Testing one or multiple: How beliefs about sparsity affect causal experimentation

Anna Coenen[1]*, Azzurra Ruggeri[2], Neil R. Bramley[1], and Todd M. Gureckis[1]

[1]Department of Psychology, NYU, New York; [2]MPRG iSearch, Max Planck Institute for Human Development, Berlin;

* Corresponding author. Please contact at coenen.anna@gmail.com.

Abstract

What is the best way of discovering the underlying structure of a causal system composed of multiple variables? One prominent idea is that learners should manipulate each candidate variable in isolation to avoid confounds (sometimes known as the "Control of Variables" or CV strategy). We demonstrate that CV is not always the most efficient method for learning. Using an optimal actor model which aims to minimize the average number of tests, we show that when a causal system is *sparse* (i.e., when the outcome of interest has few or even just one actual cause among the candidate variables) it is more efficient to test multiple variables at once. Across a series of behavioral experiments, we then show that people are sensitive to causal sparsity and adapt their strategies accordingly. When interacting with a dense causal system (high proportion of actual causes among candidate variables), they use a CV strategy, changing one variable at a time. When interacting with a sparse causal system they are more likely to test multiple variables at once. However, we also find that people sometimes use a CV strategy even when a system is sparse.

Keywords: control of variables; interventions; experimentation; causal learning; hypothesis testing

Introduction

To develop a causal understanding of the world, we often need to find out how multiple candidate variables affect an outcome of interest. This problem arises in everyday situations (e.g., "Which switch(es) control the bathroom fan?"), during scientific exploration ("Which of these treatments can affect disease x?"), and plays an important part in answering economic and social questions ("What is the impact of these policies on Gross Domestic Product?"). Often, the quickest and most effective method of resolving causal relationships is to conduct experiments that manipulate variables of a system (e.g., turning switches on or off) and to observe the resulting outcome. This kind of causal experimentation is often (but not always) required to decouple causation and correlation (Pearl, 2009; Woodward, 2005; S. Sloman, 2005).

Both children and adults can systematically leverage the outcomes of interventions to test causal hypotheses that would be indistinguishable based on observation alone (Lagnado & Sloman, 2004; Lagnado, Waldmann, Hagmayer, & Sloman, 2006; Rottman & Keil, 2012; Schulz, Gopnik, & Glymour, 2007; S. A. Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). Furthermore, people are sometimes able to come up with highly efficient experiments that optimize information gained per intervention or minimize the total number of tests needed on average to discover the true causal structure (Bramley, Lagnado, & Speekenbrink, 2015; Bramley, Dayan, Griffiths, & Lagnado, 2017; Coenen, Rehder, & Gureckis, 2015; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003).

Here we consider a specific learning situation in which people are asked to explore a causal system consisting of a number of independent variables (switches) and a dependent outcome (turning on a fan). These types of problems have played a central role in research on science education and cognitive development and are common in everyday experience (Inhelder & Piaget, 1958; Chen & Klahr, 1999; Kuhn & Brannock, 1977; Kuhn, Iordanou, Pease, & Wirkala, 2008). An example study in this area might consider how children and adults manipulate variables (such as water, fertilizer, or sunlight) that affect the health of a plant (see Klahr, Fay, & Dunbar, 1993). An important focus of science education research has been to teach basic principles of how learners *should* approach such problems in general. Educators have specifically focused on teaching students the principle of *isolating or controlling variables* (i.e., the idea that variables should be tested individually while holding everything else constant, Kuhn & Brannock, 1977). As we review below, isolating variables represents a general strategy for approaching many types of causal learning problems and often results in non-confounded evidence.

In this paper, we take a broader perspective on multivariate experimentation and consider under what circumstances testing individual variables is more or less effective. In particular, through an analysis of an optimal actor model, we show that the most efficient strategy for understanding a multivariate causal system critically depends on the proportion of causally relevant variables (which we will refer to as *causal sparsity*). The key take away from our analyses is that in causally sparse environments, where the proportion of causes among the candidate variables is low, changing multiple variables at once is more effective than controlling variables individually. We demonstrate across four experiments that naive learners adapt their intervention strategies in line with these computational predictions. We also highlight interesting ways in which people deviate from the optimal actor benchmark. Indeed, despite people's ability to adapt their strategies, we consistently find a group-level bias towards controlling individual variables even when changing multiple is more efficient. We also find that some participants seek out confirmative evidence—i.e., they make interventions they should already know the outcome of (e.g., Klayman, 1995)—and favor outcome-positive evidence—i.e., they make interventions they expect to produce the effect rather than something equally informative that they expect to not produce the effect— in line with a *positive testing strategy* (e.g., Coenen et al., 2015; Klayman, 1995; McKenzie, 2004; White, 2009). We discuss what these deviations suggests about strategy selection during causal experimentation.

Learning through experiments

We start by describing the two main strategies that are considered in the paper, before turning to the optimal actor analysis.

Test one variable at a time

The ability to learn about the effects of multiple variables has long been considered a hallmark of mature logical and inductive reasoning. Beginning with Inhelder and Piaget (1958), researchers have been particularly interested in the development of what is often called the *control of variables* (CV) strategy (for a recent review on the CV principle, see Zimmerman, 2007). This is an epistemic strategy where learners systematically test the effect of every variable in isolation to avoid confounding evidence by changing one variable at a time. For example, to find out what factors affect the health of a plant, the CV principle prescribes that one should change a single variable, say, the watering regime, without changing the amount of fertilizer, lighting, and humidity (Kuhn & Brannock, 1977). Although this is not always made explicit, the success of this strategy requires that the unchanged variables are held at their default values (e.g., same level of fertilizer and light) and not left to vary freely.

In the education literature, considerable emphasis has been placed on teaching children the CV principle (e.g., Chen & Klahr, 1999; Kuhn & Brannock, 1977; Kuhn & Angelev, 1976). In fact, its normative status is so pervasive that it features as one of the assessment criteria in national standards for science education (e.g., see *Next Generation Science Standards*, 2013, p.52).

A common finding from empirical studies is that children require extensive training to acquire the CV principle and teaching them to transfer it to novel tasks is an even bigger challenge (e.g., Kuhn et al., 1995; Klahr et al., 1993; Kuhn & Phelps, 1982). Adults and adolescents, although more likely to use the strategy spontaneously, still show a tendency to test multiple features at once instead of testing them individually (Kuhn et al., 1995). Interesting exceptions have been found in more complex tasks. For example, Bramley et al. (2017) allowed adults to fix any subset of the variables of a multivariate system and observe the consequences on the other variables with the goal of finding out the underlying causal structure. In this task, the most informative tests — according to an optimal model — typically involved leaving most variables uncontrolled. However, participants often chose to test one causal relationship at a time by holding most variables at a constant value. Note that, participants might have been more likely to test individual variables in this rather complex task (with a vast hypothesis space), because the need to reduce the cognitive load was particularly severe.

In sum, CV is a widely regarded epistemic principle for learning about causal systems composed of multiple variables. Mastery of this principle is often equated with cognitive maturity and accomplishment within Science, Engineering, Technology, and Mathematics (STEM) curricula. A key advantage of a CV strategy is that it results in unconfounded data that is easy to interpret.

Going forward, we will consider a particular instantiation of the CV principle that applies to multivariate causation. We will refer to it simply as the *Test One* strategy, which we define as the sequential causal activation of individual variables (potential causes of some outcome), while all other potential causes are kept inactive to prevent any confounding impact on the outcome. We will give a more formal definition of this strategy below.

Test half or test multiple variables

Changing variables one-by-one has the benefit of isolating the effect of every variable without the confounding influence of the others. It is therefore particularly helpful when one believes that many variables could potentially affect the outcome. However, consider the case in which a learner expects only very few, or perhaps just a single variable, to affect the outcome, and is faced with a number of equally plausible candidate variables. In that case, an alternative strategy is to *test multiple variables* at once, to see if any of them affects the outcome at all. For example, imagine trying to figure out which out of 20 switches in a poorly labeled basement fuse box controls the bedroom light. In this case, a possible strategy for identifying the correct switch is to turn on half (10 out of 20) of the switches to find out which half contains the target switch, and then continue testing half of the remaining switches until only one remains. Compared to testing switches one-by-one, this strategy will reduce the number of basement trips.

The *Test Multiple* or (more specifically) *Test Half* strategy has been studied by psychologists in a different type of information-seeking task, based on the popular "Twenty Questions" game. In this task, children or adults have to identify a target object or person among a given set by asking as few yes/no questions as possible. Here, too, the optimal strategy (in terms of expected information gain, see next section) is to ask questions targeting features that apply to *half* the possibilities under consideration (e.g., "Is the person female?", if the hypotheses are people and half are each sex), because it reduces the number of alternatives more rapidly than asking about specific identities directly (e.g., Navarro & Perfors, 2011; Mosher & Hornsby, 1966). In analogy to the switch testing example, inquiring about a feature means asking if any of the individuals sharing that feature is the target, just as turning on multiple switches asks if any of those switches has an effect on the outcome (the fan). An alternative approach to asking about shared features would be to test each person individually (e.g., "Is it Bob?"). This is similar to a Test One strategy that turns on one switch at a time. In experiments using versions of the Twenty Questions game, both adults and, to a lesser degree, children have been shown to be able to use the Test Half method successfully (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri & Feufel, 2015; Ruggeri, Lombrozo, Griffiths, & Xu, 2016; Ruggeri & Lombrozo, 2015). Analogous "divide-and-conquer" strategies have also emerged in the problem solving literature, for example the symmetry strategy described in the n-ball problem (Simmel, 1953; Ormerod, MacGregor, Chronicle, Dewald, & Chu, 2013).

Importantly, using a Test Half (or Test Multiple) strategy is sometimes considered a reasoning error (particularly in work on Control of Variables in the science education literature). For example, a child that chooses a test that simultaneously manipulates a plant's light exposure, and fertilizer, would be recorded as low-performing in a classic CV experiment (Klahr et al., 1993), because changing or setting many variables at once is thought to confound individual variables.

The Test One and Test Half/Multiple strategies are typically studied in different kinds of psychological tasks. However, as demonstrated by the switch example, they can both be reasonable approaches for testing the causal impact of multiple variables. Next, we show how the effectiveness of each strategy depends on the structure of the environment.

