y_{itr}} + p_{17} \prod_{t=1}^{27} \delta^{g_{it}} (1-\delta)^{1-g_{it}} \right) \tag{1}$$

Parameters p, β and δ are estimated to maximize the log likelihood function stated in equation 1.

[68] We make one exceptions to this in the Common Consequence (AND) high noise data set, where not all light configurations are observed. In this case, we treat the few deterministic rules that are observationally equivalent as the same. If one of these happens to be one of the top five as listed in Table 5, we present it as this rule in reporting the results.

E.3 Results at the Aggregate Level

Table 7: Population Level Estimates for Prevalence of Different Prediction Rules by Data Set

	Deterministic Rules						Non Deterministic		
	G All	*G w/ R*	*G w/ B*	*G w/ R & B*	*G w/ R or B*	Other	β	Share	δ
One Link, Low Noise	0.02	**0.64**	0.01	0.01	0.06	0.03	0.95	0.23	0.57
One Link, High Noise	0.03	**0.56**	0.01	0.02	0.07	0.05	0.93	0.26	0.58
Chain, Low Noise	0.02	**0.61**	0.01	0.13	0.08	0.01	0.96	0.14	0.59
Chain, High Noise	0.02	**0.48**	0.03	0.14	0.12	0.02	0.93	0.20	0.60
CC (AND), Low Noise	0.03	0.04	0.01	**0.70**	0.02	0.02	0.96	0.17	0.60
CC (AND), High Noise	0.00	0.02	0.04	**0.62**	0.09	0.00	0.93	0.22	0.59
CC (OR), Low Noise	0.05	0.01	0.02	0.00	**0.69**	0.04	0.95	0.18	0.57
CC (OR), High Noise	0.04	0.00	0.02	0.01	**0.52**	0.02	0.94	0.39	0.62
Full, Low Noise	0.03	0.08	0.00	0.06	**0.64**	0.01	0.95	0.18	0.62
Full, High Noise	0.04	0.03	0.03	0.02	**0.54**	0.04	0.92	0.29	0.62

Notes: Table reports estimates for the mixture model. For example, with the One Link, Low Noise data set 64 percent of subjects are classified as predicting the sound to be on only when the red light is on. The optimal rule for each data set is highlighted in bold. See Table 5 in Online Appendix C for descriptions of each prediction rule.

E.4 Typing Procedure

We use mixture model estimates to type subjects at the individual level. Namely, using mixture model estimates as a prior, given the guessing patterns of each subject for each data set, we estimate the posterior probability they are following each of the 17 rules described above. We then classify each subject as following the rule with highest posterior probability.

To demonstrate this, consider a simpler setup where the set of candidate rules consist of only two: *G w/ R* and *G w/ B*, where the former (latter) rule predicts $S = 1$ only when $R = 1$ ($B = 1$). Let c_r^i denote the number of rounds (out of 27) in which participant i's guesses are consistent with rule r and p_r denote the population level estimate for the prevalence of this rule. If implementation error rate is estimated to be ϵ, the posterior probability participant i is using rule r would be computed as follows: $\frac{p_r(1-\epsilon)^{c_r^i}\epsilon^{27-c_r^i}}{\sum_{k\in\{1,2\}} p_{r_k}(1-\epsilon)^{c_{r_k}^i}\epsilon^{27-c_{r_k}^i}}$.

E.5 Results at the Individual Level

Table 8: Type Shares for Different Prediction Rules by Data Set

	Deterministic Rules					Non Deterministic	
	G All	*G w/ R*	*G w/ B*	*G w/ R & B*	*G w/ R or B*	β	Share
One Link, Low Noise	0.02	**0.64**	0.01	0.01	0.07	0.03	0.23
One Link, High Noise	0.03	**0.57**	0.00	0.02	0.06	0.05	0.27
Chain, Low Noise	0.02	**0.59**	0.01	0.15	0.07	0.01	0.15
Chain, High Noise	0.02	**0.47**	0.03	0.17	0.10	0.02	0.19
CC (AND), Low Noise	0.03	0.04	0.00	**0.71**	0.02	0.02	0.17
CC (AND), High Noise	0.00	0.02	0.03	**0.65**	0.09	0.00	0.21
CC (OR), Low Noise	0.05	0.01	0.02	0.00	**0.70**	0.04	0.18
CC (OR), High Noise	0.03	0.00	0.02	0.01	**0.51**	0.03	0.40
Full, Low Noise	0.03	0.10	0.00	0.06	**0.62**	0.01	0.18
Full, High Noise	0.03	0.02	0.02	0.01	**0.62**	0.03	0.27

Notes: Table reports share of participants classified as using each prediction rule for each data set. For example, with the One Link, Low Noise data set 64 percent of subjects are classified as predicting the sound to be on only when the red light is on. The optimal rule for each data set is highlighted in bold. See Table 5 in Online Appendix C for descriptions of each prediction rule.

E.6 Simulation Results

The goal of this section is to demonstrate that type shares in our experiment can reliably be recovered for each data set following the typing procedure described above. To do this, for each data set we take the estimated type shares (as well as implementation error and mixing probability for the stochastic type) estimated in the paper as inputs and we simulate behavior (predictions about the sound given different light configurations) 1000 times. We then use our typing procedure exactly as described above, which involves first estimating a mixture model and then using these as a prior to type participants, to estimate type shares.

