Journal of Experimental Psychology: General
2017, Vol. 146, No. 12, 1761–1780

0096-3445/17/$12.00 http://dx.doi.org/10.1037/xge0000318

Ockham's Razor Cuts to the Root: Simplicity in Causal Explanation

M Pacer and Tania Lombrozo
University of California, Berkeley

When evaluating causal explanations, simpler explanations are widely regarded as better explanations. However, little is known about how people assess simplicity in causal explanations or what the consequences of such a preference are. We contrast 2 candidate metrics for simplicity in causal explanations: node simplicity (the number of causes invoked in an explanation) and root simplicity (the number of unexplained causes invoked in an explanation). Across 4 experiments, we find that explanatory preferences track root simplicity, not node simplicity; that a preference for root simplicity is tempered (but not eliminated) by probabilistic evidence favoring a more complex explanation; that committing to a less likely but simpler explanation distorts memory for past observations; and that a preference for root simplicity is greater when the root cause is strongly linked to its effects. We suggest that a preference for root-simpler explanations follows from the role of explanations in highlighting and efficiently representing and communicating information that supports future predictions and interventions.

Keywords: explanation, simplicity, parsimony, causal inference, inference to the best explanation

Supplemental materials: http://dx.doi.org/10.1037/xge0000318.supp

Simpler explanations are better explanations. This intuition, right or wrong, often guides both scientific and everyday reasoning, earning the moniker "Ockham's Razor" for the unnecessary complexities that it "cuts" out of explanations. While simplicity is lauded by both scientists and philosophers, there is little consensus on how simplicity should be defined. William of Ockham argued that we "not multiply entities beyond necessity," suggesting that simplicity is a matter of the *number of entities* involved in an explanation. Newton's first Rule of Reasoning in Philosophy is that "we admit no more causes" than those sufficient to explain our observations, suggesting *causes* are the units in which simplicity is measured. Einstein tells us that "the grand aim of all science . . . is to cover the greatest possible number of empirical facts . . . from the smallest possible number of hypotheses or axioms," suggesting the size of a *set of hypotheses* or *axioms* is what matters (for quotations, see Baker, 2010).

Beyond these classic examples, contemporary philosophers, statisticians, and computer scientists have developed formal definitions of simplicity that can be used to guide theory choice, model selection, and inference (for review see Sober, 2006). Simplicity is argued to lead to more accurate inference (Tenenbaum & Griffiths, 2001), better predictions (Forster & Sober, 1994), or more efficient learning (Kelly, 2007). These proposals for the virtuosity of simplicity are often grounded in particular formal systems that make expressing these advantages straightforward: some formulations include algorithmic information theory and probability (Solomonoff, 1960), Kolmogorov complexity (Kolmogorov, 1965), the cardinality of parameterized models (Akaike, 1974; Schwarz, 1978), and the (possibly implicit) size of the hypothesis space ("the size principle," Tenenbaum & Griffiths, 2001). Within psychology, these approaches to simplicity have proven useful in modeling perceptual classification (Chater, 1996), language (Clark, 2001), and the perception of hierarchically structured domains in general (Feldman, 2009).

While appeals to simplicity are widespread, they are diverse in their application: different metrics for simplicity can come apart, and these metrics vary in how well they fit different real-world applications. For example, Kolmogorov Complexity identifies simplicity with the minimal length of code required to encode a program that generates a particular object in a universal descriptor language. This metric can be readily applied to problems that involve predicting the next element in a sequence composed of characters from a fixed alphabet (e.g., predicting the next letter in the sequence "banan_"). But some scenarios cannot be easily framed as sequences of this type. Nonetheless, people reason about such scenarios.

Thus, though formal approaches to defining simplicity in well-specified domains have been fruitful, research on intuitive judgments of simplicity in everyday explanations has made considerably less progress. This is unfortunate, as explanation is a ubiquitous phenomenon (Salmon, 1989). People constantly explain the social and physical world around them, and their explanatory choices have important consequences in a variety of domains (Keil, 2006; Lombrozo, 2007, 2012, 2016). For instance, explanations for our own and others' behavior can affect judgments of responsibility and blame (Dweck, 2008; Malle, 2011; Monterosso, Royzman, & Schwartz, 2005; Kim & Ahn, 2002), and clinicians'

M Pacer and Tania Lombrozo, Department of Psychology, University of California, Berkeley.

M Pacer is now at the Berkeley Institute for Data Science.

This work was supported by an National Defense Science and Engineering Graduate Fellowship and a Berkeley Fellowship awarded to M Pacer and National Science Foundation Grant DRL-1056712 awarded to Tania Lombrozo.

Correspondence concerning this article should be addressed to M Pacer, Department of Psychology, University of California, 3210 Tolman Hall, 94720 Berkeley, CA. ORCID: 0000-0002-6680-2941. E-mail: mpacer@berkeley.edu

explanations for a patients' behavior can affect diagnoses and treatment decisions (Ahn & Kim, 2008; Ahn, Novick, & Kim, 2003; Ahn, Proctor, & Flanagan, 2008; Kim & Ahn, 2002). If people prefer simpler explanations in these domains—and there's reason to think that they do (Frances & Egger, 1999; Kelley, 1973; Read & Marcus-Newhall, 1993)—it's especially important to provide a more precise characterization of simplicity in explanations, and to better understand the implications of a preference for simpler explanations.

In this article, we consider the nature and role of simplicity in human judgment, focusing on the explicit evaluation of causal explanations, such as explanations for symptoms that appeal to underlying diseases. In four experiments, we address the following questions about simplicity in the context of causal explanation and its role in human cognition:

Q_1: What makes a causal explanation simple?

Q_2: How are explanations selected when the simplest explanation is not the one best supported by the data?

Q_3: What are the cognitive consequences of a preference for simpler explanations? For example, does the preference bias memory or inference?

Q_4: Why do people prefer simpler explanations?

We begin by differentiating two metrics for simplicity, node simplicity versus root simplicity, and motivate these questions in light of prior research. We then report four novel experiments.

Defining Simplicity: *Node* Versus *Root* Simplicity

Ockham's razor canonically applies to arguments about the number of entities or the number of *kinds* of entities postulated to exist, a notion that is often referred to as "parsimony." In contrast, previous work on simplicity in causal explanatory judgments has typically focused on "elegance" (Baker, 2010), where the *kinds* of causes are known (e.g., which diseases exist), and competing explanations differ in which of these causes they invoke in a given case (e.g., stating that a disease is present to explain a given patient's symptoms). In this work, simplicity has been measured in terms of the number of causes invoked in the explanation (Bonawitz & Lombrozo, 2012; Lagnado, 1994; Lombrozo, 2012; Read & Marcus-Newhall, 1993; Thagard, 1989).[1] We call this metric *node simplicity*, as it involves counting the total number of causal nodes that are cited as present causal entities (as opposed to absent or unspecified entities) in the explanation.

To illustrate node simplicity, consider the case of Chris, who has been *extremely fatigued* and has been *losing weight*. What explains these symptoms? Chris could have chronic fatigue syndrome, an explanation which invokes one cause to account for both symptoms. Another possibility is insomnia (to explain the fatigue) and a decrease in appetite (to explain the weight loss), thereby invoking *two* causes. On the grounds of node simplicity, the first explanation is preferable to the second—one disease is fewer than two diseases. Read and Marcus-Newhall (1993) and Lombrozo (2007) found that when the probabilities of the corresponding explanations were unspecified, participants preferred explanations consistent with this metric—that is, they preferred to explain multiple symptoms with the smallest number of diseases. However, both Lagnado (1994) and Lombrozo (2007) found that this preference was eliminated or tempered when the simplest explanation was not the most likely. In the case of Chris, chronic fatigue syndrome could in fact be less common than having the conjunction of insomnia and a decreased appetite (if, e.g., Chris happens to belong to a population of particularly sleepless and sated people).

Both Lagnado (1994) and Lombrozo (2007) investigated people's explanatory preferences in cases where simplicity and probability were in conflict, using disease examples similar to those described above. Both researchers found that when a complex explanation was explicitly identified as more likely than a simpler alternative, participants chose the more probable explanation. However, Lombrozo (2007) additionally examined cases in which participants were provided with more indirect probabilistic cues: the base-rate of each disease. While this information was sufficient to evaluate the relative probabilities of the explanations (under assumptions about independence between the diseases), participants' choices were nonetheless influenced by simplicity. In particular, participants had an overall preference for the simpler (one-cause) explanations, but this preference was tempered by probability information. This generated a pattern of judgments consistent with the interpretation that simplicity altered the prior probability assigned to explanations, with very strong probabilistic evidence required to overcome this initial bias. Bonawitz and Lombrozo (2012) found a similar pattern of results in preschool-aged children.