Sparsity determines effectiveness of learning strategies

As the switch example shows, a crucial factor determining the effectiveness of a Test One or a Test Half strategy is the *sparsity* of a causal system. We define sparsity as inversely related to the proportion of variables causally affecting the outcome. For example, when only one of the 20 switches controls the bathroom light (a sparse environment), a learner can quickly narrow in on the target switch by trying many variables at once. In contrast, where there are many effective causes (a dense environment), changing multiple at once will tend to be ineffective because the outcome will almost always be produced and little will be learned about which variable(s) are actually responsible due to the confound. Thus, the choice of an effective testing strategy in a situation is a question of ecological rationality in that it depends to a large degree on the structure of the environment (Todd & Gigerenzer, 2012; Gigerenzer, Todd, & the ABC Research Group, 1999). Past work has shown that people behave in an ecologically rational fashion when sparsity is varied in a number of non-causal hypothesis testing tasks (where sparsity applied to hypotheses or events, e.g., McKenzie, Chase, Todd, & Gigerenzer, 2012; Hendrickson, Navarro, & Perfors, 2016; Oaksford & Chater, 1994; Langsford, Hendrickson, Perfors, & Navarro, 2014; Navarro & Perfors, 2011). For example, Hendrickson et al. (2016) showed that people switch from requesting positive to negative examples of a concept when the overall proportion of positive cases increases. In the next section, we formally show why and how they should do so in causal scenarios as well.

Modeling the effect of sparsity with an optimal actor model based on Expected Information Gain. Assume that a learner is faced with a simple causal system with N binary independent input variables, I, and a single binary outcome, o. Given the subset of input variables, $C \subseteq I$ that, when active, can causally affect the outcome, the probability of the outcome given the current setting of input variables is

$$P(o = 1|C) = \begin{cases} 1, & \text{if } \exists\ c \in C \wedge (c = 1), \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In other words, the outcome occurs if and only if any of the input variables in C are currently active (this is equivalent to an *inclusive OR* relationship between causes and the outcome).

The learner must now decide how to manipulate the input variables to figure out which of them are causally relevant (that is, which variables are members of C). We assume that the learner's optimal strategy lies in choosing a switch setting, $s \in S$, that maximizes the expected *gain in information* with respect to the system.

Expected Information Gain is a common metric for quantifying the value of information-seeking actions, including causal interventions (see, Oaksford & Chater, 1994; Steyvers et al., 2003).[1] It is computed as the expected reduction in uncertainty over the

[1] While there are a range of possible measures of information (Nielsen & Nock, 2011), they disagree about the best testing strategy only in specific fringe cases that do not apply in the current context (Bramley, Nelson, Speekenbrink, Crupi, & Lagnado, 2014).

hypotheses H, after having made an intervention on the system and observed an outcome. Here, the learner's hypotheses are possible sets causally relevant variables, that is, $H = \{C_1, ..., C_m\}$. We are considering the simple case of binary outcomes ($o = 1$ or $o = 0$) with the likelihood of an outcome given by Equation (1). A learner's expected information gain is:

$$\mathrm{EIG}(s|H) = \mathrm{SE}(H) - \sum_{j=0}^{1} P(o = j|s)\ \mathrm{SE}(H|s, o = j), \tag{2}$$

where SE denotes the Shannon Entropy over a distribution of hypotheses (Shannon & Weaver, 1949), which are possible sets of causes in this application. The prior Entropy is

$$\mathrm{SE}(H) = -\sum_{i=1}^{m} P(C_i) \log P(C_i) \tag{3}$$

The updated belief for each element in H after observing an outcome follows Bayes' rule,

$$P(C_i|o) = \frac{P(o|C_i)P(C_i)}{\sum_{j=1}^{m} P(o|C_j)P(C_j)} \tag{4}$$

and the Shannon Entropy over the new set of hypotheses is is

$$\mathrm{SE}(H|s, o) = -\sum_{i=1}^{m} P(C_i|o) \log P(C_i|o) \tag{5}$$

This model is myopic, in that it only optimizes the Expected Information Gain of the *next* action without simulating additional future actions and outcomes (see below for further discussion).

To model the impact of sparsity on the model predictions, we varied the number of causes ($|C|$), and the number of total variables, N. Note that these quantities affect the EIG computation in Equation 2 by constraining the hypothesis set H. For example, if $|C| = 2$ and $N = 6$, then H contains all possible combinations of two causes in six switches, which yields 15 total hypotheses (the number of hypotheses is always $\binom{N}{|C|}$). Each hypothesis corresponds to a different set of causes, such that $H = \{\{1, 2\}, ..., \{5, 6\}\}$. We furthermore assumed a flat prior belief over all hypotheses in H, that is, each hypothesis has a prior probability of $P(C_i) = 1/|H|$.

Figure 1 shows model predictions for the number of variables this EIG-optimal actor should manipulate given different values of $|C|$ and N. The model manipulates multiple variables on the first trial (in fact, exactly $\frac{N}{2}$) when it expects only a single cause. As the number of causes increases (that is, as causal density increases), the optimal number of variables to be manipulated decreases and quickly converges to the Test One strategy. This relationship is modulated by N, which effectively decreases the degree of causal density (provided $|C|$ is constant) and consequently the number of variables that should be manipulated.

These results show that knowledge about the causal sparsity of an environment should affect a learner's strategy for manipulating binary variables to find out how they affect an

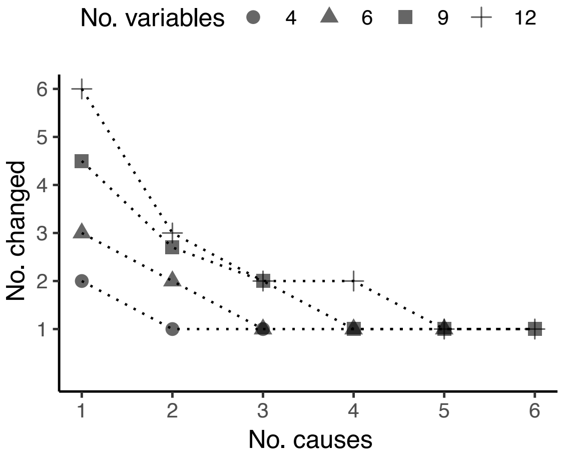

Figure 1. Effect of the number of causes ($|C|$) on the number of variables tested by an EIG-optimal actor, given different numbers of potential causes (N).

outcome of interest. This means that, even within the same kind of task, there exists a continuum of optimal strategies with respect to the number of variables to be manipulated that ranges from Test Half to Test One.

Myopic EIG vs. optimal planning. In the experiments reported in this paper, participants are given monetary incentives aligned with the goal of making efficient interventions (Brier, 1950). Specifically, they receive a fixed payoff for identifying causally relevant variables and have to pay a small cost every time they test the system.

This means that, although the EIG analysis presented here is optimal in maximizing informational value for the next trial, it does not explicitly maximize monetary reward that can result from a sequence of tests. Doing so requires a forward-looking model that takes into account the costs of additional tests and the reward for finding the correct solution. We derive predictions from such an *optimal planning* model in the Appendix.

The biggest divergence between optimal planning and myopic EIG is that the former recommends a broader number of possible strategies in sparse environments, including changing slightly fewer and slightly more variables than half. It also makes predictions for when a learner should stop making tests and guess the answer even if entropy is still nonzero. Evidence suggests that people do not typically plan multiple steps into the future when collecting information in causal systems (e.g., Bramley et al., 2015). Further, past studies suggest that even with strongly misaligned monetary and informational incentives, a large amount of task experience is required before people will maximize a monetary incentive over accuracy (Meder & Nelson, 2012; Markant & Gureckis, 2012). We therefore chose to focus on the EIG analysis for the main body of this paper. To foreshadow the results reported later, we also do not find that the optimal planning model captures behavior better than myopic EIG. We will address the differences between myopic EIG and optimal planning where appropriate in the results and discussion sections of each experiment. Crucially, we should note that, for all the experiments reported in this paper, the myopic EIG strategy is always among the set of actions that the optimal planning model recommends. That is, in sparse environments, Test Half is optimal under both models.

Throughout the rest of this paper we will refer to the myopic EIG model as the *EIG optimal actor model*, or simply EIG model. We will also occasionally use the term "optimal learner", which refers to an accurate model of belief updating (i.e. computing $P(H|o)$) in line with Bayes' rule in Equation (4).

Overview of Experiments

The results of the modeling presented in the previous section lead us to the core questions of this paper: How do people manipulate variables in multivariate single-outcome systems and how do beliefs about causal sparsity affect their inquiry strategies?

To make our predictions explicit, based on the EIG-optimal actor predictions presented in Figure 1, we hypothesize that when learning about a causal system, people will use different strategies depending on their beliefs about the sparsity of the system. When only a few of the candidate variables are causally effective, we predict that people will test multiple variables. When many of the candidate variables are causally effective, we expect people will test items individually. This result would offer a further demonstration that human intervention strategies are ecologically rational, in the sense of being well matched to the environment where they are implemented (Todd & Gigerenzer, 2012; Parpart, Jones, & Love, 2018).

In the following sections we present four experiments that investigate how sparsity affects people's causal experimentation strategies. Sparsity is manipulated in two ways, both suggested by the model results shown in Figure 1. We first vary the number of causes (i.e., variables that affect the outcome) in a system (Exp. 1 & 2) and second the number of total variables available for testing (Exp. 2). We also investigate what strategies people select given an unconstrained prior belief (when the degree of sparsity is not specified, Exp. 3). Finally, we test whether repeated interaction with sparse systems encourages learning of efficient strategy selection (Exp. 4).

Experiment 1 - manipulating number of causes

In the first experiment, participants were presented with a simple multivariate causal system consisting of a box (see Figure 2) that held a number of variables (switches) that influenced the outcome (a spinning wheel). Participants' goal was to figure out how the system worked by manipulating the variables (turning switches on/off) and then testing the effects of this manipulation. Sparsity was manipulated between-subjects by changing participant's a-priori belief about the number of causes (i.e., working switches) among the variables. In the *sparse* condition participants were told that only one switch was working (only one cause of the outcome). In the *dense* condition they were told that all but one switch were working (many individually sufficient causes of the outcome). The goal of the experiment was to assess whether and how sparsity affects people's testing strategies. The key dependent measure was the number of switches people manipulated on each test of the system to figure out how the box worked.

The goal of this first experiment was to pilot the experimental paradigm used throughout this paper in a lab setting and make sure that participants understood the main manipulation (of sparsity).

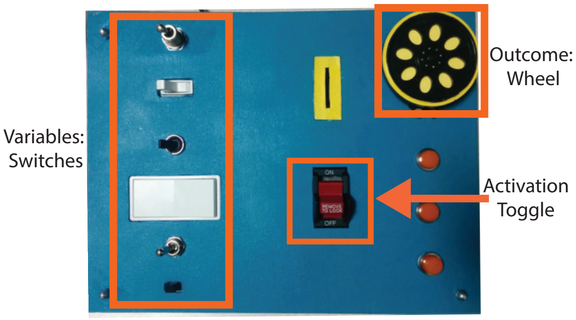

Figure 2. A photograph of the control interface on the wooden box used in Experiment 1.