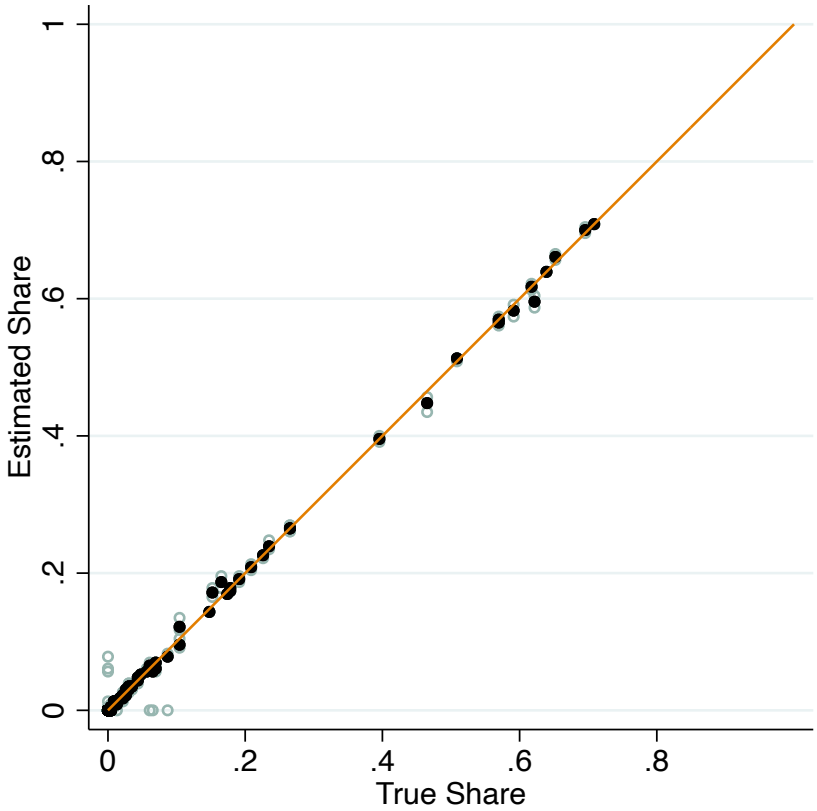


Figure 18: Estimation Results Using Simulations

Notes: Estimation results from 1000 simulated experiments with 230 subjects for 10 different data sets. *True Share* refers to input values to the simulations which correspond to estimates reported in the paper. Solid dots represent median estimate, hollow bubbles represent 25th and 75th percentile estimates.

E.7 Optimality, Mistakes and Loss in No Correlation Data Set

The typing exercise needs to be done in a more careful way in the No Correlation data set for two reasons. (1) The data set chosen for the treatment did not include observations where both lights were on. (2) The frequency of the sound when only the red light was on was 0.5. For this reason, any participant who mixes when only the red light is on, but otherwise always guesses the sound to be on, is also classified as *perfectly optimal*.

In this treatment, 30% of participants are classified as using the optimal rule. 19% of participants are perfectly optimal, i.e., all of their 27 predictions are optimal. Those who are classified as optimal, but not perfectly, achieve a guessing accuracy that is two percentage points below the best achievable benchmark.

The remainder 70% of participants who are classified as suboptimal, in aggregate, achieve a guessing accuracy that is ten percentage points below the best achievable benchmark. 12% of participants are classified as conditioning on an irrelevant variable. This group achieves a guessing accuracy that is ten percentage points below the best achievable benchmark. 57% of participants are classified as displaying suboptimal mixing. These participants achieve a guessing accuracy that is ten percentage points below the best achievable benchmark.

F Further Analysis at the Individual Level

Table 9: Prediction Optimality by Error Type and Light Configuration in Common Consequence and Full Data Sets

	Share	($l^\star$) Lights with Strongest Evidence	Others
Optimal	.63	.95	.92
Ignore both variables	.28	.69	.50
Ignore only one variable	.05	.89	.69
Other	.03	.81	.59

Notes: "Lights with Strongest Evidence" denotes the light configuration(s) in each data set for which deviating from optimal behavior (by guessing randomly or in the opposite direction) is most costly; "Others" denotes all other light configurations.

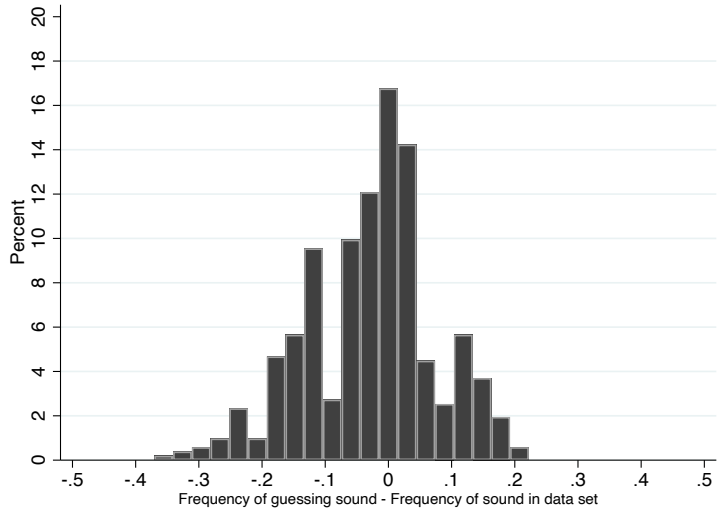


Figure 19: Distribution of Distance between Frequency of Predicting the Sound and the Frequency of the Sound in the Data Set

Notes: Includes only participant-data set observations which are typed to correspond the stochastic rule.

Table 10: Marginal Effects (Probit) of Between Subject Treatment Variation on Behavior in One Link and Chain

	Optinal	Misalignment error	Non-conditioning error
Unknown Prediction	-0.269*	0.209	0.163
	(0.152)	(0.154)	(0.173)
Short Notes	-0.0831	0.281*	-0.149
	(0.151)	(0.145)	(0.177)
Observations	920	920	920

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Table 11: Marginal Effects (Probit) of Between Subject Treatment Variation on Behavior in Common Consequence and Full

	Optinal	Misalignment error	Non-conditioning error
Unknown Prediction	-0.102	0.0809	0.0796
	(0.143)	(0.137)	(0.157)
Short Notes	0.0479	0.118	-0.109
	(0.136)	(0.123)	(0.152)
Observations	1380	1380	1380

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Computing Distance from Bayesian Behavior

Consider an agent who is uncertain about the likelihood of an event p. In our experiment, p can denote the likelihood with which $S = 1$ conditional on (R, B). The agent's prior is given by the Beta distribution and is characterized by two parameters p_0 and η, such that:

$$\mathbb{E}(p \,|\, p_0, \eta) = p_0 \quad \text{and} \quad \mathbb{V}(p \,|\, p_0, \eta) = \frac{p_0(1-p_0)}{\eta+1}.$$

While p_0 denotes the expected value of p, η captures the strength of the prior and, hence, can be interpreted as a measure of the agent's confidence.[69]

The agent updates beliefs on p using outcomes from a Bernoulli process where the probability of the event happening is the true p. The data observed by the agent can be characterized by two parameters: the number of observations n, and the observed frequency of the event among these observations f. The agent's updated posterior is still characterized by a Beta distribution with adjusted parameters $\tilde{p}$ and $\tilde{\eta}$:

$$\tilde{p} = \left(\frac{\eta}{\eta+n}\right) p_0 + \left(1 - \frac{\eta}{\eta+n}\right) f \tag{2}$$

In summary, the model describes how beliefs evolve with feedback as a function of two parameters: p_0, prior expected value of p; and η, a measure of initial confidence.