This previous work establishes that simplicity is a powerful force in determining explanatory preferences, but no empirical research (to our knowledge) has attempted to differentiate alternative metrics for simplicity in explanation choice. This is problematic given that prior results are not uniquely consistent with node simplicity. We propose an alternative metric that can also explain these results, which we call *root simplicity*. Informally, root simplicity can be defined in terms of the number of *assumed* or *unexplained* causes in an explanation, where simpler explanations are those with fewer assumed or unexplained causes (which, for present purposes, we treat as interchangeable). We call this metric *root* simplicity to reflect the fact that among the causes present in an explanation, the root causes are the initiating variables (i.e., the topmost nodes that cannot themselves be explained by reference to the presence of other causes).

This metric is related to a number of proposals from philosophy and the history of science concerning the value of simplicity and the goals of scientific theorizing, though they have not always been expressed in terms of root causes. For example, the quote from Einstein included above indicates a preference for a small number of axioms, where axioms are similarly "assumed" or unexplained. Relatedly, Friedman (1974) endorses explanations that *unify* phenomena with few assumptions, saying "science increases our understanding of the world by *reducing* the total *number of independent phenomena* that we have to accept as ultimate or *given*" (emphasis added). Friedman has in mind explanations for different *types* of properties or events. If we instead consider the enumerated

[1] Strictly speaking, Read and Marcus-Newhall (1993) and Thagard (1989) quantified simplicity in terms of the number of *propositions* involved in an explanation. However, in the stimuli used in Read and Marcus-Newhall (1993), each proposition corresponded to the presence of a cause.

number of phenomena that are given or assumed, this maps roughly onto the number of causes that are given or assumed in explaining a token event, which corresponds to root simplicity as we have defined it.[2]

Although the materials from Read and Marcus-Newhall (1993) and Lombrozo (2007) do not differentiate between *node* and *root* simplicity (both metrics predict the same judgments), there are cases for which these two metrics diverge. To illustrate, consider that depression is a known cause of *both* insomnia and loss of appetite, and suppose that we know that Billy does not have Chronic Fatigue Syndrome. This leaves us with the following two explanations for Billy's fatigue and his decreased appetite: insomnia and loss of appetite, which were themselves caused by depression, or insomnia and loss of appetite, which were not caused by depression and instead arose independently (see Figure 1). We call the first explanation the complete-choice because it includes the complete set of possible causes, and the latter explanation the proximal-choice because it includes only the most proximal causes (i.e., only the causes that directly generated the tiredness and weight-loss).[3]

In this scenario, node and root simplicity diverge. Node simplicity would say that the complete-choice has a measure of three (because it cites all three causes) and the proximal-choice a measure of two (because it cites two causes). Thus, if people employ node simplicity in evaluating explanations, they should prefer the proximal-choice explanation. However, according to root simplicity, the complete-choice has a measure of one (because we only assume that Billy is depressed) and the proximal-choice has a measure of two (because it assumes that Billy independently developed both insomnia and a reduced appetite). Root simplicity, in contrast to node simplicity, favors the complete-choice explanation.

As a second example of a scenario for which node and root simplicity generate divergent predictions, consider two candidate explanations for a heart attack. In one case, the cause is heart disease, which is itself caused by metabolic syndrome (the complete-choice explanation). In the second case, the proximal cause is heart disease, but where the heart disease was not caused by metabolic syndrome—it is itself assumed or unexplained (the proximal-choice explanation). In this case, node simplicity favors the proximal-choice explanation (one cause is fewer than two), and root simplicity does not predict a preference for either explanation (in both cases, the causal chain has one assumed cause).

Examples like these, for which the two metrics predict different preferences, allow us to investigate whether node or root simplicity better characterizes people's explanatory preferences and thus address our first research question (Q_1): what makes an explanation simple? Moreover, by varying the probabilistic evidence for different explanations using structures like those just discussed, we can address our second question (Q_2): how are explanations selected when the simplest explanation is not the one best supported by the data? These questions are the focus of Experiments 1–2.

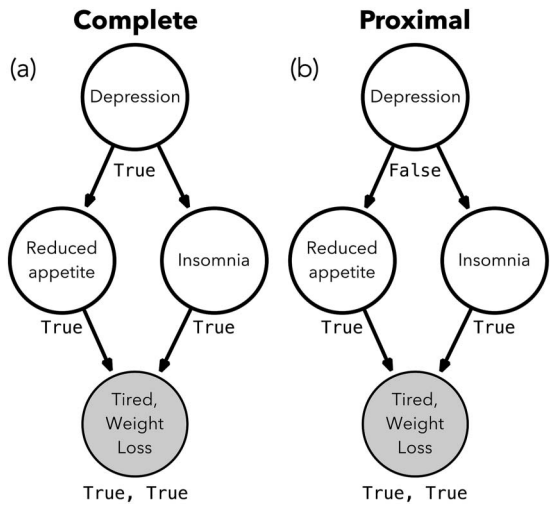

Figure 1. Illustrations of graphs corresponding to the complete (a) and proximal (b) explanations. Each circle is a variable or set of variables (e.g., disease or symptom set). The value of the node is indicated below the node; in this case nodes have values of present or not present. Arrows indicate potential causal relationships. Gray-filled circles indicate that the node's value has been observed.

The Cognitive Consequences of a Preference for Simpler Explanations

What are the implications of a preference for simpler explanations? Previous research has shown that the act of explaining can impact both learning and inference (e.g., Koehler, 1991; Sherman, Skov, Hervitz, & Stock, 1981; for reviews, see Lombrozo, 2012, 2016). Lombrozo (2007) found that participants who preferred an unlikely but simple explanation overestimated the observed frequency of the disease invoked in that simple explanation. However, Lombrozo (2007) did not differentiate node and root simplicity or go beyond an association to demonstrate a causal relationship between the act of explaining and the systematic estimation errors exhibited by some participants. Here, we vary the order in which participants explain and estimate to isolate the causal influence (if any) exerted by explanation on estimation. This allows us to address one aspect of our third research question

[2] It is important to note that *unexplained* causes are not conceptually the same as *independent* causes. For instance, the two unexplained causes of "insomnia" (to explain fatigue) and "decreased appetite" (to explain weight loss) could be probabilistically dependent. In our experiments, however, these two conceptually distinct notions are related, as the presence of a common cause "explains" two downstream causes as well as rendering them probabilistically dependent (both on each other and on the common cause). While explanatory and probabilistic dependence could be teased apart, there are some challenges to doing so: probabilistic dependence typically suggests some underlying structure (such as a direct causal relationship or an unknown common cause) that would support explanations. For present purposes, we define root simplicity in terms of which causes are unexplained, but are open to the possibility that constraints on what constitutes an unexplained cause will bring in considerations of probabilistic (in)dependence.

[3] We are contrasting the cases where something is present and where something is *not* present. Thus, ours is a discussion about a *sharp* Ockham's razor, which actively states that variables are not present, as opposed to a *dull* Ockham's razor, which is silent as to the presence or absence of variables (Sober, 2006). This assumption plays a role in our later analyses, which involve comparing evidential support for different explanations. A hypothesis consistent with a *dull* Ockham's razor will include the possibility that the variables in question are present, and (assuming it is possible that the variable is not present) will always have greater probability than the hypothesis that the variable is present. This distinction echoes the debate between Popper and Jeffreys on simplicity in the context of the prior probability of various statistical models (cf., Baker, 2010).

(Q_3): what are the cognitive consequences of a preference for simpler explanations? This is the focus of Experiment 3.

Why Do People Prefer Simpler Explanations?

Finally, why do people prefer simpler explanations? One possibility is that a preference for simpler explanations is just a human failing—perhaps a mostly harmless side effect of limited cognitive resources. Another possibility, however, is that favoring simpler explanations serves a useful cognitive function. This possibility is suggested by arguments in favor of simplicity in philosophy and statistical inference (Baker, 2010; Jeffreys, 1998; Kitcher, 1989)—some even arguing that simplicity is a foundational principle through which all of cognition can be understood (Chater, 1999; Chater & Vitanyi, 2003). However, even among those who agree on simplicity's value, it serves no single, agreed-upon role. Different roles have been proposed, and each proposal constrains (and is constrained by) the metric used to define "simplicity" (e.g., Akaike, 1974; Chater, 1999; Chater & Vitányi, 2003; Jeffreys, 1998; Kelly, 2007; Popper, 1959).

The possibility we explore is that one function of explanation is to facilitate the formation of relevant, information-rich representations of causal systems, where these representations are tailored to aiding future intervention and prediction in a variety of situations more general than the set of scenarios for which the explanation was originally invoked (Gopnik, 2000; Hacking, 1983; Lombrozo & Carey, 2006). If this is the case, a preference for simpler explanations could exist to support the acquisition, deployment, or communication of these representations. We revisit these ideas in Experiment 4, where we tackle our final question (Q_4): why do people prefer simpler explanations?