Method

Participants. Thirty participants (15 males; $M_{age} = 23, SD_{age} = 4$) were recruited via the subject pool of New York University's Department of Psychology. Participants were paid at a rate of $5 per hour and could win an additional bonus of up to $3 (see below). Because this study was very short (it took around 5 minutes per participant), it was conducted with participants who had completed an unrelated memory experiment immediately prior to this one. The sample size was chosen to accommodate the number of participants in the memory experiment. Approval for this study was obtained by New York University's Institutional Review Board (IRB) under the protocol "Active Learning in Dynamic Task Environments" (IRB-FY2016-231).

Design and apparatus. Participants were presented with the wooden box depicted in Figure 2. The dimensions of the box was approximately 35cm by 25cm. All of the sides of the box were painted blue. The top of the box had a number of vintage electronic components arranged as shown in the photograph. The motivation behind using a physical box instead of a digital computer was that we planned to run a similar experiment with children. However, subsequent adult experiments were run online and we did not observe any systematic differences in behavior (see Experiment 2).

The box had six different toggle switches (variables), a yellow wheel (outcome), and a red activation toggle. Each switch could be turned to the left (off) or the right (on). The yellow wheel could either spin (outcome present) or not spin (outcome absent). The activation toggle controlled whether the box was currently active (if inactive, the outcome could never occur). A row of three lights along the lower left side of the box would turn on when the box was activated by the activation toggle. In addition a yellow slot above the activation toggle provided a place for people to insert "tokens" into the box.

Participants were randomly assigned to one of two experimental conditions. In the *sparse* condition, participants were told only one of the switches caused the wheel to spin, whereas the remaining five switches were broken. In the *dense* condition, participants were told five switches caused the wheel to spin and one switch was broken. A single working switch was sufficient to activate the wheel, and the position of the broken switches had no effect whatsoever.

In both conditions, the wheel could only be activated if the activation toggle was currently in its *on* position. Otherwise, participants were told that the box was turned off. Thus, participants experimented with the system by first setting the variable switches in different ways and then setting the activation toggle to *on* to see what happened to the wheel. Which exact switches were broken or working was randomly determined for each participant via a microcomputer hidden inside the box. At the beginning of the experiment, participants were given six plastic tokens, each of which was worth $0.50. Participants had to pay one token every time they wanted to turn on the box via the activation toggle (see below) by inserting the coins into the yellow coin slot.

Procedure. Participants were first familiarized with the components on the box through verbal explanation of the experimenter. They were told about the binary (left=off, right=on) nature of the switches, and the difference between broken and working switches. Depending on the condition, participants were then told that they had to identify the one working switch (*sparse* condition) or the one broken switch (*dense* condition). Before starting the task, participants in both conditions were shown the same two demonstration trials. First, while the activation toggle was turned off, the experimenter turned all six switches to their *on* position and subsequently turned on the activation toggle, causing the wheel to spin. Second, after turning the activation toggle off again, the experimenter set all switches to their *off* state and turned the activation toggle back on, which did not cause the wheel to spin.

In the main part of the experiment, participants could repeatedly test different settings of the switches to find out which one was broken/working. On each trial, they could change the switches in any way they liked while the activation toggle was turned off. They could then test their chosen switch setting by turning the activation toggle on and observing the effect on the wheel. Before the start of each new trial, the activation toggle had to be turned off again.

To incentivize participants to be efficient (i.e., to use as few trials as possible), they had to pay one of their six plastic tokens (worth $0.50 each) each time they performed a test by inserting it into a coin slot on the box. Participants could test the box up to six times (hence the use of six tokens), but could stop whenever they were ready to make their judgment. After their final test, they had to indicate to the experimenter which of the switches they thought was broken/working. If their choice was correct, they could trade in any remaining tokens for their corresponding monetary value. If their choice was incorrect or they used up all their tokens, they received no bonus.

Results and discussion.

Performance. Before comparing strategies in detail, we evaluated the overall performance in the task. Final accuracy in identifying the working/broken switch at the end was 100% in the sparse condition and 80% in the dense condition. This difference was not statistically significant (Fisher's exact test, 95% CI $[0.43, \infty], p = 0.22$). On average, participants made 3.6 ($SD = 1.45$) interventions in the sparse condition, compared to 4.6 ($SD = 1.92$) interventions in the dense condition. Note that this difference is in line with the predictions of the EIG analysis reported above (the EIG-optimal strategy in the dense condition requires more steps), but given the number of participants in this experiment this was not statistically significant ($t(28) = 1.61$, 95% CI$= [-0.27, 2.27], p = 0.12$).

Strategy classification. To characterize participants behavior over multiple trials, we used the same strategy classification scheme across all experiments. For a sequence of example trials that would have been classified as each strategy, see Table 1. Note that in the upcoming strategy definitions we will use the term *zero EIG* to talk about trials in which the participants' test could not yield any additional information according to the Expected Information Gain equation described above (Equation (2)). Zero EIG could result, for example, from a participant testing the same switch twice or turning on multiple variables in the dense condition (which would yield any additional information since the outcome always occurred). Furthermore, we will use the term *potential causes* to designate variables that, according to the EIG-optimal actor, could still be causes of the outcome. Conversely, *noncauses* are variables that a participant has ruled out through prior tests.

- *Test One*: The participant followed the Controlling Variables principle and tested a single switch on every trial, while turning off any potential causes.
  - *Pure Test One*: The participant used Test One on all trials.
  - *Noisy Test One*: The participant used Test One with interspersed zero EIG trials.
- *Test Multiple* (sparse condition only): The participant turned on several switches on every trial. Importantly, *this strategy could only be used in the sparse condition*, because manipulating more than one switch in the dense condition always led to zero EIG.
  - *Pure Test Half*: Participant manipulated exactly half of the remaining potential causes on every trial (rounding odd numbers up or down).
  - *Pure Test Multiple*: The participant turned on several (but not always exactly half) of the remaining potential causes on every trial. Note that it was possible that a participant in this category used a forward looking strategy, which does not require testing exactly half (see Appendix A).
  - *Noisy Test Multiple*: The participant used a Test Half or Test Multiple strategy but with interspersed zero EIG trials.

  Note that with an odd number of potential causes, both rounding up and down from the central value gets counted as a viable testing Half/Multiple strategy. That means that on a trial with *three* remaining variables, changing *one* can actually be part of the Test Half/Multiple strategy (see the last trial of *Pure Test Multiple* in Table 1, for example).
- *Other*: Any strategy that does not fall into the above categories. also included participants who switched back and forth between Testing One and Testing Multiple. For example, this category would include participants who started testing variables one-by-one and then changed their strategy to changing half of the variables, and vice versa.

The goal of these definitions was to strike a balance between accurately summarizing our participants' behavior and parsimony in terms of the number of categories we assigned. Throughout this paper, we will report additional behavioral markers of strategy use in addition to these categories to ensure we provide a comprehensive picture of the data.

Table 1

Example sequences of the first three trials for different strategies given six binary variables. Trials on which the outcome occurred are denoted with an asterisk (∗). The outcome of the third trial has no impact on the actions depicted in this table.

Strategy	Example
Pure Test One – Test One on all trials	$\begin{array}{cc} & \begin{array}{ccc} t1{*} & t2{*} & t3? \end{array} \\ \begin{array}{c} s1 \\ s2 \\ s3 \\ s4 \\ s5 \\ s6 \end{array} & \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{array} \right] \end{array} \ldots$
Noisy Test One – Test One with interspersed 0-EIG trials	$\begin{array}{cc} & \begin{array}{ccc} t1{*} & t2{*} & t3? \end{array} \\ \begin{array}{c} s1 \\ s2 \\ s3 \\ s4 \\ s5 \\ s6 \end{array} & \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \end{array} \ldots$
Pure Test Half – Test Half on all trials	$\begin{array}{cc} & \begin{array}{ccc} t1- & t2{*} & t3? \end{array} \\ \begin{array}{c} s1 \\ s2 \\ s3 \\ s4 \\ s5 \\ s6 \end{array} & \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{array} \right] \end{array} \ldots$
Pure Test Multiple – Test Multiple on all trials	$\begin{array}{cc} & \begin{array}{ccc} t1- & t2{*} & t3? \end{array} \\ \begin{array}{c} s1 \\ s2 \\ s3 \\ s4 \\ s5 \\ s6 \end{array} & \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{array} \right] \end{array} \ldots$
Noisy Test Multiple – Test Multiple with interspersed 0-EIG trials	$\begin{array}{cc} & \begin{array}{ccc} t1- & t2{*} & t3? \end{array} \\ \begin{array}{c} s1 \\ s2 \\ s3 \\ s4 \\ s5 \\ s6 \end{array} & \left[\begin{array}{ccc} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{array} \right] \end{array} \ldots$

Note that participants were classified based on their sequence of tests up to the point at which an optimal learner would have been able to correctly identify the working or broken switch (i.e., the point at which the Expected Information Gain from Equation 2 was zero for every kind of test). Some participants made further unnecessary tests that we report and analyze in separate parts of the result sections.

Strategy. Table 2 shows the number of participants classified into all strategy types described in the previous section. Note that not all of the strategy types appeared in this experiment (there were no Test Multiple or Noisy Test Multiple participants).

To simplify analyses, we further grouped participants as either using a Test One (includes Test One and Noisy Test One), Test Multiple (includes Test Half, Test Multiple, and Noisy Test Multiple), or Other strategy. Figure 3 shows the number of participants in each of those broader strategy categories. The number of participants using a Test One strategy was lower in the sparse condition (4 in 15 vs. 14 in 15, Fisher's exact test, 95% CI [0.0006, 0.31], $p < 0.001$). However, even in the sparse condition around a quarter of the participants decided to change one variable at a time.

Note that all of the participants in the Test Multiple group of the sparse condition actually manipulated exactly half of the switches (see Table 2). This finding supports the fact that the myopic EIG model is a better description of behavior than the optimal planning model that is briefly mentioned above (see Appendix A for details). Indeed, although the optimal planning model assigns equal value to manipulating 2, 3, and 4 variables on the first trial of Experiment 1, we only observe people manipulate exactly half.