Note that it is optimal to predict $S = 1$ when $\tilde{p} \geq 0.5$. This happens when

$$n(f - 0.5) > \eta(0.5 - p_0) \tag{3}$$

While we do not observe the agent's prior p_0 and η in the experiment, we observe $n(f - 0.5)$. Let's refer to $n(f - 0.5)$ as the *strength of evidence*. One important implication of this simple model is that the agent's prediction must be monotonic in $n(f - 0.5)$, the strength of evidence. That is, if the agent faces distinct data sets with different features, the optimal prediction for the agent (conditional on a light configuration) will be $S = 0$ for values of $n(f - 0.5) < \bar{e}$ for some $\bar{e} := \eta(0.5 - p_0)$ and $S = 1$ for values of $n(f - 0.5) > \bar{e}$. While we do not know $\bar{e}$, we can search for whether there exists a value of $\bar{e}$ that rationalizes the agent's decisions.

Figure 20 depicts the behavior of four different participants to display variation in behavior. Participant 209 is perfectly optimal, consistently predicting $S = 1$ only when the evidence in the

[69] In the standard formulation, the Beta distribution is characterized by two parameters: α, β such that $\mathbb{E}(p \,|\, \alpha, \beta) = \frac{\alpha}{\alpha+\beta}$ and $\mathbb{V}(p \,|\, \alpha, \beta) = \frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$. The mapping to p_0 and η are such that $p_0 = \frac{\alpha}{\alpha+\beta}$ and $\eta = \alpha + \beta$.

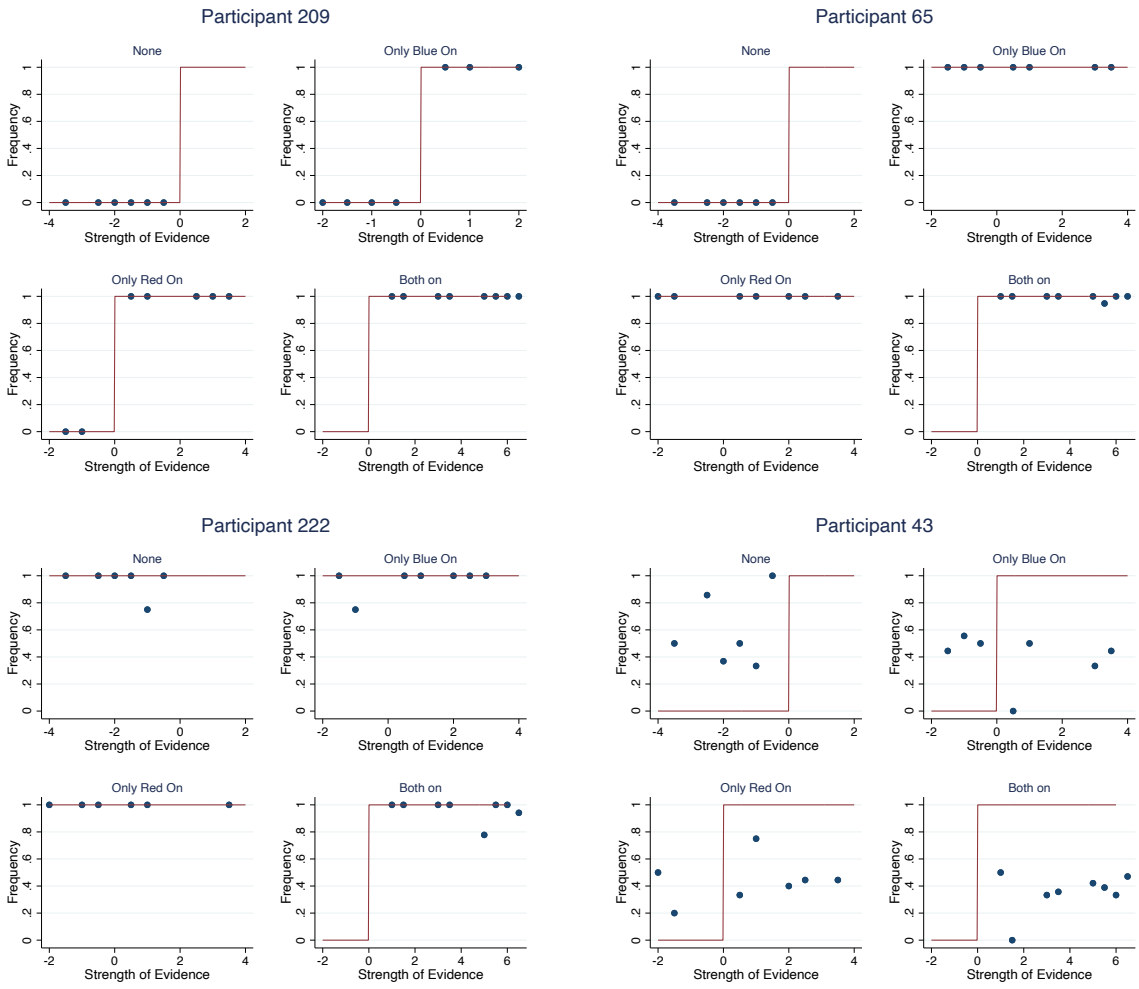


Figure 20: Examples of Participant Behavior

Notes: Examples on how participants predictions (frequency of guessing $S = 1$ changes with the strength of evidence. Red lines for participants 209, 65 and 222 depict how behavior can be rationalized as Bayesian with respect to some p_0 and η.

data points in that direction. Participant 65's predictions perfectly follow *G w/ R or B* in all data sets independent of the strength of evidence. But this can be rationalized with a strong prior as depicted with the red lines. This participant would be classified as conditioning on an irrelevant variable in some of the data sets. Participant 222's predictions very closely follow *G All* in all data sets independent of the strength of evidence. But this can be rationalized with a strong prior as depicted with the red lines. This participant would be classified as ignoring a relevant variable in some of the data sets. Participant 43, in contrast to the other three examples, displays highly stochastic behavior. It is easy to see why such behavior cannot be rationalized as a Bayesian response to some prior.