Experiment 1

In Experiment 1 we test the predictions of node simplicity versus root simplicity against human judgments. Participants learn one of two causal structures involving novel diseases and are asked to provide the most satisfying explanation for an individual's symptoms. The causal structures are designed to support two alternative explanations for which node and root simplicity generate divergent rankings. In Experiment 1 we do not provide information about the relative probabilities of different explanations. However, we introduce this information in Experiment 2.

Method

Participants. Sixty-eight participants were recruited online using Amazon Mechanical Turk and paid $.60 for their participation. Of these, 53% passed reading checks described below, leaving 36 participants for analysis. Participation was restricted to individuals with IP addresses from the United States and with approval ratings of 95% or higher on previous tasks completed through Amazon Mechanical Turk.

Conducting an a priori power analysis was challenging because there is not much related work on the topic of simplicity in causal explanation. The experiment that most closely resembles Experiment 1 is the first experiment from Lombrozo (2007). We therefore conducted a power analysis based on the data reported in this experiment, which had an observed proportion of 96% of participants choosing an explanation with one cause rather than an explanation with two causes. By contrasting this against a comparison proportion of .5 (a uniform choice between two options) using a one-sample chi-squared test, to obtain a power of $\beta = .8$ with an $\alpha = .05$, we needed 9 participants. This places our sample size per condition at twice the sample size indicated by this power analysis, suggesting that we have sufficient power in this study.

Materials and procedure. Participants were asked to imagine that they were doctors on an alien planet, Zorg. Their task was to assist in the diagnosis of alien diseases. Participants read information about the causal relationships between diseases that afflict the aliens living on Zorg. This causal information varied across the diamond-structure and chain-structure conditions.

Each participant learned about two symptoms that were chosen at random, one from a set of meaningful symptoms ("purple spots," "low fluid levels," "cold body temperature") and one from a set of "blank" symptoms ("itchy flippets," "swollen niffles," and "sore mintels"; see Lombrozo, 2007). For ease of presentation, we use purple spots and itchy flippets as sample symptoms throughout the article.

In the chain-structure condition, there were two diseases, Hummel's disease and Tritchet's disease, that could cause these symptoms under some conditions. Specifically, participants read the following information:

> Tritchet's disease always causes itchy flippets and purple spots. One of several ways to contract Tritchet's disease is to first develop Hummel's disease, which causes Tritchet's disease. Aliens can also develop Tritchet's disease independently of having Hummel's disease. Nothing else is known to cause itchy flippets and purple spots, that is, only aliens who have Tritchet's disease develop itchy flippets and purple spots.

In the diamond-structure condition, there were three diseases, Hummel's disease, Tritchet's disease, and Morad's disease. Participants in the diamond-structure condition read the following information:

> Morad's disease and Tritchet's disease together always cause itchy flippets and purple spots. If either disease is not present, neither symptom will occur.
>
> One of several ways to contract Tritchet's disease and Morad's disease is to first develop Hummel's disease, which causes both Tritchet's disease and Morad's disease. Hummel's can only cause both of these diseases or neither of them. It will never cause *just* Morad's disease or *just* Tritchet's disease.
>
> Aliens can also develop Tritchet's disease and/or Morad's disease independently of having Hummel's disease.
>
> Nothing else is known to cause itchy flippets and purple spots, that is, only aliens who have Tritchet's *and* Morad's disease develop itchy flippets and purple spots.

We chose these structures because node and root simplicity support different predictions across these cases. As illustrated in Figure 2, node simplicity always supports a preference for the Proximal case regardless of structure as there is always one less node taken to be true (diamond: complete, 3 vs. proximal, 2 and chain: complete, 2 vs. proximal, 1). The predictions supported by root simplicity differ between the two structures. In the diamond-structure condition, root simplicity favors choosing the Complete explanation over the Proxi-

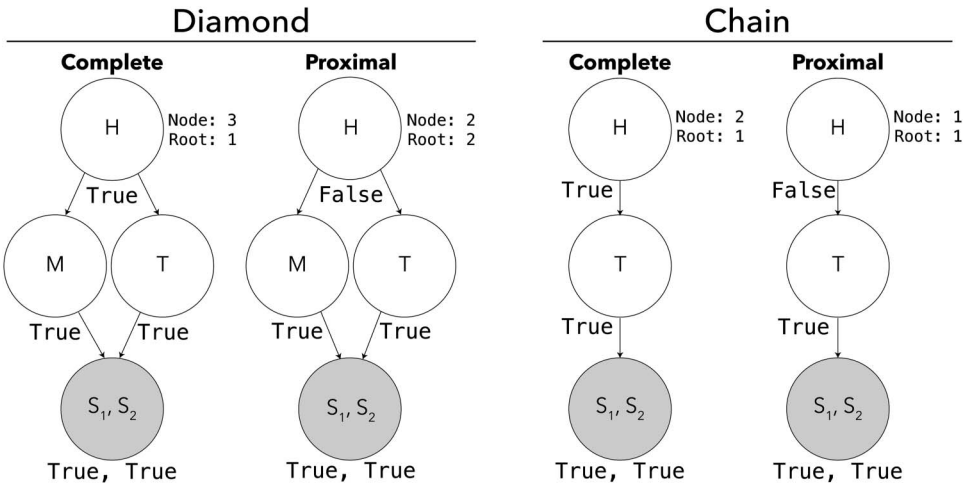

Figure 2. Illustration of the different predictions supported by root and node simplicity for the diamond and chain cases. H = Hummel's disease; M = Morad's disease; T =Tritchet's disease; S_1 and S_2 = two symptoms in question.

mal explanation (3 vs. 2). In the chain-structure condition, root simplicity treats the two explanations equivalently (1 vs. 1).

Explanation choice. Participants in both conditions were told that a particular alien, "Treda," was suffering from the two symptoms. They were asked to choose what they thought was the "most satisfying explanation" for the symptoms from a set of three explanations: the proximal-choice explanation (which included only proximate causes of the symptoms), the complete-choice explanation (which included all the causes they learned about), and an *unknown cause* explanation (see Table 1). They could select only one of the three response options. The order in which these explanations appeared was independently, randomly sampled from a uniform distribution over all possible orderings for each participant.

Explanation choice justifications. After indicating their explanation choice, participants were asked: "Why did you choose this explanation?" and could type a few sentences in a text box. We call this the *justification* of their explanation choice.

Reading Checks. Throughout the experiment, participants were asked a series of questions probing whether they accurately understood the causal information presented to them and ensuring they were reading the scenario closely. For example, in the chain-structure condition participants were asked whether it is possible to develop Tritchet's disease without having Hummel's disease (the answer is "yes"). If participants failed any reading checks their data were excluded from analyses (but see the online supplementary materials, Part D for analyses for all experiments involving less stringent exclusion criteria). The full set of reading checks and exclusion criteria can be found in the online supplementary materials, Part A, along with the proportion of participants failing reading checks across all experiments.

Results

Explanation choices. No participants selected the *unknown cause* explanation. As a result, a percentage of participants selecting the complete-choice (e.g., 80%) implies that the remaining participants (e.g., 20%) selected the proximal-choice. Overall,

Table 1
Explanation Choices, Prompts, and Response Options: Sample Stimuli From Experiment 1

Type of explanation choice	Chain-structure	Diamond-structure
	Prompt: What do you think is the <u>*most satisfying*</u> explanation for the symptoms Treda is exhibiting?	
Complete-choice	Treda **has Hummel's disease**, which **caused Tritchet's disease**, which caused the itchy flippets and purple spots.	Treda **has Hummel's disease**, which **caused Tritchet's disease and Morad's disease**, which together caused the itchy flippets and purple spots.
Proximal-choice	Treda **does not have Hummel's disease**, and **independently developed Tritchet's disease**, which caused the itchy flippets and purple spots.	Treda **does not have Hummel's disease**, and **independently developed Tritchet's disease and Morad's disease**, which together caused the itchy flippets and purple spots.
Unknown	Treda developed itchy flippets and purple spots but has **neither** of the aforementioned diseases.	Treda developed itchy flippets and purple spots but has **none** of the aforementioned diseases.

Note. Participants were randomly assigned to either the chain-structure condition or the diamond-structure condition; the explanation labels (e.g., complete-choice) were not presented to participants. The bolded words in the table are the diseases from our stimuli as styled when presented to participants.

participants selected the complete-choice 44% of the time in chain-structure, and 83% of the time in diamond-structure. We analyzed responses with χ^2 tests.

Participants selected the proximal-choice and the complete-choice about equally often in the chain-structure condition, $\chi^2(1) = 0.22$, $\varphi = 0.08$, $p > .5$, but selected the complete-choice significantly more often than the proximal-choice in the diamond-structure condition, $\chi^2(1) = 8.00$, $p < .01$. Reponses across the two causal structure conditions additionally differed significantly from each other, $\chi^2(1) = 5.89$, $\varphi = 0.40$, $p < .05$, with the complete-choice chosen more often in diamond-structure than in chain-structure. These findings are consistent with the predictions of root simplicity, but not with the predictions of node simplicity.