Stopping. We found that one third ($\sim 33\%$) of participants made at least one additional intervention at the point when, according to the EIG-optimal actor, they should have identified the solution already. The number of such "unnecessary" tests was higher for participants in the sparse group ($\sim 47\%$ compared to 20% in the dense group), but this difference did not reach statistical significance (Fisher's exact test, 95% CI [0.55, 26], $p = 0.245$). The sample size of this experiment makes it unfeasible to look for factors that affect whether participants with no remaining uncertainty made additional interventions. We will return to this question in the larger-N experiments reported below.

When modeling each participant's sequence of tests, it was revealed that only 60% of participants in the dense group had completely resolved their uncertainty (had zero Shannon Entropy according to (5)) about the broken switch by the time they stopped conducting tests and made a choice. Conversely, in the sparse group all participants had resolved their uncertainty completely (the difference between conditions was significant; Fisher's exact test, 95% CI $[1.5, \infty]$, $p = 0.017$). Out of the six participants who stopped before reaching certainty in the dense group, four did so because they used a noisy Test One strategy and had reached the last trial (trial six), which meant they had to stop because they used up all the given tokens. One participant stopped after only two tests with a high risk of making an incorrect guess. Finally, the last participant among the early stoppers chose to guess with only two candidate switches remaining (thus with a 0.5 chance of guessing correctly), which is actually among the optimal stopping strategies predicted by an optimal planning model (see Appendix A).

In sum, as predicted by the EIG-optimal actor model, this experiment found that instructing participants to expect either a sparse (one cause in six variables) or a dense (five causes in six variables) environment had an effect on how they manipulated the set of six variables. This confirms that people's prior beliefs about the sparsity of their environment can induce changes in strategy selection. However, another intriguing result of this experiment is that even in the sparse condition a proportion of participants adhered to the Controlling Variables principle and used a Test One strategy. We will address this finding in the results and discussions of the remaining experiments (particularly Exp. 3 & 4).

Because this experiment was designed to test in-lab instructions and the sparsity manipulation, we lacked sufficient sample sizes for some of the analyses based on subsets of participants. However, this study still revealed some trends indicating potentially interesting patterns in stopping behavior that differed between conditions. In particular, participants in the sparse group were more likely to make further, unnecessary tests. We will continue exploring these differences in the next experiments.

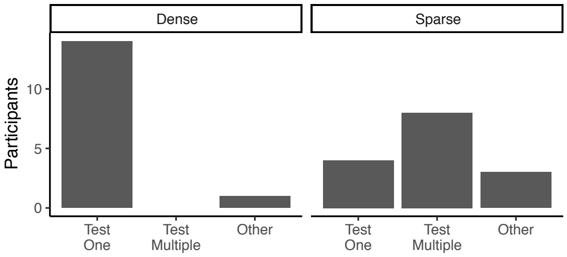

Figure 3. Strategy classification in Experiment 1.

Table 2
Detailed strategy use in Experiment 1.

condition	Pure Test One	Noisy Test One	Pure Test Half	Other
dense	7	7	0	1
sparse	3	1	8	3

Experiment 2 - manipulating the number of variables

In addition to the number of causes, our computational analyses above show that the total number of variables affects the sparsity of causal systems. In Experiment 1, the benefit of testing multiple variables over testing variables one-by-one was relatively modest. In fact, testing half of the variables in the sparse condition would save participants less than one step (2/3 of a step) on average, compared to testing variables individually (this difference translated to an average saving of ~$0.33). This may not have provided sufficient incentive for participants to realize that a Test Half strategy would be more advantageous. As discussed above, one way to amplify the potential impact of the sparsity manipulation is to include more variables (see Figure 1). Figure 4 shows the average number of trials (causal tests) needed to find the working switch for a learner in a sparse (one cause) environment employing either a Test One or a Test Half strategy depending on the number of switches available. We can see that as the number of switches increases, so does the benefit of the Test Half strategy over the Test One strategy.

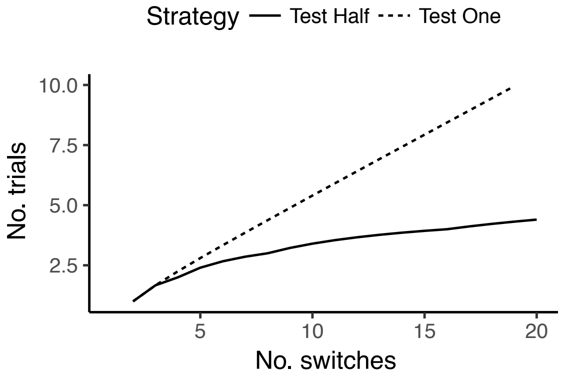

Figure 4. Expected number of trials needed to find the working switch in the sparse condition, when using a Test One or Test Half strategy.

To test if people are sensitive to the degree of sparsity, Experiment 2 manipulated the number of variables (switches). Participants on Amazon Mechanical Turk completed the same task as in Experiment 1 (modified for presentation over the web), but were presented with either 4, 6, 10, or 20 switches (all manipulations were between-subjects). As before, they were given either sparse (one switch working) or dense (all but one working) instructions (see Appendix B for details). Although adding variables should have no effect on behavior in the dense condition, we decided to keep the manipulation to ensure that adding variables does not encourage a general increase in the number of variables participants would test on each trial. By including the six switches condition again, this experiment also served to replicate the results from Experiment 1 with an online sample.

Method

Participants. One hundred and thirty one participants (73 males) were recruited on Amazon Mechanical Turk (15 to 18 per cell).[2] Recruitment was restricted to AMT workers within the United States aged 18 or above. Participants were paid $0.50 for their participation, with the possibility of earning an additional bonus of up to $1 (see below). Approval for this study was obtained by New York University's Institutional Review Board (IRB) under the protocol "Active Learning in Dynamic Task Environments" (IRB-FY2016-231).

[2] Sample size was chosen to match the previous experiment (fluctuation in sample size resulted from Mechanical Turk drop-outs), because it was sufficient to detect differences between the sparse and dense groups in terms of strategy use. Furthermore, the data for some analyses (e.g., stopping, see below) could be pooled across conditions, which is why we still anticipated benefitting from greater power based on the observed in this experiment compared to Experiment 1. A posthoc power calculation suggested by a reviewer, of the nonsignificant main effect of condition on number of interventions in Experiment 1 yielded .37. This indicates the study may have been somewhat underpowered for detecting this particular effect. The power of Experiment 2 to find the commensurate effect in its larger sample (pooling across number of switches) was .94. Since we expect this effect to be larger for devices with more switches and the setting in Experiment 1 was at the low end of the range, this is a conservative estimate. Participants indicated their age, but it was not recorded due to a coding error.

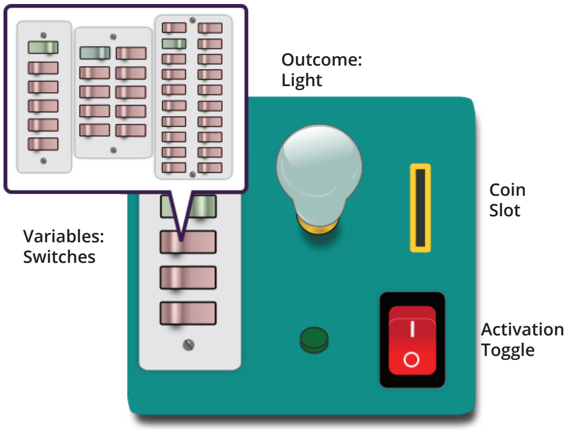

Figure 5. Switchboard presented to participants in Experiments 2-4. Experiment 2 varied the number of switches on the board (4, 6, 10, or 20), as shown above.

Stimuli. The task from Experiment 1 was adapted as faithfully as possible to be run on the web with some minor changes (See Figure 5). Again, switches could be turned on (green) or off (red). Instead of a wheel, the outcome of interest was a light bulb, which lit up when it was turned on, and remained gray otherwise. The red activation toggle needed to be in its “on” position for the switchboard to work. When the switch-box was turned on, the green indicator light to the left of the activation toggle shone bright green. The coin slot on the top right corner of the switchboard would show a brief animation of a coin being inserted whenever a participant made an additional test (see below).

Procedure. The experiment followed a 4 x 2 between-subjects design. Participants received different versions of the task with either 4, 6, 10, or 20 switches (as shown in Figure 5), and were given either the sparse or the dense instructions.

The procedure was the same as in Experiment 1. Participants received similar instructions (but written) and were also asked to perform two demonstration trials in which first all and then none of the switches were turned on, to show that the light bulb would turn on and stay off, respectively. The per-trial payment was adjusted depending on condition, such that participants had to pay either $0.25, $0.16, $0.1, or $0.05 per additional test in the 4, 6, 10, or 20 switches conditions, respectively. These payments were chosen so that the total potential bonus (starting at $1) would be zero if participants decided to test every single switch in isolation. The remaining bonus was shown to participants to the right side of the switchboard.

At the beginning of a trial, participants could click on as many switches as they wanted to turn them on or off. They then had to turn the activation toggle on to observe the effect of the light bulb. Every time they turned on the activation toggle and saw the outcome, the cost of this test was automatically deducted from the total bonus shown on the side of the screen. Participants could then either make another test or click a button to proceed to the final choice. Before each new test, they had to turn the activation toggle off again. In the choice phase, participants were asked to click on the one switch that was broken/working and confirm their choice before receiving feedback.

Results.