The examples provided in Figure 20 depict the spectrum of behavior observed in our experiment. While behavior of participants 209 and 65 are different, they can each be perfectly rationalized as Bayesian behavior with respect to some prior. For participant 222, we see some small deviations from Bayesian behavior. One way to see this is that, only a few predictions of this participant would need to be modified (flipped) for them to be perfectly consistent with Bayesian behavior. Participant 43's predictions are far from Bayesian; namely, many predictions of this participant would need to be changed to reconcile with Bayesian behavior.

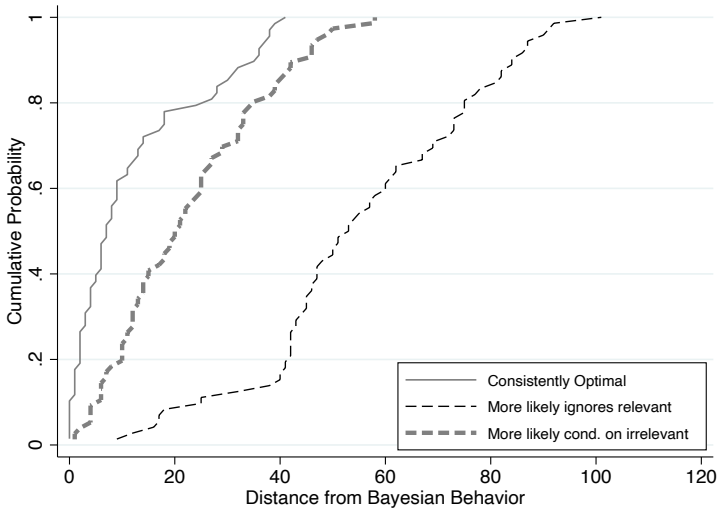


Figure 21: Distribution of Distance to Bayesian Behavior

Notes: Focusing on One Link and Chain data sets we classify participants into three categories: Most likely to be classified as using the optimal prediction rule, most likely to be classified as ignoring a relevant variable, and most likely to condition on an irrelevant one. Distance to the Bayesian Behavior measure is computed over all ten data sets.

Building on this observation, we define Δ_i, *distance to Bayesian behavior* for agent i, as the minimum number of predictions (among 27×10 observed for agent i) that would need to be flipped such that the behavior of the agent can be rationalized with some value of $\bar{e}$ as described before. Figure 21 plots the distribution of this measure for three group of participants. We observe that very few subjects behave in a way that is perfectly aligned with Bayesian behavior. Nonetheless, the third group of participants, those who are more likely to display non-conditioning errors (ignore all relevant variables) appear to be very different from the other two. Namely, it is much more difficult to reconcile their behavior as a Bayesian response to some prior. This is consistent with the example of Participant 43 as depicted in Figure 20. Overall, we find that ignoring all relevant variables is often associated with stochastic behavior in our data.

Behavior With Partial Information on Lights

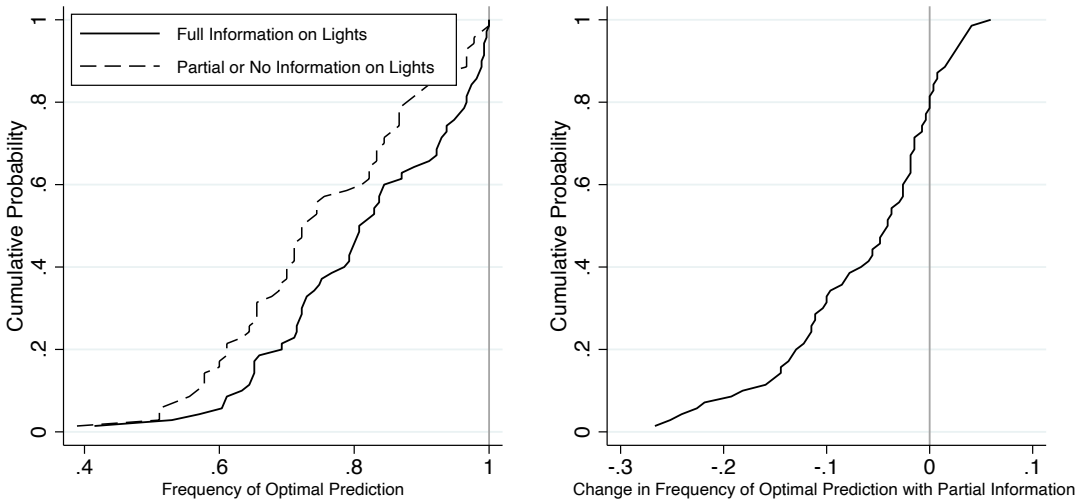


Figure 22: Behavior with Partial Information on Lights

Notes: Left panel plots distribution of frequency of optimal prediction contrasting rounds 1-27 (where there is always full information on the status of the lights) with rounds 28-36 (where there is partial or no information on the status of the lights). Right panel plots the distribution of decline in frequency of optimal predictions when comparing the first 27 rounds to the last nine rounds. In both graphs a unit of observation is a participant's behavior in a single data set.

Examples on Classification of Notes

Codes All Data

srb nrb srb n srb*4 n srb nrb nrnr srb*3 nr n nrb srb n srb sb nrb srb*2 sr
$|BD|N|A|D|N|RD|BD|RD|N|D|N|BD|BD|N|N|RD|BD|BD|N|D|A|N|D|R|BD|D|B|$

Summarizes Frequency

b+s=1 b only=3 r+s=7 r only=2 r+b+s=8 r+b only=2 none+s=1 none=3

12/27 Both - 12D 13/27 none - 3D 1/27 R only - 1D 1/27 B only - 1D

Identifies correlations

both on sound, B no sound, R no sound

ding when both are on

Does not identify correlations

blue

Many dings, many all three

F.1 Notes and Mistakes

Table 12: Notes and Mistakes in One Link and Chain

	Share	Deterministic	Optimal (no errors)	Optimal (w/ errors)	Ignoring relevant variable	Cond. on irrelevant variable
Codes All Data	.21	.73	.24	.34	.30	.12
Summarizes Frequency	.25	.95	.53	.35	.06	.06
Identifies Correlations	.26	.88	.28	.27	.13	.32
Other	.28	.62	.09	.21	.44	.26

See Table for definition of categories.