Explanation choice justifications. Three coders classified all participants' justifications for their explanation choices into one of four coding categories: "simplicity," "probability," "misunderstood," and "other." Justifications that explicitly appealed to simplicity, complexity, or the number of causes included in the explanation were coded as "simplicity." Justifications that referred to one of the options as being more "probable" or "likely" than the others were classified as "probability." Explanations that suggested the participant misunderstood some aspect of the experiment were classified as "misunderstood," and participants whose explanations fell into this category were excluded from additional analyses. For example, a participant would be classified as "misunderstood" in the chain-structure condition if she indicated that Treda must have Tritchet's disease *and* Hummel's disease because that is the *only* way to develop the symptoms. Finally, justifications that did not fall into one of the previous designations were classified as "other." Many of these restated the explanation choice (e.g., "Treda had Tritchet's disease and Hummel's disease which caused itchy flippets and purple spots"), or provided a response that appealed to neither simplicity nor probability, such as "it's what I remember reading from the paragraph," or claiming that it made most sense. Disagreements between coders were resolved in favor of the majority, with rare three-way ties resolved through discussion (Fleiss $\kappa = 0.63$, $z = 13.19$, $p < 10^{-4}$).

Overall, 11% of participants justified their choice by appeal to Simplicity, 33% by appeal to Probability, 0% were classified as misunderstood, and the remainder, 56%, fell under Other. The distribution of justifications did not vary as a function of causal structure; in fact, the frequencies of response types were identical across the two conditions. Of the small number of justifications that did appeal to simplicity ($N = 4$), two were used to support the proximal-choice in the chain-structure condition, none to support the complete-choice in the chain-structure condition, one to support the proximal-choice in the diamond-structure, condition, and one to support the complete-choice in the diamond-structure condition.

Discussion

The findings from Experiment 1 challenge the predictions of node simplicity but support those of root simplicity. Had participants been selecting explanations according to node simplicity, they should have preferred the proximal-choice (i.e., the explanation with fewer causes) in both conditions. Instead, participants were equally likely to choose the proximal-choice and the complete-choice in chain-structure, and significantly *less* likely to choose the proximal-choice in diamond-structure. These findings conform to the predictions of root simplicity (i.e., that people will prefer explanations with fewer *unexplained* causes), and thereby suggest that root simplicity better describes people's explanatory preferences than node simplicity, at least in these cases.

It is worth noting that only a small minority of participants (11%) explicitly justified their explanation choice by appeal to simplicity. This suggests that although root simplicity fits naturally within a philosophical and scientific tradition of characterizing *simplicity*, it may not correspond to laypeople's explicit beliefs about the extension of the term. Such a mismatch between intuitive judgments and explicit justifications is unlikely to be unique to simplicity in the evaluation of causal explanations—simplicity is invoked in explaining perceptual interpretations and concept learning (Feldman, 2000), for example, although it seems unlikely that participants would attribute their interpretation or inference to simplicity itself.

Experiment 2

While Experiment 1 challenges the claim that people choose explanations on the basis of node simplicity, the findings cannot differentiate two possibilities for why judgments were consistent with root simplicity. First, it could be that participants' explanatory preferences were a consequence of evaluating each explanation's root simplicity per se. Second, it could be that participants' preferences did not result *directly* from a preference for root-simpler explanations, but instead from assumptions about the relative probabilities of the complete-choice and the proximal-choice explanations; assumptions that happened to align with root simplicity. For example, participants could have assumed (in diamond-structure) that Morad's and Tritchet's diseases were unlikely to co-occur except in the presence of Hummel's disease, and therefore opted for the complete-choice over the proximal-choice on purely probabilistic grounds (i.e., without any recourse to simplicity per se). In Experiment 2, we address this possibility by providing frequency information to indicate how often different diseases occur together in the population at large. This allows us to flexibly adjust the baseline probability of alternative explanations before soliciting participants' explanation choices.

In the present experiment we did not present participants with isolated base-rates (as in Lombrozo, 2007), but instead with frequency information that represented the full joint distribution on diseases. Participants first learned a causal structure (chain-structure or diamond-structure) and then observed a random sample of aliens from the full population. Each alien's disease status (present/absent) was indicated for all diseases. By varying the disease status of the sampled aliens, we could control the relative probabilities of the proximal-choice and the complete-choice explanations for the target alien's symptoms, which participants were asked to explain as in Experiment 1. In this way participants received the frequency information necessary for assessing the

probability of each explanation without being told, explicitly, which explanation was most likely.[4]

Finally, this design allowed us to investigate the effects of explanation choice on memory for frequency information. After the explanation choice task, participants reported back the number of times they remembered having previously observed each combination of diseases in the alien population. Lombrozo (2007) found that some participants who selected simple explanations (specifically, those who selected simple explanations which were unlikely to be true) overestimated the frequency of the disease that figured in the simple explanation. Experiment 2 allowed us to investigate whether this effect would extend to cases in which participants were presented with information about the full joint distributions of diseases. It also provided an additional opportunity to differentiate root and node simplicity: if simplicity drives biases in memory for the frequency of causes invoked in simple explanations, these biases should track the simplicity metric that informs people's explanation choices.

Method

Participants. Using Amazon Mechanical Turk, 575 participants were recruited online as in Experiment 1. Of these, 50.6% passed the reading checks described below, leaving 291 participants for analysis.

We conducted a power analysis for this study based on the findings from the diamond condition of Experiment 1. If we consider the population size needed to achieve a power of $\beta = 0.8$ and $\alpha = .05$ based on a chi-squared test in the case where the data we present suggests (a priori) that complete and proximal explanations are equally likely, we obtain an estimate of 18 participants in that condition. This suggests that, with our average number of participants per condition of 29.1, we had sufficient power in Experiment 2.

Materials and procedure. Experiment 2 followed the materials and procedure from Experiment 1 closely. However, before participants were told about Treda and asked to make an explanation choice, they were provided with information about the frequencies at which the different diseases co-occurred. Specifically, participants observed 120 aliens that were described as having been randomly sampled from the population and tested for the presence of each disease. Participants were taught how to interpret images with multiple aliens, with the boxes below each alien indicating the presence or absence of the disease with the corresponding initial letter (see Figures 3a and 3b). If a box was yellow, that meant that the alien above had the disease indicated by that initial. Otherwise, the alien did not have that disease. Participants were tested to ensure that they understood the representation system as part of our larger set of reading checks.

Participants were told that "the particular incidence rates of these diseases are unknown," but that "to address this issue, the hospital you work in is running diagnostic tests on a random sample of the population" (for complete instructions, see the online supplementary materials, Part B). The aliens were then presented in 12 groups of 10, with each group appearing for 3.5 seconds between 2-s breaks. In pilot testing we confirmed that these intervals allowed participants to view all aliens while discouraging explicit counting.

We varied the actual frequency information that participants viewed across five between-subjects conditions, the 3:1, 2:1, 1:1, 1:2, and 1:3 conditions, named for the corresponding ratios of the degree to which the evidence supports choosing the proximal-choice versus complete-choice (all frequency counts can be seen in Table 2). To compute these *support ratios*, we defined the probability of an explanation as the percentage of times that the exact pattern of diseases corresponding to that explanation appeared in the data that participants observed. For example, in the chain-structure in the 3:1 condition, the *support ratio* is:

$$\mathrm{P}(\textit{proximal}\,|\,\mathrm{data}) : \mathrm{P}(\textit{complete}\,|\,\mathrm{data}) = \mathrm{P}(\neg \mathrm{H}, \mathrm{T}\,|\,\mathbf{D}_{3:1}, \mathrm{S}) : \mathrm{P}(\mathrm{H}, \mathrm{T}\,|\,\mathbf{D}_{3:1}, \mathrm{S}) = 3 : 1,$$

where

- H and T mean that Treda has Hummel's disease and Tritchet's diseases,
- $\neg$ is the negation operator,
- $\mathbf{D}_{3:1}$ is the frequency data from the 3:1 condition,
- and S indicates the presence of the observed symptoms.

Analogously, in diamond-structure the *support ratio* would be between

$$\mathrm{P}(\neg \mathrm{H}, \mathrm{T}, \mathrm{M}\,|\,\mathbf{D}_{3:1}, \mathrm{S}) : \mathrm{P}(\mathrm{H}, \mathrm{T}, \mathrm{M}\,|\,\mathbf{D}_{3:1}, \mathrm{S}) = 3 : 1.$$

The frequencies in Table 2 were chosen such that the number of cases supporting the proximal-choice versus complete-choice corresponded to the *support ratio* appropriate for each condition, and for diamond-structure, so that the frequencies of M and T were equally likely and approximately conditionally independent given ¬H (to avoid inadvertently suggesting that there existed an additional common cause for these diseases' co-occurrence).