Table 3

Detailed strategy use in Experiment 2.

variables	sparsity	Pure Test One	Noisy Test One	Pure Test Half	Pure Test Multiple	Noisy Test Multiple	Other
4	dense	10	5	0	0	0	2
4	sparse	9	1	5	0	0	1
6	dense	7	4	0	0	0	4
6	sparse	7	1	4	1	3	0
10	dense	7	5	0	0	0	3
10	sparse	3	0	4	6	3	0
20	dense	2	13	0	0	0	3
20	sparse	2	2	2	7	3	2

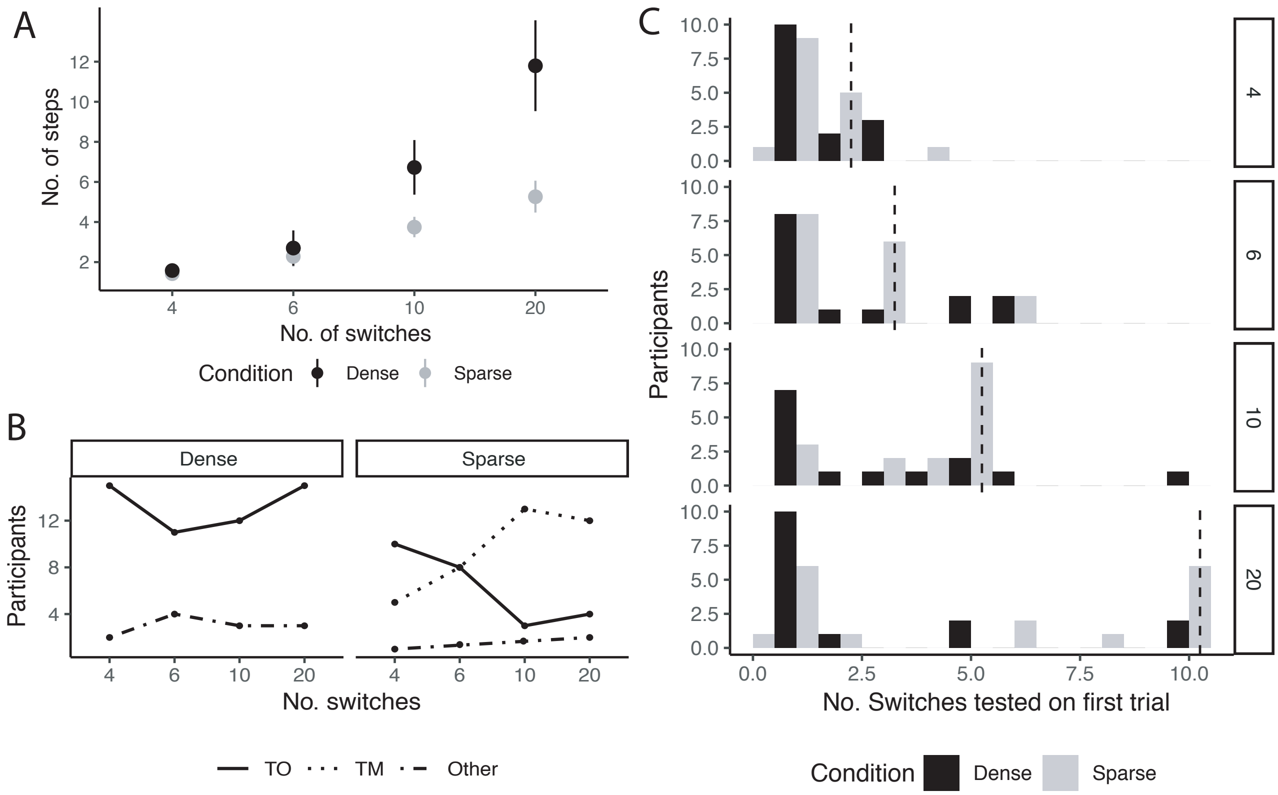

Figure 6. Results from Experiment 2. (A) Average Number of interventions that participants in each group needed to find the working/broken switch. Recall, *TO = Test One*, *TM = Test Multiple*. (B) Strategy classification. (C) Number of switches manipulated on trial 1. Dotted lines indicate 1/2 of the total number of variables.

Performance. Pooling across the four groups with different numbers of switches, participants in the sparse conditions made fewer interventions ($M = 4.17$, $SD = 2.98$) than in the dense condition ($M = 8.91$, $SD = 8.01$, $t(129) = 4.5$, 95% CI $[2.7, 6.8], p < 0.001$). They were also more likely to make the correct choice at the end (76% vs. 43% correct, Fisher's exact 95% CI [1.8, 9.4], $p < 0.001$). As the EIG analysis predicts, this efficiency difference was disproportionately driven by participants in the groups with larger numbers of switches. To illustrate this, Figure 6A plots the average number of tests participants needed to find the correct solution in every experimental group. It shows that that the *number of trials participants saved* in the sparse condition compared to the dense condition increased from 0.15 trials in the 4 variable condition to 6.53 trials in the 20 variable condition. This finding qualitatively matches the EIG-optimal actor predictions in Figure 4.

Strategies. Table 3 shows the detailed strategy use per group, based on the definitions described in the results section of Experiment 1. Again, we further collapsed these data into three strategy types (Test One, Test Multiple, and Other). The results for these summary strategies is shown in Figure 6B. As expected, the vast majority of participants in the dense group adopted a Test One strategy for all levels of the number of variables. In the sparse condition, on the other hand, the proportion of Test One users varied with the number of switches (Fisher's exact test with null hypothesis of equal proportion of TO to TM participants in each group of number of variables, $p < 0.023$).

Figure 6C shows how many variables participants manipulated on their very first trial. As in Experiment 1, participants who did not use the Test One strategy were particularly likely to change *exactly half* of the switches (see black dotted lines). However, we found some variance in the number of switches changed, particularly in the 10 and 20 switches group. Some participants chose to manipulate slightly fewer than half, and some considerably fewer (e.g., 3 in 10 or 6 in 20 switches). Recall that the cost-minimizing strategy from an optimal planning model (described in Appendix A) actually reveals that testing slightly fewer or more than half is an equally good strategy in terms of expected value (Table 5 in the Appendix). This is because it can increase the likelihood of a "quick win" if the working switch happens to be in the smaller of the two subsets. Thus, some of these participants were still following an optimal strategy. Furthermore, even among suboptimal participants, some still acted much more efficiently than Test One participants. For example, testing 6 in 20 variables will still eliminate 8.4 variables, on average, from the set of potential causes (there is a 6/20 chance of the light turning on, thus eliminating 14 variables and a 14/20 chance of the light staying off, thus eliminating 6 variables), while manipulating 1 in 20 eliminates only 1.9 on average. In sum, adding more variables increased the number of participants in the sparse condition who adopted more efficient strategies (i.e., testing multiple variables) than testing one variable at a time.

Stopping. As in Experiment 1, we again found that some participants (21% overall, with 23% in the dense and 18% in the sparse condition) chose to perform further tests after they had identified the broken/working switch (at least from the perspective of an optimal learner yoked to their choices). With the larger sample size in this experiment we could further analyze factors that influenced whether or not participants conducted unnecessary tests. In particular, we were interested in whether observing the outcome of interest (i.e., the light turning on in the sparse condition or the light not turning on in the dense condition) influenced people's stopping decisions. What allowed us to do this analysis is that sometimes the working or broken switch can be identified without ever seeing the relevant outcome, simply by ruling out all other possibilities. The analysis was conducted separately for the sparse condition (where the relevant outcome was the light turning on) and the dense condition (the relevant outcome was light not turning on). In the sparse condition, we found that among participants who, according to an optimal learner, had identified the correct switch, those who had *not* yet observed the outcome were more likely to make additional interventions than those who had activated the working switch and seen it cause the outcome (6 out of 7, vs. 6 out of 48; Fisher's exact test 95% CI [3.7, 1947], $p < 0.001$). However, in the dense condition among those participants with enough information to identify the broken switch, the group that had not yet tested it were no more likely to make additional tests than those that had tested it (2 out of 7, vs. 13 out of 31, Fisher's exact test, 95% CI [0.05, 4.1], $p = 0.68$).

Among those participants who stopped before finding the broken switch in the dense condition, we found relatively little evidence for such "rational guessing". Only 2 out of 24 of those who stopped before their bonus ran out (while they could still gain from guessing), did so at a point where it was optimal (that is, guessing was at least as good as continuing).

Discussion. These results provide further evidence that information about sparsity affects how people intervene on multiple-variable systems. On average, participants in the sparse group were more likely to manipulate multiple variables at a time, whereas those in the dense group were more likely to follow a Controlling Variables strategy (Test One). This replicates the main result from Experiment 1 and qualitatively matches the predictions of the EIG-optimal actor analysis above. Furthermore, this effect was strongly affected by the total number of variables in the system. The more switches were presented to participants (the more sparse the environment in the sparse condition, and the more dense the environment in the dense condition), the more prominent was sparse participants' use of a Test Multiple strategy. Again, this finding is in line with our EIG analysis.

However, this experiment also showed, even more strongly than Experiment 1, that the Test One strategy is a common choice even for participants in the sparse condition, who would be better off manipulating multiple variables. In fact, in the 4 and 6 switch condition, testing one variable was at least as common as testing multiple. Again, this suggests that in the absence of a *strong* incentive to do otherwise, many people have a tendency to change variables individually. We will further address this finding in the following experiments.

Finally, the experiment revealed an interesting pattern in participants' decisions to stop or continue making tests once they identified the relevant switch. In the sparse condition, participants often chose to make such additional tests when they had not yet observed the light turn on. This finding mirrors research on many other hypothesis-testing tasks suggesting that people have a bias to verify or confirm their hypotheses, even if doing so does not lead to additional information (e.g., Klayman & Ha, 1987; Nickerson, 1998; Ruggeri et al., 2016). In causal intervention tasks, such a tendency has specifically manifested itself in a preference for producing the positive effects (variables turning on) that a particular hypothesis entails (Coenen et al., 2015; Bramley et al., 2017). Because we did not find that participants in the dense group tried to create the expected non-effect, this experiment provides further evidence for this preference to verify positive outcomes in causal systems. We will continue to address this question in Experiment 4 and the General Discussion.

Experiment 3 - no sparsity information

Although many subjects automatically switched to a Test Multiple strategy in Experiments 1 and 2, a distinct subset continued to test one variable at a time even in sparse environments. One possible explanation for this finding is that Controlling Variables acts as a kind of behavioral default when people are not given any information about the sparsity of a system. If this is the case, then some participants may have used a Test One strategy in earlier experiments if they were unsure or did not pay enough attention to the sparsity instructions. To explore this possibility, Experiment 3 explored what strategy people use to test a multivariate system when they are offered no explicit expectations about the number of causes in the instructions. If Test One is indeed people's default strategy, we would expect participants to use it when testing under this new condition.

Method

Participants. Fifty-seven participants (35 males; $M_{age} = 36, SD_{age} = 13$) were recruited on Amazon Mechanical Turk.[3] Recruitment was restricted to AMT workers within the United States aged 18 or above. Participants were paid \$0.50 for their participation, with the possibility of earning an additional bonus of up to \$1. Approval for this study was obtained by New York University's Institutional Review Board (IRB) under the protocol "Active Learning in Dynamic Task Environments" (IRB-FY2016-231).

Stimuli. Materials were the same as the 6 switch condition of Experiment 2. In a between-subject design, participants were again randomly assigned to a switchboard that either had one broken or one working switch.

Procedure. The procedure was the same as in the previous experiment with the exception that participants were given the *same* set of instructions in both conditions. Instead of being told to find the one broken or one working switch, they were instructed to "find out which switch(es) are working or broken". After the switch testing phase, participants were asked to indicate which switch(es) were working or broken, now being able to make multiple selections.