Table 13: Notes and Mistakes in Common Consequence and Full

	Share	Codes All Data	Summarizes Frequencies	Identifies Correlations	Other
Optimal	.23	.13	.59	.15	.13
Ignores both	.56	.26	.16	.25	.33
Ignores one	.13	.18	.13	.41	.28
Residual	.07	.16	.00	.65	.20

Notes: 'Optimal' refers to participants who consistently use the optimal prediction rule (in all six data sets associated with these DAGs. 'Ignores both' refers to participants who are more likely to ignore both variables . 'Ignores one' refers to participants who are more likely to ignore only one variable. 'Residual' refers to participants are equally likely to ignore both variables or only one variable. See Table for definition of categories associated with different note taking strategies.

Table 14: Notes and Mistakes in Common Consequence AND/OR and Full

	Share	Deterministic	Optimal (no errors)	Optimal (w/ errors)	Ignores (only one)	Ignores (both)
Codes All Data	.21	.69	.19	.41	.06	.34
Summarizes Frequency	.25	.92	.50	.37	.02	.10
Identifies Correlations	.28	.82	.31	.33	.15	.21
Other	.27	.63	.15	.28	.10	.46

Notes: See notes for Table 12.

Table 15: Behavior in Common Consequence and Full

		High Noise			
		Optimal (no errors)	Optimal (w/ errors)	Non-conditioning error (Ignore relevant)	Misalignment error (Cond. suboptimally)
	Optimal (no errors)	**.52**	.35	.08	.04
Low	Optimal (w/ errors)	.12	**.46**	.36	.06
Noise	Non-Conditioning error	.04	.18	**.70**	.07
	Misalignment error	.16	.37	.39	**.12**

Notes: The table reports likelihood of different categories of behavior in the high noise data set of each DAG as a function of category of behavior in the low noise data set of the same DAG. Example: 70 percent of subjects who ignored both relevant variables in CC (AND) L, CC (OR) L or Full L also ignore both relevant variable in CC (AND) H, CC (OR) H or Full H.

Note length, Response Time, and Note Type

Table 16: Determinants of Prediction Accuracy and Optimality (OLS)

	Prediction Accuracy	Prediction Optimality
Number of characters used in notes	0.000152	0.000232
	(0.000112)	(0.000168)
Observations	2300	2300

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Table 17: Average Note Length

	All Data	Treatments with Short Notes
	Note length	Note length
Codes All Data	115	64
Summarizes Frequency	69	56
Identifies Correlations	67	47
Other	55	52

Notes: Reported values are in number of characters. Includes all data sets where optimal behavior involves conditioning on the lights.

Table 18: Disagreement Rate by Note Type

	Codes All Data	Summarizes Freq.	Identifies Correlations	Other
Codes All Data	.30	.23	.27	.34
Summarizes Freq.	.23	.14	.19	.29
Identifies Correlations	.27	.19	.24	.32
Other	.34	.29	.32	.37

Notes: Table shows average disagreement rate for any two randomly matched participants by their note taking style. For instance, a participant who summarizes frequencies in their notes disagree 19 percent of the time in their predictions with a participant who is classified as identifying correlations in their notes.

Table 19: Determinants of Prediction Accuracy and Optimality (OLS)

	Prediction Accuracy	Prediction Optimality
Time spent taking notes	-0.000706	-0.000961
	(0.00109)	(0.00171)
Time spent making predictions	0.00938***	0.0141***
	(0.00242)	(0.00383)
Observations	2300	2300

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Table 20: Average Response Time

	All Data		Treatments with Short Notes	
	Time taking notes	Time making predictions	Time taking notes	Time making predictions
Codes All Data	3.21	2.81	3.55	2.70
Summarizes Frequency	3.49	2.25	3.32	1.75
Identifies Correlations	2.37	1.84	2.11	1.54
Other	3.08	2.10	3.55	1.77

Notes: Reported values are in minutes. Includes all data sets where optimal behavior involves conditioning on the lights.

F.2 Learning

Table 21 indicates only a minor decline in prediction accuracy and optimality in later data sets than in earlier ones. Furthermore, Table 22 suggests the likelihood of conditioning on a variable may be influenced by certain features of the previous data set. For instance, we find that participants are more likely to condition on a light if doing so in the previous data set was optimal, although this effect, while statistically significant, is limited in scope.

Table 21: Determinants of Prediction Accuracy and Optimality (OLS)

	Prediction Accuracy	Prediction Optimality
Case	-0.00105*	-0.00189*
	(0.000574)	(0.000980)
Observations	2530	2530

Case (from 1 to 11) reflects the order in which data set was observed.

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Table 22: Marginal Effects (Probit) on Determinants of Conditioning on a Light

Optimal to Cond. on Light	1.038***
	(0.0753)
Optimal to Cond. on Opposite Light	-0.506***
	(0.0574)
Optimal to Cond. on Same Light Previous Case	0.0750**
	(0.0366)
Optimal to Cond. on Opposite Light Previous Case	-0.0374
	(0.0335)
Case	-0.0173**
	(0.00722)
Observations	4600

Case (from 1 to 11) reflects the order in which data set was observed.

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Disagreement Rates

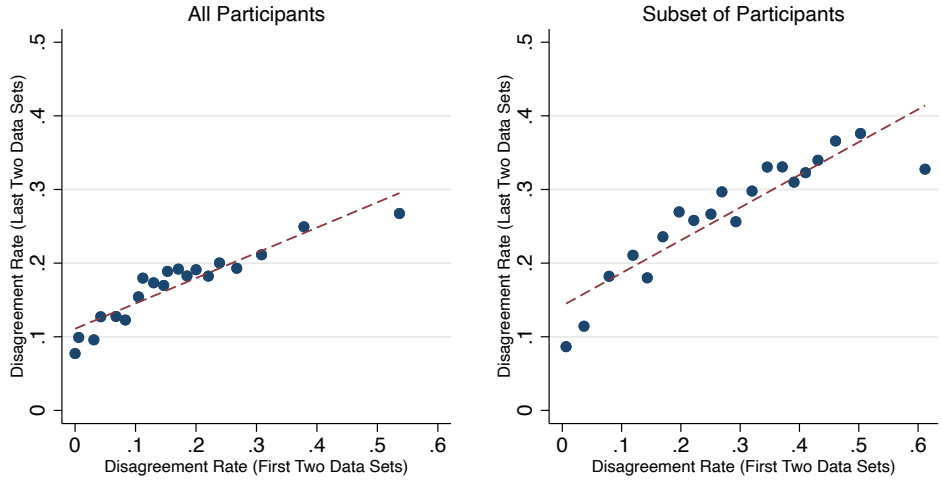


Figure 23: Disagreement Rates in Early and Late Data Sets Among Participants Classified as Using a Deterministic Prediction Rule

Notes: See Figure 16.