Explanation choice. As in Experiment 1, participants were asked to identify the most satisfying explanation for Treda's two symptoms.

Explanation choice justification. Also as in Experiment 1, we asked participants to justify their choice in a free-response format.

Estimated frequency counts. Participants were told that they originally observed 120 aliens and asked to indicate how many of these observed aliens belonged to each diagnostic option (presented with its corresponding image), with four possible disease combinations in chain-structure (e.g., H but not T) and eight in diamond-structure (e.g., H, M, and T).

Reading checks. The reading checks from Experiment 1 were employed again in Experiment 2. In addition, if participants' responses to the frequency estimate question did not add up to 120 (the correct number) or to 100 (implying a probabilistic interpretation of the question, which we renormalized to add up to 120),

[4] It is worth noting that the central findings from Lombrozo (2007) involved two sources of probabilistic uncertainty. On the one hand, participants may have been unsure whether the two diseases in the two-disease condition were probabilistically independent, and therefore whether their joint probability was well approximated by the product of their probabilities. On the other hand, participants' "uncertainty" could have stemmed from a more global tendency to rely on an intuitive evaluation of probability when one has to deal with a complex evaluation of multiple sources of evidence. In Experiment 2, we isolate the role of the latter source of uncertainty by eliminating the first: we present participants with data about the full joint probability distribution for the diseases relevant to chain-structure and diamond-structure.

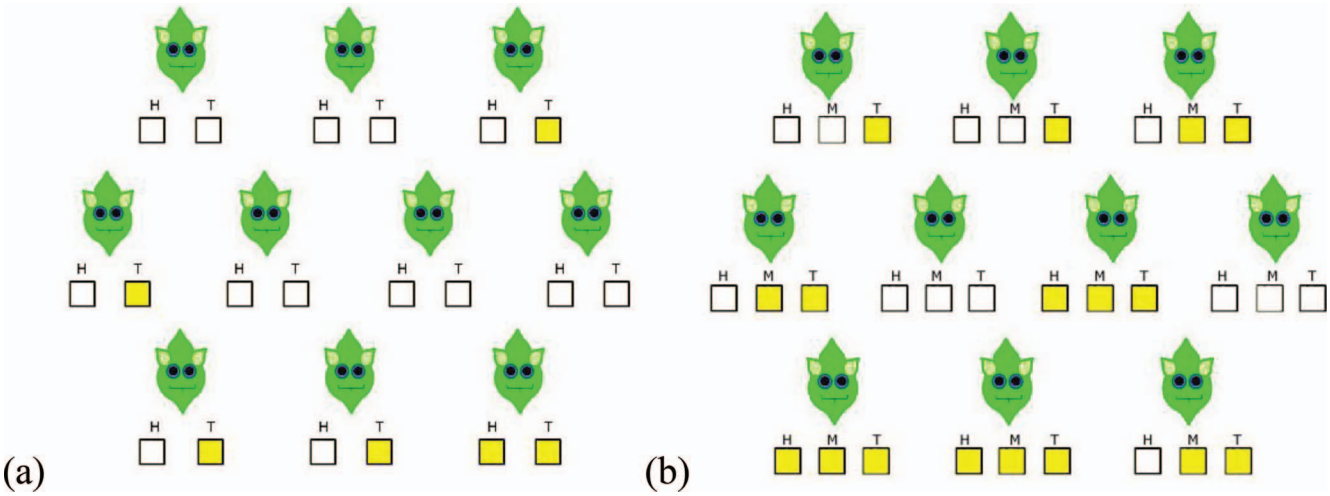

Figure 3. (a) Example of a group of 10 aliens from chain-structure. (b) Example of a group of 10 aliens from diamond-structure. Yellow boxes indicate the presence of a disease. For example, the top right alien in 2b has Tritchet's disease (T) and Morad's disease (M), but not Hummel's disease (H).

they were excluded for failing to follow instructions, and because their inclusion would considerably complicate the analysis and interpretation of the data.

Results

Explanation choices. All participants who passed the reading checks selected either the proximal-choice or complete-choice explanations. To analyze explanation choices, we first computed the logarithm of the *support ratio* (*log-support-ratio*) in each condition. The *log-support-ratio* should account for explanation choices under two assumptions: first, that participants' explanation choices were a function of the true frequency information provided, and not (e.g.) a preference for node or root simplicity, and second, that participants "probability matched"—that is, that they chose explanations in proportion to their probability of being true, which is a common strategy in many human judgments (cf. Eberhardt & Danks, 2011) and was a useful assumption in interpreting the findings from Lombrozo (2007) and Bonawitz and Lombrozo (2012). A systematic deviation from the explanation choices predicted by probability matching would therefore suggest that something other than frequency information (e.g., root simplicity) plays a role in explanation choice.

Table 2
The Frequencies With Which Each Disease Combination was Presented for Each Support Ratio for the Diamond-Structure (Experiments 2 and 3) and Chain-Structure (Experiment 2) Conditions

	Frequency				
Event type	3:1	2:1	1:1	1:2	1:3
Experiments 2 and 3: Diamond-structure					
¬H, ¬M, ¬T	7	17	33	50	57
¬H, M, T	54	36	18	9	6
H, ¬M, ¬T	1	1	1	1	1
H, M, T	18	18	18	18	18
¬H, M, ¬T	20	24	25	21	19
¬H, ¬M, T	20	24	25	21	19
H, M, ¬T	0	0	0	0	0
H, ¬M, T	0	0	0	0	0
Experiment 2: Chain-structure					
¬H, ¬T	47	65	83	92	95
¬H, T	54	36	18	9	6
H, ¬T	1	1	1	1	1
H, T	18	18	18	18	18

Note. All columns add to 120 (the total sample). H = Hummel's disease; M = Morad's disease; T = Tritchet's disease; "¬" indicates the absence of a disease.

We conducted a regression (a generalized linear model) on explanation choices with three predictors: *log-support-ratio*, a categorical variable designating each participant's structure (chain-structure or diamond-structure), and an interaction term to assess whether participants used frequency data differently across structures.

The regression, with the proportion of complete-choice selections as the dependent variable, revealed no significant intercept, $t(287) = -0.286$, $\beta = -0.0515$, $p > .7$, a significant coefficient for *log-support-ratio*, $t(287) = 5.110$, $\beta = 1.185$, $p < 10^{-4}$, a significant effect of the categorical variable corresponding to causal structure, $t(287) = 3.299$, $\beta = 0.849$, $p < .001$, and a significant interaction between *log-support-ratio* and causal structure, $t(287) = -2.075$, $\beta = -0.6811$, $p < .05$ (see Figure 4).[5]

The effect of *log-support-ratio* suggests that frequency information had a significant effect on participants' explanation choices, increasing the probability of choosing the complete-choice explanation when it was more frequent in past observations. However, the interaction between *log-support-ratio* and causal structure suggests that the influence of frequency information on

[5] Technically, this analysis necessitates calculating different interaction effects at each point in question (see Ai & Norton, 2003); however, an interaction effect at the intercept is sufficient for our purposes.

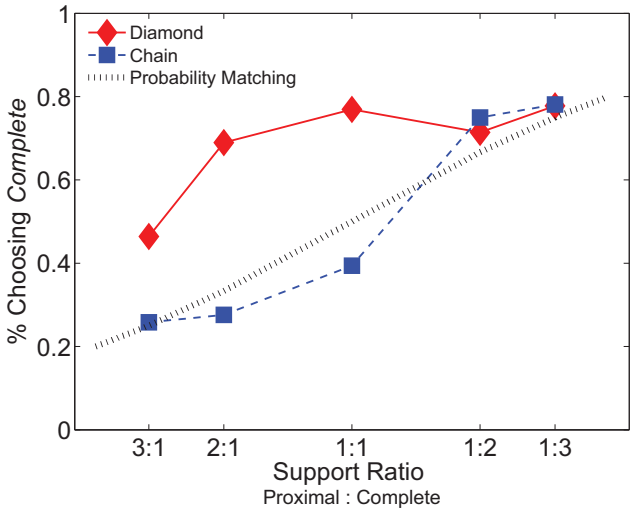

Figure 4. Graph of explanation choices, % of Participants Choosing Complete × Support Ratio (mapped to the *x*-axis as log(Y/X) for Y:X, centered at 0 = log(1/1)).

explanation choices was not equivalent across conditions. We therefore conducted two subsequent regression analyses, treating participants from chain-structure and diamond-structure independently.

For chain-structure, the analysis revealed no significant intercept, $\beta = -0.0515$, $t(151) = -0.2861$, $p > .7$, but a significant effect of *log-support-ratio*, $t(151) = 5.110$, $\beta = 1.186$, $p < 10^{-4}$. The coefficient for *log-support-ratio* did not differ significantly from 1 (95% confidence interval [0.722, 1.650]). This analysis suggests that participants' explanation choices in chain-structure were well captured by probability matching based on the frequency information that participants received. In other words, the data from chain-structure provide no evidence of a preference for either the proximal-choice or complete-choice explanations (above and beyond their frequency), which contrasts with the predictions of node simplicity, but is consistent with those of root simplicity.