Results and discussion. In analyzing behavior from this experiment, we focused on the number of switches manipulated on the first trial of the experiment, which is the most informative trial to reveal their naive expectations. Analyzing their trial-by-trial behavior using the classification scheme from earlier experiments was infeasible here, because it would require knowing participants' prior beliefs over the number of causes. Figure 7 shows the proportion of participants that chose to turn on any possible number of switches on the very first trial. Data is collapsed over both conditions, because the initial instructions were the same and hence the first trial should not lead to different behaviors. The majority of participants (58%) chose to manipulate a single switch, with only 10% manipulating half. The frequency of different numbers of manipulations differs substantially from the probability of manipulating different numbers of switches at random $[.02, .09, .23, .31, .23, .09.02], \chi^2(5) = 170, p < .001$.

[3] Based on previous experiments, we had no expectation about people's behavior in the absence of any sparsity instructions and we, therefore, decided to use a larger sample size. With an initial goal of 60, the final number came about through irregular posting of AMT tasks.

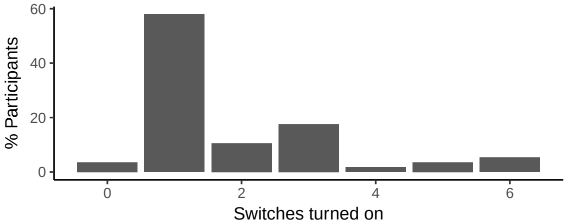

Figure 7. Experiment 3: Number of switches tested on the first trial after being given instructions to “find which switch or switches were working or broken.”

This supports the idea that Test One may be many people’s default strategy in the absence of more specific knowledge about their environment.

Note that an optimal learner initialized with a flat prior over all possible distributions of working or broken switches also assigns higher Expected Information Gain to testing one over testing multiple variables. Therefore, preference for Controlling Variables found in this experiment might be a consequence of participants selecting the most informative strategy given their knowledge of the system. Similarly, it is possible that participants who tested variables individually in the sparse condition of earlier experiments, simply ignored our instructions about the number of causes and acted as if they knew nothing about the sparsity of the system. If that was the case, their behavior would still be in line with the optimal actor analysis presented above and not necessarily a consequence of a Test One “default”.

Experiment 4: repeated interaction with sparse system

Experiment 3 showed that, when they are not offered specific prior information about sparsity, people have a strong preference for testing variables individually. In Experiment 4, we want to know to what extent using Test One in sparse conditions of earlier experiments was driven by a similar lack of knowledge about the system (e.g., a flat prior over all hypotheses) versus a default bias towards testing variables individually.

To test this, we had participants interact *repeatedly* with multiple sparse systems. Unlike Experiment 3, we also gave them the explicit sparsity instruction again (see Appendix B). This allowed them to learn about sparsity not only through instructions, but also through direct experience. If people’s use of Test One was due to them doubting (or not paying attention to) our instructions about sparsity, we expect that direct experience should mitigate this effect of prior belief and motivate them to develop the optimal strategy over time. If, on the other hand, some participants have a genuine bias towards Test One, we expect them to not change their behavior over multiple exposures with sparse systems.

Method

Participants. Thirty-seven participants (26 male; $M_{age} = 25, SD_{age} = 11$) were recruited on Amazon Mechanical Turk[4]. Recruitment was restricted to AMT workers within

[4] The choice of sample size was due to an initial goal of 40 participants, 3 of which were not collected due to the AMT assignments timing out. The initial goal was based on the observation that just under half of the participants in the sparse conditions of previous experiments used a Test One strategy. Since those were the participants we were interested in, we decided to collect a little more than double the number of

the United States aged 18 or above. Participants were paid \$1.00 for their participation, with the possibility of earning an additional bonus of up to \$1. Approval for this study was obtained by New York University's Institutional Review Board (IRB) under the protocol "Active Learning in Dynamic Task Environments" (IRB-FY2016-231).

Stimuli. Materials were the same as the sparse condition with six switches in Experiment 2. However, each participant tested five different switchboards sequentially. Each board had a different color and the order of colors was randomly chosen for each participant. As before, the working switch of every switchboard was randomly generated.

Procedure. Participants received the same instructions as in the sparse condition of Experiment 2, with the slight modification that they were told they would be testing five switchboards, not one. They were still told to expect one working switch per switchboard and were reminded of this again before interacting with each of the five switchboards. The final bonus payment was determined by randomly selecting one of the five switchboards at the end of the task and paying participants the bonus gained from testing it.

Results and discussion.

Performance. Across all switchboards, participants made 3.01 intervention per board on average and there was no significant increase or decrease with board number ($r = -0.03, n = 185$, 95% CI $[-0.17, 0.11], p = 0.68$). Overall, participants isolated the correct working switch 88% of the time (from an optimal learner's perspective). Although this number was lowest on the very first switchboard (78%), there was no significant overall effect of board number on the probability of isolating the correct switch (logistic regression with board number predictor, $z = 1.56$, 95% CI $[-0.06, 0.57], p = 0.12$).

Strategy. The crucial question we asked with this experiment was whether participants who started out using the less efficient Test One strategy would learn over time that they could be more efficient by manipulating multiple switches. Table 4 shows the detailed breakdown of the number of participants using each of the strategies defined in the Results section of Experiment 1. Figure 8A and B shows the number of switches participants manipulated on the first trial of each of the five switchboards, as well as their higher-level strategy classifications. First, note that the results from the first board (top row) fall somewhere between the strategy distributions found in Experiment 1 and Experiment 2 (6 switches condition). Although more participants adopted a strategy of changing multiple variables, a substantial number of participants used a Test One strategy. Second, this pattern remained remarkably stable as participants interacted with the remaining four switchboards. Even on the final iteration of the experiment, more than a third of the participants tested variables one-by-one. The distrubution of strategies did not depend on board number, Fisher's exact test $p = .96$. To check that this pattern holds at the level of individual participants, consider Figure 8C which compares the number of switches participants manipulated on the first trial of board 1 and on the first trial of board 5 as a transition heatmap. It shows that the majority of participants manipulated the same number of switches on the first and last board (black squares). In fact, of the 14 participants who tested a single variable on board 1, only one switched to testing half by the time they were interacting with board 5.

Interestingly, there was no significant strategy difference in terms of the number of tests participants conducted ($M = 2.90$ for Test Multiple participants, $M = 3.06$ for Test

participants compared to the cell sizes in Experiments 1 and 2. We had no a priori expectation of propensity to switch strategies with experience on which to base additional power analyses

One participants). However, participants who tested multiple variables were more likely to have isolated the working switch by the time they finished testing (95% compared to 78%, Fisher's exact test, $p < 0.001$). Thus, there was still a clear advantage of using the Test Multiple strategy, even if it did not reflect in the number of tests people made.

Stopping. As in the previous experiments, participants sometimes conducted additional tests after they found the solution. There was a suggestive decreasing trend of the probability of making further unnecessary tests with board number (logistic regression with board number predictor, $z = -1.93$, 95% CI $[-.68, 0.006], p = 0.054$). While on the first board, 19% of participants chose to do so, only 5% did so on the final (fifth) board. Overall, we found that participants were more likely to conduct unnecessary tests when they had not yet observed the light turn on. They made additional tests on 10 out of 23 of those trials compared to 11 out of 139 (Fisher's exact test 95% CI [2.8, 28], $p < 0.001$). This replicates the stopping results found in Experiment 2.

Experiment 4 found that people's strategies remained largely static over multiple instances of interacting with sparse systems. It therefore further strengthens the finding that a tendency to control variables can act as a persistent behavioral default. Recall that, before testing each new switchboard, participants were once again reminded of the fact that only a single switch was working and their repeated exposure of the task should have strengthened this belief. Thus, it seems unlikely that, in prior experiments, Test One participants ignored our prior instructions about sparsity and based their strategy on a flat prior belief over all possible structures.

Table 4
Detailed strategy use in Experiment 4.

Board no.	Pure Test One	Noisy Test One	Pure Test Half	Pure Test Multiple	Noisy Test Multiple	Other
board 1	13	0	14	3	3	4
board 2	12	1	15	4	2	3
board 3	14	0	15	4	2	2
board 4	13	0	17	4	1	2
board 5	12	0	13	7	4	1

General Discussion

In this paper, we investigated how people test multivariate causal systems with a single outcome of interest. Using an optimal actor model, we showed that the most efficient strategy crucially depends on a learner's belief about the *causal sparsity* of the system, that is, on the density of causes among the total number of variables. When there are many possible causes, learners are best off adhering to a Controlling Variables principle — operationalized here as a strategy of turning on variables one-by-one — to isolate the specific effect of every variable without the confounding influence of the others. If there are only very few causes among variables, it becomes more beneficial to test multiple variables at once to narrow down the space of actual causes quickly. Our behavioral experiments investigated what strategies people use to test multivariate systems and whether their knowledge about

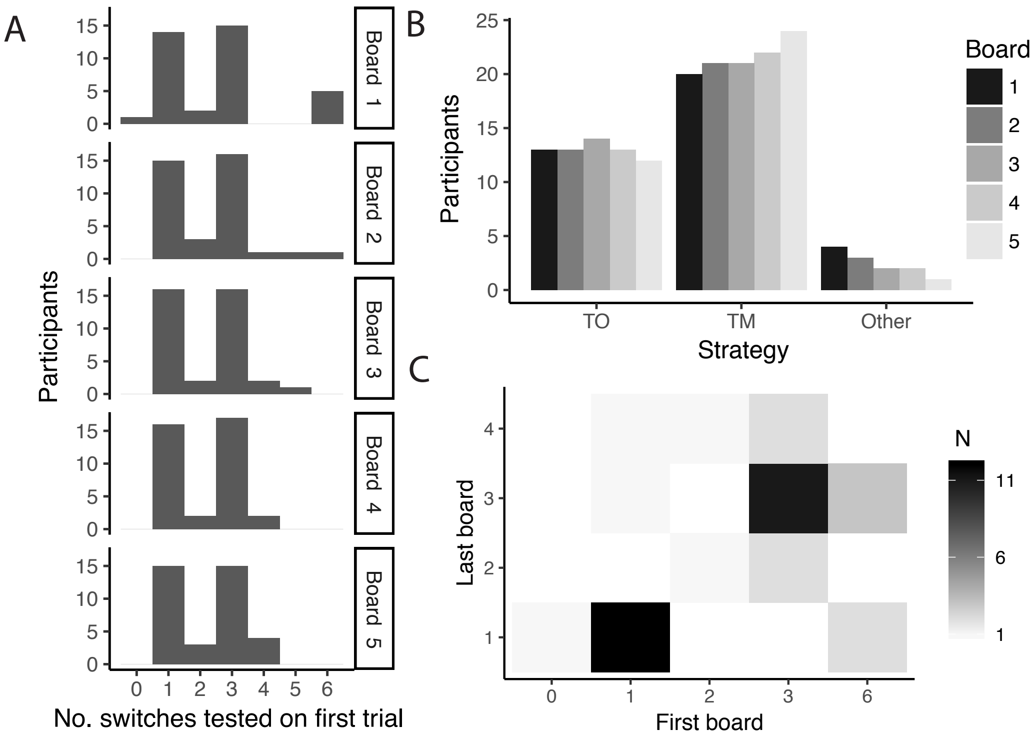

Figure 8. Experiment 4 Results. (A) Number of switches turned on during the first trial of every switchboard (board 1 corresponds to first, board 5 to last switchboard that participants interacted with). (B) Strategy classification. Note that the number of Test One participants remains constant. (C) Number of participants choosing to manipulate a given number of switches on the first trial of board 1 (x) and board 5 (y). The vast majority used the same strategy on both boards.

causal sparsity had any effect on their choices. Two main findings emerged from these experiments.