Table 23: Use of Prediction Rules by Data set and Note Type

Data set type	Note type	Share	*G All*	*G w/ R*	*G w/ B*	*G w/ R & B*	*G w/ R or B*	*Stochastic*
No Correlation	Codes All	0.21	0.24	0.00	0.00	0.00	0.04	0.65
	Summarize Freq.	0.23	0.37	0.00	0.00	0.02	0.02	0.48
	Identify Cor.	0.20	0.13	0.00	0.00	0.02	0.07	0.42
	Other	0.36	0.20	0.00	0.00	0.01	0.05	0.66
One Link	Codes All	0.22	0.03	0.57	0.01	0.00	0.03	0.33
	Summarize Freq.	0.24	0.00	0.90	0.00	0.00	0.03	0.06
	Identify Cor.	0.25	0.02	0.67	0.00	0.03	0.10	0.13
	Other	0.29	0.06	0.32	0.01	0.03	0.09	0.43
Chain	Codes All	0.21	0.03	0.58	0.01	0.09	0.05	0.21
	Summarize Freq.	0.25	0.01	0.86	0.01	0.03	0.03	0.04
	Identify Cor.	0.27	0.00	0.43	0.04	0.31	0.09	0.11
	Other	0.27	0.05	0.27	0.02	0.18	0.16	0.32
CC AND	Codes All	0.20	0.05	0.03	0.02	0.63	0.00	0.27
	Summarize Freq.	0.25	0.01	0.01	0.02	0.93	0.00	0.03
	Identify Cor.	0.27	0.05	0.04	0.02	0.71	0.02	0.15
	Other	0.27	0.11	0.04	0.01	0.46	0.02	0.32
CC OR	Codes All	0.21	0.02	0.01	0.04	0.00	0.56	0.36
	Summarize Freq.	0.25	0.04	0.00	0.00	0.00	0.82	0.12
	Identify Cor.	0.27	0.03	0.02	0.02	0.01	0.61	0.22
	4	0.27	0.07	0.00	0.02	0.02	0.42	0.45
Full	Codes All	0.21	0.02	0.06	0.00	0.00	0.61	0.30
	Summarize Freq.	0.24	0.01	0.02	0.01	0.00	0.88	0.09
	Identify Cor.	0.29	0.02	0.12	0.01	0.07	0.58	0.17
	Other	0.26	0.07	0.04	0.03	0.07	0.43	0.34

Notes: For each case, values represent share of participants classified as using that prediction rule. We aggregate over low and high noise data sets.

G Protocol for Coding Participants' Notes

Two undergraduate research assistants were provided with a spreadsheet that contained the notes written by participants all treatments. Research assistants were not informed of any aspect of the research beyond the information in the spreadsheet, the instructions to the experiment, and the coding protocol that we attach in this appendix. The coding protocol describes the spreadsheet that coders were provided with as well as the questions that they were asked to answer. Both assistants were asked to code all notes independently. For any differences in coding, they were asked to have a meeting, discuss each difference and reconcile them. We reproduce the protocol next. The screenshot referred to in the protocol corresponds to Figure 1.

Protocol for coding notes

You will code notes from an experiment. The purpose of this protocol is to explain you how to do the coding. Before describing your task, we will explain how the notes that you will code were generated. A participant in the experiment that generated the notes was provided with data from 11 machines. Each machine produced 27 trials. The screenshot on this page provides an example for machine 1. In every trial the machine makes a sound or does not make a sound (last column). A machine has lights of multiple colors, including always one red and one blue light. In every trial, the participant can see whether the red light is on or off (first column), whether the blue light is on or off (second column), but cannot see other lights, which can be on or off (third column). Their task in a later part of the experiment is to make predictions. They will see the status of the red and blue light and their task is to predict whether the machine will make a sound or not. When they face the prediction task, the data that you see in the screenshot below is NOT available to them. But in this first part of the experiment, when they do see the data, they can write notes on the left side of their screens. These notes will be available when they face the prediction task. Each participant wrote notes for themselves on each of the 11 machines that they faced. Your task is to read each note and determine whether the note has certain properties that we will describe later.

Make yourself familiar with the spreadsheet (textdata_tocode.xls). The first column (column A) is the treatment. In Treatments 1 and 2, the task the prediction task consists of we providing them with the status of the red and blue lights and they having to predict whether the machine makes a sound or not. The difference is that in Treatment 1 notes can be up to 280 characters long. In Treatment 2 notes can be up to 75 characters long. In Treatment 3, the prediction task is broader: they are given partial information about a trial (e.g., one light) and are asked to predict information that they are not provided (e.g., the other light or the sound). In Treatment 3, notes

are 75 characters long. What treatment the participant is, does not change the questions you have to have to answer to code each note. The second column, Column B, is the number id that a specific participant was given. The third column, Column C, captures the specific machine the note corresponds to. There are 11 machines or cases, so this variable goes from 1 to 11. Column D (dagnotes) contains the written note that you have to code. For each row, you will fill in several columns, starting from the fifth column onward. If a note is blank (the participant did not write down anything), move on to the next note. For non-blank notes, you will fill in Columns E to H:

E. FullData: Did the participant code the full 27 rows of data in their own words (for example, using their own coding)? [Note: it does not matter if you can or cannot make sense of the way in which they coded the data. All you have to assess is if the participant seemed to have used an approach such that when they read their own notes, they could use their notes to reconstruct the 27 data points.]

- Enter 1: if the participant's notes would let you reconstruct the order in which the data points were presented and the participant does not try to use any code. Example, the participant would use something like 'Red On, Blue On, Sound On', for a trial in which that is what happens. Or a small modification like 'R On, B On, S On.' And the message would then have 27 instances in which the full event is written. If the notes attempt to do this but there is a mistake (for example because they record 26 instead of the 27 trials), enter 1.5 instead of 1.

- Enter 2: if the participants notes would let you reconstruct the order in which the data points were presented, and the participant uses a code to describe the 8 possible cases in the trials. Example, someone uses letter 'A' for the case Red On, Blue On, Sound On; letter 'a' for the case Red On, Blue On, Sound Off; uses letter 'B' for the case Red On, Blue Off, Sound On; letter 'b' for the case Red On, Blue Off, Sound Off; uses letter 'C' for the case Red Off, Blue On, Sound On; letter 'c' for the case Red Off, Blue On, Sound Off; uses letter 'D' for the case Red Off, Blue Off, Sound On; letter 'd' for the case Red Off, Blue Off, Sound Off. Then, the message would involve 27 characters: d c a A D B b a C d c a A D D d a C d c a a c A b a C. This is an example in which the way that the participant coded the message fully lets you reproduce the data set, including the order in which the trials appeared, and the participant uses code for each possible case. If the notes attempt to do this but there is a mistake (for example because they record 26 instead of the 27 trials), enter 2.5 instead of 2.