For diamond-structure, an equivalent analysis revealed a significant intercept, $\beta = 0.797$, $t(136) = 4.224$, $p < 10^{-4}$, as well as a significant effect of *log-support-ratio*, $\beta = 0.504$, $t(136) = 2.192$, $p < .05$. In this case, the coefficient for *log-support-ratio* did differ from 1 (95% confidence interval for $\beta = [0.044, 0.965]$). These results suggest that *log-support-ratio* accounted for some variation in explanation choices, but that participants were significantly more likely to choose the complete-choice explanation than expected on the basis of the frequency information alone. An analysis of the nonzero intercept suggests that participants effectively operated with a prior probability of 0.69 (95% confidence interval for $\beta = [0.603, 0.764]$) favoring the explanation deemed simpler according to root simplicity. This concords with the estimates of the prior probability of a simpler explanation as reported in Lombrozo (2007). These findings also challenge the predictions of node simplicity, but support those of root simplicity.

Deviations from the predictions of *log-support-ratio* in diamond-structure were not uniform across support ratios. Post hoc, one-sample *t* tests comparing the proportion of complete-choice explanations for each *log-support-ratio* to the proportion expected from probability-matching revealed significantly higher selection of the complete-choice explanations in the 3:1, $t(27) = 2.619$, Cohen's $d = 1.008$, $p < .01$; 2:1, $t(28) = 4.071$, Cohen's $d = 1.539$, $p < 10^{-4}$; and 1:1, $t(25) = 2.746$, Cohen's $d = 1.098$, $p < .01$, cases, but not for 1:2, $t(27) = 0.535$, Cohen's $d = 0.205$, $p > .5$; or 1:3, $t(26) = 0.333$, Cohen's $d = 0.131$ $p > .7$. In other words, participants' explanation choices involved a significant departure from the predictions of probability matching only when frequency information did not favor the root-simpler explanation.

Explanation choice justifications. Explanation choice justifications were coded as in Experiment 1 (see also the online supplementary materials, Part C). There was moderate agreement among the raters (returning all instances of "Misunderstood" to the dataset that were not excluded for other reasons; Fleiss $\kappa = 0.4415$, $z = 29.46$, $p < 10^{-4}$). The distribution of explanation justifications can be found in Table 3. We found a significant difference between the overall justification distributions across causal structures, $\chi^2(308) = 8.7738$, $p < .05$, with participants more likely to invoke probability in chain-structure than in diamond-structure.

As in Experiment 1, the proportion of justifications that appealed to simplicity was quite small (8%, $N = 25$). Of these, 14 were used to support the proximal-choice in the chain-structure condition, zero to support the complete-choice in the chain-structure condition, eight to support the proximal-choice in the diamond-structure, condition, and three to support the complete-choice in the diamond-structure condition.

Reported frequencies: Bias for complete-choice over proximal-choice. The frequency estimates that participants reported at the end of the task were analyzed as a function of both the actual frequencies (corresponding to each *log-support-ratio* condition) and participants' individual explanation choices. We considered the extent to which participants overestimated the complete-choice explanation relative to the proximal-choice explanation.

First, we computed the *true difference* between the number of observed cases that corresponded to the proximal-choice explanation and subtracted that from the number of cases corresponding to the complete-choice explanation for a given support ratio. Next, we computed an *estimated difference* by subtracting the number of proximal-choice-consistent cases that a participant estimated having seen from their estimate of the number of complete-choice-consistent cases. Because *estimated difference* should reflect a combination of *true difference* and any biases in memory or reporting, we subtracted the *true difference* from *estimated difference* to create a normalized measure of participants' memory bias for the complete-choice explanation, which we refer to as *bias*. A positive value for the *bias* term would result from overestimating the complete-choice-consistent cases or underestimating the

Table 3
Distribution of Explanation Justifications for Experiment 2

Justification type	Overall	Chain-structure	Diamond-structure
Simplicity	8.0%	8.9%	7.2%
Probability	52.4%	58.2%	46.4%
Other	33.1%	29.7%	36.6%
Misunderstood	6.4%	3.1%	9.8%

proximal-choice-consistent cases (or both), whereas a negative value suggests the reverse. A perfect estimate would receive a score of 0 in all frequency conditions.

We analyzed *bias* with a linear regression model, using *log-support-ratio* and causal structure (chain-structure vs. diamond-structure) as continuous and categorical independent variables, respectively, and *choosing-complete* as a categorical factor. This analysis revealed a nonsignificant intercept, $t(286) = 0.430$, $\beta = 0.695$, $p > 0.6$, a significant effect of *log-support-ratio*, $t(286) = -4.553$, $\beta = -7.477$, $p < .001$, a significant effect of causal structure, $t(286) = 2.525$, $\beta = 4.772$, $p < .05$, a significant effect of *choosing-complete*, $t(286) = 4.243$, $\beta = 8.636$, $p < 10^{-5}$, and a significant interaction between *log-support-ratio* and causal structure, $t(286) = -3.990$, $\beta = -9.051$, $p < .001$. No other interactions were significant (*p*s > 0.4). Given the interaction between *log-support-ratio* and causal structure, and to facilitate the interpretation of these results, we conducted follow-up analyses restricted to each of the causal structure conditions and analyzed with respect to *log-support-ratio* and *choosing-complete* (see Figure 5).

In the chain-structure condition, the analysis revealed a nonsignificant intercept, $t(150) = 0.71$, $\beta = 1.40$, $p > .4$, and significant effects of both *log-support-ratio,* $t(150) = -3.90$, $\beta = -7.079$, $p < .01$, and *choosing-complete*, $t(150) = 2.40$, $\beta = 7.178$, $p < .05$. The more strongly the data supported choosing complete the less bias we observed, and those who chose complete were more biased toward complete in their estimates.

In the diamond-structure condition, the analysis revealed a marginally significant intercept, $t(135) = 1.95$, $\beta = 4.402$, $p < .10$, and significant effects of both *log-support-ratio*, $t(135) = -10.74$, $\beta = -16.684$, $p < 10^{-4}$, and *choosing-complete*, $t(135) = 3.71$, $\beta = 10.19$, $p < .01$. As in chain-structure, the more strongly the data supported choosing complete the less bias was observed, and those who chose complete were more biased toward complete in their estimates. However, as indicated by the original interaction, *log-support-ratio* had a stronger effect in the diamond-structure than in the chain-structure.

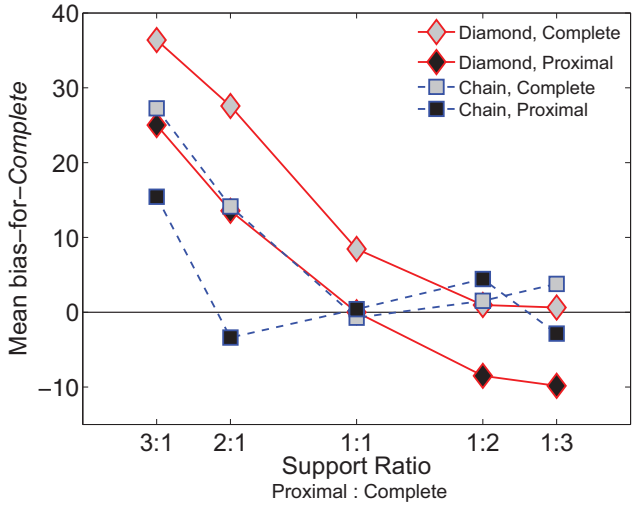

Figure 5. Graph of average bias-for-complete values by support ratio, split by causal structure and explanation choice.

Discussion

Experiment 2 replicated and extended our findings from Experiment 1. First, participants were no more likely to select the proximal-choice explanation than expected on the basis of probability matching in any condition, challenging the predictions of node simplicity. However, in the diamond-structure condition, participants were more likely to select the complete-choice explanation than expected on the basis of probability matching, consistent with the predictions of root simplicity. Thus, even when participants are presented with information that (noisily) supported the alternative hypothesis, we see a preference for root simplicity. However, as in Experiment 1, participants rarely justified their explanation choice by explicit appeal to simplicity.

Second, consistent with the findings from Lombrozo (2007), participants' explanation choices were a function of *both* simplicity and frequency data. In particular, the proportion of participants selecting the complete-choice explanation was influenced by *log-support-ratio* in *both* chain-structure and diamond-structure. However, a systematic deviation from the predictions of probability matching emerged in the three *log-support-ratios* involving diamond-structure for which the probability information did not warrant the complete-choice: the 3:1, 2:1, and 1:1 conditions.