First, as predicted by the EIG-optimal actor model, participants adapted their behavior based on knowledge about the causal sparsity of a system (Experiments 1 & 2). When causes were sparse (only one cause) they frequently chose to manipulate multiple (often half of the) candidate variables. In dense tasks (N-1 causes) they were more likely to test one variable at a time. We also found that increasing the degree of sparsity/density, by increasing the total number of variables, amplified this effect on people's strategies, because it magnifies the benefits of the more effective strategy (Experiment 2). In sparse systems with more variables, people were more likely to manipulate multiple or half of the variables.

Second, we also found a number of ways in which subjects deviated from the predictions of the EIG analysis. Most strikingly, across all experiments with a sparsity manipulation (i.e., all except Experiment 3), many participants showed a tendency to test variables one-by-one even if it was more efficient to test multiple variables at once. Even after several repetitions of the task, those who started testing variables one-by-one were very unlikely to switch to the more efficient strategy (Experiment 4). We also found that participants using either strategy often continued to collect redundant (costly) information after having already learned enough to know the correct answer. Like in Experiment 2, they were more likely to do so when they had not yet observed the light turn on, which reveals a particular preference for collecting confirmatory evidence that has been found in other hypothesis testing tasks.

In the remainder of this article, we will discuss how these findings contribute to our understanding of how people should and do approach multivariate systems and how they shed light on important determinants of strategy use during causal exploration.

Ecological rationality of experimentation strategies

In the educational literature the Controlling Variables principle (here: the Test One strategy) is treated as a hallmark of scientific thinking and optimal experimentation (Inhelder & Piaget, 1958; Kuhn et al., 1995; Chen & Klahr, 1999). As we show in our modeling analyses, this gives an incomplete picture of the right approach to experiment on multivariate systems. While it is true that controlling variables is the most efficient strategy under a variety of different assumptions, it is clearly inferior in sparse systems, especially when there are many variables. Our analysis thus demonstrates the importance of examining the ecological validity of strategies (Goldstein & Gigerenzer, 2002; Todd & Gigerenzer, 2012; Ruggeri & Lombrozo, 2015; Parpart et al., 2018), that is, the fit between a strategy and the environment it is used in.

Taking this ecological perspective allowed us to detect a novel finding with respect to the adaptive nature of people's causal experimentation strategies. This finding tallies with other recent work on causal interventions, which showed that people's strategy choices were made adaptively with respect to internal constraints, like cognitive load, and external factors like the match of a strategy and the task environment. For example, Coenen et al. (2015) found that people were more likely to choose a simpler, heuristic intervention strategy when put under time pressure, but adopted the more complex information-maximizing strategy in an environment that was designed to yield poor results from using the heuristic. The current finding adds to this result by demonstrating that explicit (instructed) prior knowledge about the environment can also affect people's strategies in an adaptive manner.

In finding that sparsity affects behavior, our experiments also add to recent evidence from a non-causal information search task that manipulated hypothesis sparsity. In a spatial search task, similar to the game "Battleship", Hendrickson et al. (2016) gave people different instructions on the number of tiles on the game board that were taken up by hidden ships. Participants were then given the option to reveal either a ship tile or a non-ship (i.e., water) tile. Participants changed their strategy depending on the prior instructions. When they expected few ship tiles (sparse hypothesis space) they chose to reveal ship tiles and when they expected many ship tiles (dense hypothesis space) they chose to see water tiles. Although their definition of "sparsity" is slightly different than the one used in this paper (because of the different nature of the two tasks), both papers show how people's meta-beliefs about a hypothesis space impact their information seeking behavior.

Controlling Variables default

The fact that a substantial number of participants chose to test one variable at a time even in sparse environments was not a finding we anticipated. It is particularly surprising given the previous literature on scientific experimentation. Teaching children to control variables has frequently been shown to be a rather arduous task (Chen & Klahr, 1999) and even adults sometimes have a hard time adopting this strategy (Kuhn et al., 1995). However, participants in our experiments were not only able to use the strategy successfully when it maximized information, but they sometimes persistently used it when it did not. This shows that there is a potential limit to the ecological rationality claim of the previous section. Whereas the degree of sparsity in the environment had an impact on people's *initial* strategy selection, this selection appears immutable for at least some time, even when experience should demonstrate that a Test One strategy is inefficient.

We consider a number of explanations for this finding. First, in the education literature, experimentation strategies are often probed by giving participants a single forced choice between a non-confounded test (one variable changed) and a confounded test (multiple variables changed, see Chen & Klahr, 1999; Kuhn & Brannock, 1977). However, it is less common to allow participants to conduct and observe the results of multiple experiments in sequence, which poses the challenge of interpreting the outcomes of multiple experiments, integrate them with existing knowledge, and memorize past results (Fernbach & Sloman, 2009). Because of these additional cognitive requirements, it is possible that the sequential nature of our experiments motivated a portion of participants to use a strategy that keeps the demands on storage and integration of evidence manageable. Ruling out variables one-by-one has an advantage over Testing Half in this respect, since it only requires reasoning about a single variable on every trial. This explanation would be in line with a recent finding by Bramley et al. (2017), who showed that in more complex and open-ended causal intervention experiments people have a preference for holding many variables constant in order to test fewer variables at a time, presumably to minimize cognitive effort. Note that repeated exposure to different systems, as in Experiment 4, does not change the fact that Testing Multiple is more effortful, so this interpretation explains the observed behavior (people not changing strategy even with repetition).

Another contributing factor might be that changing one variable at a time is explicitly taught in STEM classes as the main method for scientific experimentation (e.g., National Academy of Sciences, 2013; National Research Council, 1996). It is possible that students are so well trained in this strategy that they start adopting it as a default for any experimental situation and only alter it if the environment seems particularly unsuitable (e.g., when it's very sparse). If so, the present studies raise the question whether this curriculum standard might in some cases hinder efficient experimentation by promoting a narrow focus on the idea of testing variables individually irrespective of the specific situation. This perspective lines up with recent work arguing that the kind of broad exploration often exhibited by children is actually beneficial for their learning, because it allows them to explore a wider number of hypotheses, before becoming more targeted in their information search behavior (e.g., Lucas, Bridgers, Griffiths, & Gopnik, 2014). Also, testing multiple variables at once can be a reasonable strategy from a purely scientific perspective. For example, during exploratory research, it is common to change multiple factors and see if any of them have an effect on some outcome, before honing in more precisely on the factor(s) that are responsible. Thus, rather than teaching students a single strategy for designing experiments, we should offer them a toolbox of strategies and teach them to to choose the *right* strategy for their question and prior knowledge about a domain. To substantiate such speculations about the impact of instruction on the Test One bias, it would be worthwhile to investigate strategy use in children of different ages and at different stages of their education. The authors of this paper are currently pursuing this developmental direction using a similar paradigm to the one used in this paper.

In addition to explicit instruction of the Controlling Variables principle, it is possible that people learned it autonomously through repeated exposure to systems like the ones used in our experiments (e.g., switches and lights). Such exposure could have led people to develop an inductive bias (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010) to expect that more than one variable (switch) can bring about the relevant outcome (light). If this assumption was deeply ingrained, encountering five different sparse systems in Experiment 4 may not have been sufficient to override it. Evidence in support of the idea that people have *some* prior expectations about causal sparsity comes from a recent study by Strickland, Silver, and Keil (2016) who show that people have different beliefs about the average number of causes of an effect depending on the domain. Specifically, participants expected more causes (e.g., a less sparse environment) in psychological domains, compared to physical domains. Given that our current experiments are based on a physical system (switches) it would be interesting to further explore if the default bias towards testing one variable at a time is exacerbated when variables are psychological states.

Stopping decisions

In addition to deviations from the most efficient strategy, we found that participants sometimes made additional (costly) tests even if — from the perspective of an optimal learner — they already had enough information to determine which of the variables could affect the outcome. Oversampling with respect to the cost structure is not an isolated finding in information search tasks (e.g., Tversky & Edwards, 1966; Juni, Gureckis, & Maloney, 2016). However, it is curious that in this task participants were, in principle, getting *no* information from their additional tests. A similar result was recently found in a causal learning version of the Twenty Questions game that required participants to ask questions or make interventions in order to sequentially narrow down causally relevant objects (Ruggeri et al., 2016). In this task particularly young children (7-year-olds), but to lesser degree adults as well, were found to collect additional information when they already knew enough to make a correct choice.

Our experiments also give at least some indication of the source of this suboptimal stopping behavior. Importantly, we found that, in the sparse condition, additional tests were much more likely when a participant had not yet observed the working switch affect the outcome (i.e., turn on the light), but the same was not the case for the non-working effect in the dense condition. First of all, this suggests that suboptimal stopping was not merely a memory effect. If participants conducted extra tests because they forgot about prior ones, they should have done so irrespective of how many variables one should manipulate in order to gather information most efficiently. Instead, the fact that participants in the sparse group sought to confirm the effects of the working switch might be the consequence of a type of *positive testing strategy*, that is, the desire to confirm the expected outcome(s) — here, the causal effect — predicted by one's hypothesis. Evidence for this kind of strategy has been found in different kinds of hypothesis testing experiments in which participants frequently test instances that can confirm their current hypothesis instead of trying to falsify it (e.g., Klayman & Ha, 1987; Wason, 1960; Mckenzie & Mikkelsen, 2000; White, 2009). Recently, a similar tendency was found in a number of causal learning experiments showing that participants were particularly likely to intervene on variables that would yield many effects predicted by a structure that the learner currently considers to be plausible (Bramley et al., 2017; Coenen et al., 2015). There also exists evidence that people in some cases perceive the presence of causal effects as more informative (Coenen & Gureckis, 2015) and more likely (Davis & Rehder, 2017) than non-effects that objectively have the same informational value or likelihood. Our experiments thus further corroborate this asymmetry of effects and non-effects.