- Enter 3: if the participants message uses a code but would not let you reproduce the order

in which the trials appeared. Following the previous example, it could be something like: A 4/27; a 3/27, and so on. If the notes attempt to do this but there is a mistake (for example because they divided by 26 instead of 27), enter 3.5 instead of 3.

- Enter 4 if the participant uses numerical information that would not let you fully reproduce the data set. For instance, the participant just says that the sound takes place 14/27 times. If the notes attempt to do this but there is a mistake (for example because they divided by 26 instead of 27), enter 4.5 instead of 4.
- Enter 99: if you are unsure.
- If the answer is no, leave blank.

F. Generic: Do the notes use words to describe the data?

- Enter 1: if the note mentions that the sound is random or something similar. [For example, 'the sound is sometimes on, sometimes off, for certain light combinations.']
- Enter 2: if the note mentions that the sound is 'mostly' or 'overall' on without specifying the frequency or connecting it to a combination of lights. [For example, 'in most of the trials the sound is on.']
- Enter 3: if the note mentions that the sound is always on or off for certain light combinations. [For example, the sound is always on when the blue light and the red light are on.]
- Enter 4: if the note uses mostly words but cannot be classified as any of the previous three. Here is an example of a note that would fall in this category. 'Mach 8: Intervals of 5x ding followed by 1-2 silence. As more trials done, more ding. Seemingly no relation between r/b light and ding. Most r&b=n = no ding. Most r&b=y = ding. There are exceptions.' It does not use the word 'random' directly, so it is not captured by '1.' It does mention a word like 'most' but it also connects the sound to the lights, so it does not fall into '2.' It does not convey that the light is on or off always for some combination of lights, so it does not fall into '3.' But while there are some numbers in this note, it involves mostly words. Since it does not fall into 1, 2 or 3 and it does use mostly words, it qualifies as 4.
- If none of the above aspects is mentioned, leave blank.

G. Correlation: Does the participant uses the word correlation, association, or a related word to describe the data?

- Enter 1: if the answer is yes.
- Enter 99: if you are unsure.
- If the answer is no, leave blank and jump to column F.

H. Causality: Does the participant use the word 'causality' or the idea that one item causes another item? [Example: 'When the blue light is on, the machine makes a sound' or 'When the machine makes a sound, then the red and the blue lights are on'] Does the participant use the word 'lack of causality' or the idea that one item does not cause another item? [Example: 'When the blue light is on, the machine makes a sound' or 'When the machine makes a sound, then the red and the blue lights are on']

- Enter 1: if the answer to either question above is yes.
- Enter 99: if you are unsure.
- If the answer is no, leave blank and jump to column G.

I. Model [Answer this question only if you did not answer E as 1, 2, 3 or 4.] Assume you have to make predictions, as participants in this experiment did. Looking at the notes they wrote, evaluate if any of the following is a good match to what the notes suggest as best strategy for predictions for when the sound is on:

- Enter 0: if the notes suggest predicting that the machine makes the sound all the time.
- Enter 1: if the notes suggest predicting that the machine makes the sound only when the blue light is on.
- Enter 2: if the notes suggest predicting that the machine makes the sound only when the red light is on.
- Enter 3: if the notes suggest predicting that the machine makes a sound only when either the red light or the blue light is on.
- Enter 4: if the notes suggest predicting that the machine makes a sound only when both the red light and the blue lights are on.
- Enter 5: if the notes suggest a clear prediction strategy that is not one of the above (example: notes suggest predicting that the machine makes the sound only when the blue light is off).

- Enter 6: the notes clearly do not suggest a prediction strategy.
- Enter 99: if you are unsure. If you dont know whether it is possible to classify as any of the first 5, enter 99. If you think it may be one of the five but are unsure, do the following. If you think it may be classified as 2, then enter 99.2. If you think it may be classified as 3, enter 99.3. Example: If the note says, Sound is on when red light is on but not otherwise, then you would clearly code that as 2. But if the note says: Red on, sound on 20/27; when Red off, sound not there very often, it may be unclear because the note does not specify what to do if the red light is off. In this case, you may code it as 99.2.

J. Other aspects:

- Enter 1: if the note is very difficult to understand (incomprehensible).
- Enter 2: if the note compares this machine to previous machines.
- Enter 3: if the note conveys that there is a time dependency on trials (e.g., for a given machine, the participant thinks that after a trial in which the blue light was on and there was a sound, there will be another trial in which the red light will be on and there will not be a sound).

H Instructions

The followings are instructions from the Explicit Treatment.

INSTRUCTIONS

You are about to participate in an experiment on decision-making. Please turn off cell phones and similar devices now. Please do not talk or in any way try to communicate with other participants. We will start with a brief instruction period. If you have any questions during this period, raise your hand and your question will be answered so everyone can hear.

What you earn in the session depends partly on your decisions, and partly on chance. This experiment consists of two parts. Part 1 will provide information for the decisions that you make in Part 2. At the end of the session one of the decisions you will make in Part 2 is randomly selected for payment with equal chance. Your payment is equal to $10 plus the earnings in the randomly selected Part 2 decision.

Part 1

- In this part you will be presented with data produced by 11 different *machines*. For each machine, we will show you 27 trials generated by that machine.
- In every trial:
  - the machine either makes a sound or doesn't make a sound.
- A machine has lights of multiple colors, including always one red and one blue light. In every trial:
  - you can see if the red light is on or off;
  - you can see if the blue light is on or off;
  - you cannot see the other lights, which could be on or off.
- The lights and the sounds may or may not be related to each other.
- That is, a trial includes information for lights that can be observed (blue and red) and on whether the machine made a sound or not. It does not provide information for lights that cannot be observed.
- Your task in Part 1 is to take summary notes (at most 75 characters) for each machine because in future parts you will not have access to the data on the trials presented in Part 1.
- In Part 2 you will face the same 11 machines again. For each machine, you will have access only to the notes you took in Part 1, and you will make predictions. For each prediction, you will see:
  - Whether the red light is on or off.
  - Whether the blue light is on or off.