Third, Experiment 2 revealed a systematic bias in memory: participants who chose the root-simpler explanation sometimes misremembered their observations as more consistent with the root-simpler explanation than they in fact were. This bias emerged in the three conditions for which the evidence did not independently support the root-simpler explanation: the 3:1, 2:1, and 1:1 conditions. These were also the conditions for which explanation choices deviated from probability matching. In Experiment 3, we consider whether estimation bias was a *consequence* of participants' explanation choices.

Experiment 3

Experiment 2 found that those participants who chose a simple explanation that was not supported by observed data also systematically misremembered the data as more consistent with their explanation than it actually was (see also, Lombrozo, 2007). This finding is consistent with the idea that explanation choices can systematically alter memory, but it could also be that systematic distortions in memory have implications for explanation choice (or that both explanation choice and memory for observations have a common cause).

Here we address these alternatives by varying the order in which participants choose an explanation and are asked to estimate the observed frequency data. Those who estimate before explaining provide a baseline against which we can compare the estimates of those who explain first. If we find that memory distortions are larger for participants who *explain first* than for those who *estimate first*, this suggests that the act of explaining causes (or contributes to) these distortions. If we instead find that distortions are equivalent in both groups, that

would suggest that distortions causally contribute to explanation choices and/or that both distortions and explanation choices have a common cause.

Method

Participants. Three-hundred-eighty-nine participants were recruited via Amazon Mechanical Turk as in Experiments 1–2. Of these, 43.2% passed the reading checks, leaving 168 participants for analysis. This provides 28 participants on average per condition (for our 6 conditions), which is comparable to Experiment 2's 29.1 participants on average per condition.

Materials and procedure. The materials and procedure mirrored those from the 3:1, 1:1, and 1:3 diamond-structure conditions of Experiment 2, with the following changes. First, we varied the order in which participants were asked to provide their explanation choice and frequency estimates: participants in the *explain-first* condition learned about Treda and indicated an explanation choice before providing frequency estimates (as in Experiment 2). Participants in the *estimate-first* condition were asked to report observed frequencies before they learned about Treda or explained Treda's symptoms.

Second, to ensure that there were equal time delays between observing and reporting frequency data in both ordering conditions, we added an additional explanation choice question which took the place of the alien explanation choice in the *estimate-first* condition. This new question was equivalent to the explanation choice task in terms of time and structure, but irrelevant to the subsequent frequency estimation task. We told participants that a human named Pat was sneezing and asked them for a diagnosis from the following possibilities: "Pat has the flu, which caused her sneezing," "Pat has a cold, which caused her sneezing," and "Pat does not have either of these diseases, her sneezing was caused by something unknown." Because this question is irrelevant to the aims of the study, we do not analyze these data.

Reading checks. Experiment 3 employed the same reading checks as Experiment 2.

Results

Explanation choices. Explanation choices replicated those of Experiment 2 (see Figure 6), with a significant intercept ($p < .001$), a significant effect of log-support-ratio ($p < .005$) and no significant effect of task order ($p > .5$).[6]

Explanation choice justifications. Justifications were coded as in Experiments 1–2, yielding moderate agreement among coders ($\kappa = 0.577$, $z = 35.58$, $p < 10^{-4}$). The justification distributions differed between the explain-first and the estimate-first conditions, $\chi^2(185) = 7.9078$, $p < 0.05$, with participants more likely to provide Other justifications in *estimate-first* (see Table 4). As in Experiments 1–2, the proportion of justifications that appealed to simplicity was quite small (4.8%, $N = 9$), with the following distribution across conditions and explanation choices: two were used to support the proximal-choice in the *explain-first* condition, two to support the complete-choice in the *explain-first* condition, three to support the proximal-choice in the *estimate-first* condition, and two to support the complete-choice in the *estimate-first* condition.

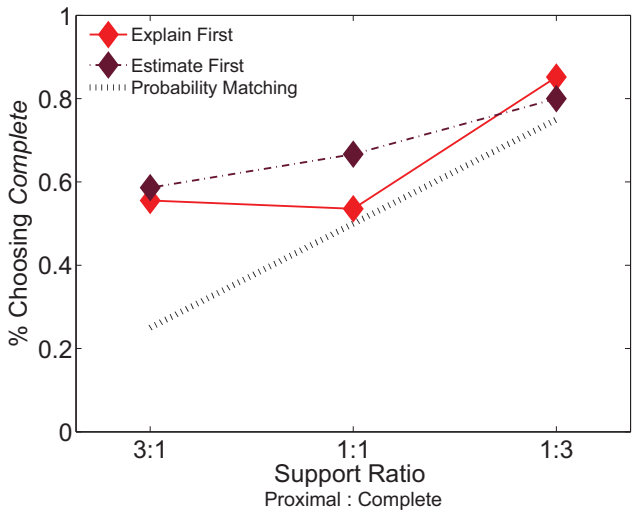

Figure 6. Graph of explanation choices, % of Participants Choosing Complete × Support Ratio (mapped to the *x*-axis as log(Y/X) for Y:X, centered at 0 = log(1/1)).

Frequency estimates: Bias for complete-choice over proximal-choice. As in Experiment 2, we analyzed the magnitude of a bias for the complete-choice (see Figure 7). We used *task-order*, *choosing-complete*, *log-support-ratio,* and paired interactions between these variables as predictors of the extent to which participants overestimated the frequency of evidence consistent with the complete-choice over the proximal-choice. As in Experiment 2, the measure of overestimation that we used was "bias," the difference between the estimated difference and the true difference in the number of observations favoring the complete-choice over the proximal-choice.

Most critically, Experiment 3 revealed a significant interaction between *choosing-complete* and *task-order*, $t(162) = 2.961$, $\beta = 17.090$, $p < .01$, although the effect of *task-order* itself was only marginally significant, $t(162) = -1.929$, $\beta = -9.089$, $p < .06$.

[6] As in Experiment 2, we analyzed explanation choices using logistic regression with *log-support-ratio* as a predictor for the proportion of participants selecting the complete-choice explanation. However, we additionally included *task-order* (*explain-first* vs. *estimate-first*) as a predictor, as well as an interaction term between *log-support-ratio* and *task-order*. This analysis revealed a significant intercept, $t(165) = 3.413$, $\beta = 0.819$, $p < 0.001$, as well as a significant coefficient for *log-support-ratio*, $t(165) = 2.815$, $\beta = 0.536$, $p < 0.005$. This suggests that participants chose the complete-choice more often than expected on the basis of probability matching, but were additionally sensitive to *log-support-ratio*, with a larger proportion of participants selecting the complete-choice when it was more likely to be true. These findings replicate those from the diamond-structure condition in Experiment 2. We did not find significant effects of *task-order*, $t(165) = -0.543$, $\beta = -0.183$, $p > 0.50$, suggesting that this manipulation did not have a large impact on explanation choices.

The remaining effects[7] of the regression largely replicated those of Experiment 2, and average error across all event types was not influenced by task order.[8]

To better understand the interaction between *choosing-complete* and *task-order*, we performed separate analyses for each *support ratio* condition. For the 1:3 condition, in which the frequency evidence supported the complete-choice over the proximal-choice, we found a significant intercept, $t(53) = -3.091$, $\beta = -16.667$, $p < .01$, a significant effect of *choosing-complete,* $t(53) = 2.691$, $\beta = 16.225$, $p < .01$, and no significant effect of *task-order (explain-first),* $t(53) = -0.420$, $\beta = -3.583$, $p > .6$. Participants had an overall bias for the complete-choice, and had a larger bias if they in fact chose the complete-choice. However, there was no interaction between *choosing-complete* and *task-order* (*explain-first*), $t(53) = -0.245$, $\beta = -2.295$, $p > .8$, suggesting that in the 1:3 case, where the evidence was sufficient to justify the complete-choice, the bias that emerged was not a consequence of committing to the complete-choice in the explanation task.

In the remaining two *support ratio* conditions, 1:1 and 3:1, the evidence did not favor the complete-choice, and we would therefore anticipate a greater role for explicit explanation choices on frequency estimation, as found in Experiment 2. Consistent with this prediction, in both the 1:1 and 3:1 conditions we found significant intercepts, $t(51) = -3.670$, $\beta = -21.778$, $p < .001$, and $t(52) = 12.110$, $\beta = 35.500$, $p < 10^{-4}$, and significant interaction effects between *choosing-complete* and *task-order* (*explain-first*), $t(51) = 2.706$, $\beta = 26.836$, $p < .01$, and $t(52) = 2.284$, $\beta = 12.536$, $p < .05$. In these *support ratio* conditions, participants exhibited a larger estimation bias if they completed the explanation choice task prior to the frequency estimation task and chose the complete-choice.