However, the positive testing we observed was not uniquely indicative of confirmatory testing. When participants in the dense condition had enough information to identify the non-working switch, they did not generally confirm this by producing confirmatory case in which only the broken switch is activated and the effect does not occur (see analyses of Experiment 2). This is consistent with people having a *positive testing strategy* (Klayman, 1995; McKenzie, 2004), favoring tests in which their hypothesis predicts a positive outcome to tests in which their hypothesis predicts a negative outcome.

Complexity

A question we have not yet touched on is how optimal strategies change when some of the simplifying assumptions about the nature of the causal system no longer hold. For example, consider dropping the assumption that causes are independent of one another, in which case there may exist interactions between them (Lucas & Griffiths, 2010; Eberhardt, Glymour, & Scheines, 2012). In that situation, learners eventually *have to* test combinations of variables to ensure they uncover the complete causal structure of a system. Another possibility is that causes take on more values than present (here: "on") or not absent (here: "off"). For example, causes might be disabling (i.e., preventing the outcome from occurring, see Walsh & Sloman, 2011). In that case, the meaning of controlling a variable changes from turning it off, to finding a neutral state in which it neither enables nor disables the outcome. Another complicating factor that often holds in real-world causal systems is the probabilistic nature of the relationship between causes and outcomes. Compared to the deterministic case used in this paper, probabilistic causal links can make intervention-planning much more difficult (e.g. Bramley et al., 2015; Steyvers et al., 2003). In future work, we hope to explore how people plan strategies to interact with such more complex multivariate systems.

Conclusions

This paper contributes two sets of findings to our understanding of causal experimentation.

First, the model-based analyses demonstrate that, contrary to some debates in the education literature, the principle of Controlling Variables (testing variables one-by-one) is not always the most efficient strategy for discriminating among multiple causes. Instead, we show that the *causal sparsity* of a system determines whether the information-maximizing strategy is to manipulate few or many variables. Second, our empirical findings highlight some of the major determinants of self-directed causal learning behavior. They show that, in line with the optimal actor analysis, learners can adapt their behavior based on knowledge about abstract features of their environment (here: causal sparsity) when planning causal experiments. However, our experiments also revealed a tendency towards habitual strategy use, which might be based on people's prior learning history, instructions received in the past, or strong environmental priors. Finally, we found asymmetric preferences for seeking effects rather than non-effects, which adds to growing evidence that biases found in other areas of self-directed hypothesis testing also affects causal experimentation.

Acknowledgments. This research was supported by NSF grant BCS-1255538 and a John S. McDonnell Foundation Scholar Award to TMG.

Appendix A

In addition to the informational analysis presented above, we also considered the strategy that optimizes the expected payoff over the entire sequence of tests. Assuming that the cost for making a test and the payoff for finding the correct solution are defined in advance, this kind of decision problem can be solved via dynamic programming.

Here, we will derive predictions of this optimal planning model for the same basic example used in the introduction of a system of n binary variables which can be either (deterministic) causes or non-causes. To simplify the analysis, we only consider the most extreme cases of a sparse system with a single cause and a dense system with N-1 causes. These cases also correspond to most of the experiments presented in this paper. We assume that there exists a cost, c, for conducting each test, as well as a final reward, r, for finding the correct solution and no reward otherwise. On every trial, a learner must choose from a set of possible actions, where an action corresponds to either turning on any number of variables or stopping and guessing the solution, that is $a \in A = \{1, ..., n, \text{stop}\}$. The state of a learner's current belief can be summarized as the number of remaining variables that might be the cause (in the sparse case) or non-cause (in the dense case), that is $x \in X = \{1, ..., n\}$. We assume that this belief state is updated after every observation such that proven (non-)causes are eliminated. The value of a state at trial t is:

$$V_{x_t} = \max_{a \in A} U_{(x_t, a)}, \tag{6}$$

where the utility for a further test is

$$U_{x_t, a \neq \text{stop}} = \mathbb{E}_{x_{t+1}}[V(x_{t+1})|a] - c, \tag{7}$$

and the utility of stopping and guessing is

$$U_{x_t, \text{stop}} = \frac{1}{x_t} r. \tag{8}$$

We use this model to compute the optimal solution for the first trial of a sparse (one cause) system of 4, 6, 10, or 20 variables, with a final reward $r = 1$ and a sampling cost of

Table 5

Predictions of the Optimal Planning Model for the First Trial of the Sparse Condition (one cause).

Number changed:	n=4	n=6	n=10	n=20
sparse	2	2, 3, 4	4, 5, 6	8, 9, 10, 11, 12
dense	1	1	1	1

$c = \frac{r}{n}$. Again, we assume that the learner starts with a uniform belief over which variables are causes and non-causes. This corresponds to the 8 conditions of Experiment 2. Note that the solution when $n = 6$ also applies to Experiments 1 and 4.

Predictions are shown in Table 5. Unlike the myopic EIG model, the optimal planning strategy for the sparse case includes other methods of dividing the initial hypothesis space, besides the exact half-split. In these cases, the added risk of needing more interventions to find the correct solution is offset by the probability of the working switch being part of the smaller of the two subsets, which speeds up the process of finding it. The optimal solution for the dense case obviously remains the same as the myopic EIG solution since any test of more than one variable yields completely uninformative evidence.

In addition to predicting the number of variables to manipulate in a given state, this model also predicts when a learner should stop making tests altogether. In addition to the obvious case of having found the solution, it is sometimes possible that it becomes equally or more valuable to stop earlier and make a guess about the remaining possibilities. For example, when there are two remaining possibilities, stopping early results in a 0.5 probability of guessing correctly and winning the remaining reward. If this value happens to be greater than or equal too the remaining reward minus the cost of an additional test, stopping early is a viable option. Given the same cost structure as Experiment 2, the model predicts that an optimal actor could stop and guess if no solution is found after 2 tests on 4 switches, 4 tests on 6 switches, 8 tests on 10 switches, and 17 tests on 20 switches. This holds assuming that, before that point, all tests were optimal (here: testing one variable at a time).

Appendix B

The following instructions about the sparsity of the system were used in the Mechanical Turk Experiments 2 and 4. In particular participants were instructed twice and tested once on the number of working or broken switches, before proceeding to the main task. In the beginning, participants were told (in the **dense/sparse** condition, respectively):

> "On each switchboard, one of the switches is **broken/working** and all others are **working/broken**".

Then, during the demo phase, they were told again that

> "Again, **only one of the switches is broken/working**, the rest are **working/broken**".

In the quiz, participants were asked:

"How many of the switches on the switchboard are working?"

They had to answer correctly in order to proceed to the main task.

References

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.

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

Bramley, N. R., Nelson, J. D., Speekenbrink, M., Crupi, V., & Lagnado, D. A. (2014). *What should causal learners value?* Poster presented at the Annual Meeting of the Psychonomic Society, Long Beach, CA.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Chen, Z., & Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.

Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? In R. D. Noelle, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Davis, Z. J., & Rehder, B. (2017). The causal sampler: A sampling approach to causal representation, reasoning and learning. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society.*

Eberhardt, F., Glymour, C., & Scheines, R. (2012). On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence.*

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple Heuristics that Make us Smart.* Oxford: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision*, *3*(1), 62.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking.* New York: Basic Books.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2016). Information sampling behavior with explicit sampling costs. *Decision*, *3*(3), 147.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*(1), 111–146.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation*, *32*, 385–418.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

Kuhn, D., & Angelev, J. (1976). An experimental study of the development of formal operational thought. *Child Development*, 697–706.

Kuhn, D., & Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. *Developmental Psychology*, *13*(1), 9–14.

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., & Carver, S. M. (1995). Strategies of knowledge acquisition. *Monographs of the Society for Research in Child Development*, i–157.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development*, *23*(4), 435–451.

Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. *Advances in child development and behavior*, *17*, 1–44.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2006). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.*

Langsford, S., Hendrickson, A., Perfors, A., & Navarro, D. J. (2014). People are sensitive to hypothesis sparsity during category discrimination. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.

Markant, D., & Gureckis, T. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th annual meeting of the cognitive science society.*

Mckenzie, C. R., & Mikkelsen, L. A. (2000). The psychological side of hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, *7*(2), 360–366.

McKenzie, C. R. M. (2004). Hypothesis testing and evaluation. *Blackwell handbook of judgment and decision making*, 200–219.

McKenzie, C. R. M., Chase, V. M., Todd, P. M., & Gigerenzer, G. (2012). Why rare things are precious: The importance of rarity in lay inference. *Ecological rationality: Intelligence in the world*, 309–334.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148.

Mosher, F. A., & Hornsby, J. R. (1966). *On asking questions.* New York, NY: Wiley.

National Academy of Sciences. (2013). *Next generation science standards: For states, by states.* Washington, DC: The National Academies Press.

National Research Council. (1996). *The national science education standards.* National Academies Press.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological review*, *118*(1), 120–134.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175.

Nielsen, F., & Nock, R. (2011). A closed-form expression for the sharma–mittal entropy of exponential families. *Journal of Physics A: Mathematical and Theoretical*, *45*(3), 032003.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Ormerod, T. C., MacGregor, J. N., Chronicle, E. P., Dewald, A. D., & Chu, Y. (2013). Act first, think later: The presence and absence of inferential planning in problem solving. *Memory & Cognition*, *41*(7), 1096–1108.

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as bayesian inference under extreme priors. *Cognitive Psychology*, *102*, 127–144.

Pearl, J. (2009). *Causality.* Cambridge university press.

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*(1), 93–125.

Ruggeri, A., & Feufel, M. A. (2015). How basic level objects facilitate question asking in a categorization task. *Frontiers in Psychology*, *6*.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, *52*(12), 2159-2173.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–332.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of information.* University of Illinois Press.

Simmel, M. L. (1953). The coin problem: a study in thinking. *The American journal of psychology*, *66*(2), 229–241.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives.* Oxford University Press.

Sloman, S. A., & Lagnado, D. A. (2005). Do we "Do "? *Cognitive Science*, *29*, 5–39.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive science*, *27*(3), 453–489.

Strickland, B., Silver, I., & Keil, F. C. (2016). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. *Memory & Cognition*, 1–14.

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world.* Oxford University Press.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216.

Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind &amp; Language*, *26*(1), 21–52.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

White, P. A. (2009). Accounting for occurrences: An explanation for some novel tendencies in causal judgment from contingency information. *Memory & Cognition*, *37*(4), 500–513.

Woodward, J. (2005). *Making things happen: A theory of causal explanation.* Oxford University Press.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223.