  Your task is to make a prediction about the sound.
- For payment, we will randomly select one of your predictions. If your prediction is correct, we will add $25 to your payoffs.

Part 2 [On the screen]

- In Part 2 you will face the same 11 machines (you saw in Part 1) in random order. For each machine, you will have access only to the notes you took in Part 1.
- For each machine, you will see whether the red or the blue light is on or off and you will have to make a prediction about the sound.
- We will randomly select one of your predictions. If your prediction is correct, we will add $25 to your payoffs.

References

Aitken, Michael RF, Mark JW Larkin, and Anthony Dickinson, "Super-learning of causal judgements," *The Quarterly Journal of Experimental Psychology: Section B*, 2000, *53* (1), 59–81.

Alloy, Lauren B and Lyn Y Abramson, "Judgment of contingency in depressed and nondepressed students: Sadder but wiser?," *Journal of experimental psychology: General*, 1979, *108* (4), 441.

_ **and** _ , "Learned helplessness, depression, and the illusion of control.," *Journal of personality and social psychology*, 1982, *42* (6), 1114.

_ **and Naomi Tabachnik**, "Assessment of covariation by humans and animals: the joint influence of prior expectations and current situational information.," *Psychological review*, 1984, *91* (1), 112.

Bramley, Neil R, David A Lagnado, and Maarten Speekenbrink, "Conservative forgetful scholars: How people learn causal structure through sequences of interventions.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2015, *41* (3), 708.

Chapman, Loren J, "Illusory correlation in observational report," *Journal of Verbal Learning and Verbal Behavior*, 1967, *6* (1), 151–155.

Chapman, Loren J. and Jean Chapman, "Tests results are what you think they are," *in Judgement under uncertainty: Heuristics and Biases, D. Kahneman, P. Slovic and A. Tversky eds.*, 1982.

Coenen, Anna, Bob Rehder, and Todd M Gureckis, "Strategies to intervene on causal systems are adaptively selected," *Cognitive psychology*, 2015, *79*, 102–133.

Crocker, Jennifer, "Judgment of covariation by social perceivers.," *Psychological Bulletin*, 1981, *90* (2), 272.

_ , "Biased questions in judgment of covariation studies," *Personality and Social Psychology Bulletin*, 1982, *8* (2), 214–220.

Dickinson, Anthony, David Shanks, and John Evenden, "Judgement of act-outcome contingency: The role of selective attribution," *The Quarterly Journal of Experimental Psychology*, 1984, *36* (1), 29–50.

Erlick, Dwight E, "Human estimates of statistical relatedness," *Psychonomic Science*, 1966, *5* (10), 365–366.

Heider, Fritz, "The Psychology of Interpersonal Relations," *New York: Wiley*, 1958.

Inhelder, Bärbel and Jean Piaget, "The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures," 1958, *22*.

Jenkins, Herbert M and William C Ward, "Judgment of contingency between responses and outcomes.," *Psychological monographs: General and applied*, 1965, *79* (1), 1.

Jennings, Dennis L, Teresa M Amabile, and Lee Ross, "Informal covariation assessment: Data-based versus theory-based judgments," *in Judgement under uncertainty: Heuristics and Biases, D. Kahneman, P. Slovic and A. Tversky eds.*, 1982.

Kelley, Harold H, "Attribution theory in social psychology.," *in Nebraska symposium on motivation, D. Levine ed.*, 1967.

Lagnado, David A and Steven A Sloman, "Time as a guide to cause.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2006, *32* (3), 451.

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

Lejarraga, Tomás and Ralph Hertwig, "How experimental methods shaped views on human competence and rationality.," *Psychological Bulletin*, 2021, *147* (6), 535.

Mata, André, "Judgment of covariation: A review," *Psicologia*, 2016, *30* (1), 61–74.

McCormack, Teresa, Caren Frosch, Fiona Patrick, and David Lagnado, "Temporal and statistical information in causal structure learning.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2015, *41* (2), 395.

Pavlov, Ivan Petrovich, "Conditioned reflexes.," *London: Oxford University Press*, 1927.

Pearl, Judea, *Causality*, Cambridge university press, 2009.

Pelley, Mike E Le, Oren Griffiths, and Tom Beesley, "Associative accounts of causal cognition," *The Oxford handbook of causal reasoning*, 2017, pp. 13–28.

Perales, José C, Andrés Catena, Antonio Cándido, and Antonio Maldonado, "Rules of causal judgment: Mapping statistical information onto causal beliefs.," *The Oxford handbook of causal reasoning*, 2017, pp. 29–51.

Peterson, Cameron R and Lee Roy Beach, "Man as an intuitive statistician.," *Psychological bulletin*, 1967, *68* (1), 29.

Peterson, Christopher, "Recognition of noncontingency.," *Journal of Personality and Social Psychology*, 1980, *38* (5), 727.

Rottman, Benjamin M and Frank C Keil, "Causal structure learning over time: Observations and interventions," *Cognitive psychology*, 2012, *64* (1-2), 93–125.

Rottman, Benjamin Margolin, "The acquisition and use of causal structure knowledge," *The Oxford handbook of causal reasoning*, 2017, pp. 85–114.

_ **and Reid Hastie**, "Reasoning about causal relationships: Inferences on causal networks.," *Psychological bulletin*, 2014, *140* (1), 109.

Shanks, David R and Anthony Dickinson, "Instrumental judgment and performance under variations in action-outcome contingency and contiguity," *Memory & Cognition*, 1991, *19* (4), 353–360.

Sher, Varshita, Karen G Bemis, Ilaria Liccardi, and Min Chen, "An empirical study on the reliability of perceiving correlation indices using scatterplots," 2017.

Smedslund, Jan, "The concept of correlation in adults," *Scandinavian journal of Psychology*, 1963, *4* (1), 165–173.

Steyvers, Mark, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum, "Inferring causal networks from observations and interventions," *Cognitive science*, 2003, *27* (3), 453–489.

Thorndike, Edward L, "Animal intelligence: An experimental study of the associative processes in animals.," *The Psychological Review: Monograph Supplements*, 1898, *2* (4), i.

Tversky, Amos and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, 1974, *185* (4157), 1124–1131.

Ward, William C and Herbert M Jenkins, "The display of information and the judgment of contingency.," *Canadian Journal of Psychology/Revue canadienne de psychologie*, 1965, *19* (3), 231.