These analyses also revealed that in the 1:1 condition, there was a marginal main effect of *choosing-complete*, $t(51) = 1.972$, $\beta = 14.333$, $p < .1$. There was no main effect of *task-order* (*explain-first*), $t(51) = -0.517$, $\beta = -3.992$, $p > .5$, on bias. In the 3:1 condition, there was a marginal main effect of *task-order* (*explain-first*), $t(52) = -1.729$, $\beta = -7.167$, $p < .1$: when the explanation choice task was first, there may have been lower bias among participants who chose the proximal-choice than that for those who chose complete (see Figure 6). There was no main effect of *choosing-complete* ($p > .9$).

Discussion

Experiment 3 replicated key findings from Experiment 2: participants were significantly more likely to choose the complete-choice in diamond-structure than predicted on the basis of the support ratios and probability matching, with explanation choices modulated by the actual support ratios. This is consistent with the idea that root simplicity and frequency information jointly inform explanation choices.

Table 4
Distribution of Explanation Justifications for Experiment 3

Justification type	Overall	Explain-first	Estimate-first
Simplicity	4.8%	4.2%	5.4%
Probability	43.6%	49.0%	38.0%
Other	41.0%	32.3%	50.0%
Misunderstood	10.6%	14.6%	6.5%

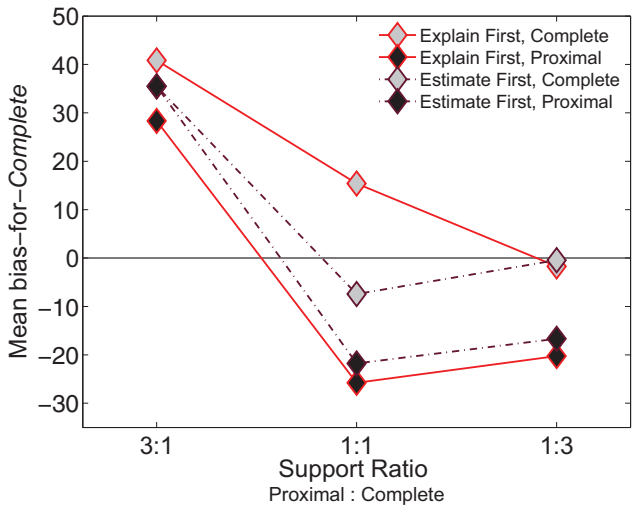

Figure 7. Graph of average bias-for-complete values by support ratio, split by task-order and Explanation Choice.

Experiment 3 also went beyond Experiment 2 in considering the causal relationship between choosing an explanation and memory. Participants who provided an explanation before estimating frequencies, who chose the complete-choice explanation, and who were in a *support ratio* condition that did not favor the complete-choice were most likely to exhibit a large bias for the complete-choice (over the proximal-choice) in their frequency estimates. Notably, these effects were additive: participants were most biased when all three factors co-occurred. These findings not only provide converging evidence of a powerful human preference for root simplicity, but support the idea that its effects extend to judgments beyond the explicit evaluation of explanations.

[7] In Experiment 3, the interaction term between *task-order* and *log-support-ratio* was not a significant predictor, and was thus removed from the analysis, $t(161) = -1.112$, $\beta = -3.444$, $p > .2$. In the resulting analysis, the intercept was not significant, $t(162) = -0.588$, $\beta = -2.036$, $p > .5$, suggesting that overall bias did not differ from zero. However, *choosing-complete*, $t(162) = 2.892$, $\beta = 12.028$, $p < .01$, and *log-support-ratio*, $t(162) = -9.663$, $\beta = -28.062$, $p < 10^{-9}$, were significant predictors of bias: participants had a greater bias for the complete-choice in their reported frequencies if they chose the complete-choice as the better explanation or if they were in a *support ratio* condition that favored the proximal-choice. These findings mirror those from Experiment 2, though here we additionally found an interaction between the effects of *choosing-complete* and *log-support-ratio*, $t(162) = 3.209$, $\beta = 11.011$, $p < .005$, with the greatest bias favoring the complete-choice-consistent cases among participants who selected the *complete-choice* when it was unlikely to be true.

[8] We used a generalized linear model with *task-order*, *choosing-complete*, and *log-support-ratio* as predictors for participants' average absolute error rates across all eight event types. This analysis revealed a significant intercept, $t(164) = 20.317$, $\beta = 95.181$, $p < 10^{-9}$, indicating that error was significantly greater than zero, and a significant coefficient for *log-support-ratio*, $t(164) = -4.380$, $\beta = -11.430$, $p < 10^{-4}$, indicating that error was greater for conditions that favored the proximal-choice. Neither *task-order*, $t(164) = -0.471$, $\beta = -2.164$, $p > .5$, nor *choosing-complete*, $t(164) = -1.189$, $\beta = -5.936$, $p > 0.2$, were significant predictors of absolute error.

Experiment 4

Why might people favor simpler explanations, especially when doing so appears to have negative consequences for the fidelity of memory (Experiment 3)? Experiment 4 explores one hypothesis about why explanations with few *root* causes, in particular, might be preferred. We propose that, in general, explanations with fewer root causes provide a useful way to compress information about a causal system for the purposes of memory storage, diagnosis, communication, and intervention. This proposal is related to *explanation for export* (Lombrozo & Carey, 2006), a hypothesis that suggests explanations are tailored to support predictions and interventions, and so explanations should privilege *exportable* causal information—that is, information that can be exported from the current situation to support prediction and intervention in novel scenarios (see also Lombrozo, 2006). Root causes are prima facie good candidates for exportable causes: they can be used to predict downstream effects, and they make good candidates for interventions intended to have wide-reaching effects (for information-theoretic analyses of interventional loci, see *causal information flow* in Ay & Polani, 2008; in the context of explanation choice, see Pacer, Williams, Lombrozo, & Griffiths, 2013).

If root simplicity is instrumentally valuable—via its relation to effective prediction and intervention—then a preference for root simplicity should be moderated by the degree to which a "root" cause predicts and controls its effects. Specifically, the preference for root simplicity should vary as a function of causal strength, with a stronger preference as the strength of a root cause increases (for more about causal strength see also, Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). We test this prediction in Experiment 4.

Method

Participants. Two-hundred-and-five participants were recruited via Amazon Mechanical Turk as in Experiments 1–3. Of these, 57.1% passed the reading checks, leaving 117 participants for analysis.

Because one of the primary intents of Experiment 4 is to demonstrate a weakening of the effect of root simplicity due to nonstructural factors that were not manipulated in previous experiments, Experiments 1–3 do not provide ideal bases for a power analysis. The closest comparison is the sample size needed to achieve $\beta = 0.8$ and $\alpha = .05$ in the 2:1 diamond condition of Experiment 2 (which had an observed proportion of 0.6897 choosing Complete compared with an expected proportion of .3333), which yields a sample size of at least 17 participants. This suggests that our sample size of 39 participants per condition in Experiment 3 provides sufficient power to replicate the effect.

Materials and procedures. The materials and procedure were very similar to the 2:1 diamond-structure condition from Experiment 2, and the support ratio was held constant across conditions in Experiment 4. However, the frequency data were varied across three conditions that corresponded to different levels of causal strength between *H* and *M & T*: *weak*, *moderate*, and *strong*.

There are several different metrics for causal strength, all of which try to capture the intuition that some causal relationships are stronger than others. Common metrics include ΔP (Cheng & Novick, 1990) and causal power ("Power-PC"; Cheng, 1997), which once modified to apply to our case scenarios, could be defined as:

$$\Delta P = P(M,\ T|D,\ H) - P(M,\ T|D,\ \neg H).$$

$$Power = \frac{\Delta P}{1 - P(M,\ T|D,\ \neg H)},$$

for positive values of ΔP, and for negative values, Power-PC is calculated as:

$$Power = \frac{\Delta P}{P(M,\ T|D,\ \neg H)}.$$

Across strength conditions, participants received data consistent with a *weak* causal relationship (ΔP ≈ .02 and Power-PC ≈ .03), a *moderate* causal relationship (ΔP ≈ .28 and Power-PC ≈ .45), or a *strong* causal relationship (ΔP ≈ .59 and Power-PC ≈ .91). The final case was identical to the 2:1 diamond-structure condition from Experiment 2 (see Table 5 for exact frequency counts).

Reading checks. Experiment 4 involved the same reading checks as Experiment 2.

Results

Explanation choices. We conducted analyses similar to those in Experiment 2 (see Figure 8). However, because the *support ratio* was held constant while *causal strength* varied, we used the latter as a predictor for explanation choices. Participants were more likely to choose the complete-choice as causal strength increased, whether causal strength was measured using ΔP, $t(115) = 2.900$, $\beta = 2.668$, $p < .005$, or Power-PC, $t(115) = 2.895$, $\beta = 1.709$, $p < .005$.[9] The intercepts were not significantly different from 0, suggesting that there was not a baseline preference for one explanation over another across all conditions in this experiment ($ps > 0.9$). However, even when the causal strength was weak, participants selected the complete-choice explanation more often than the frequency predicted by probability matching ($\mu = 0.525$, $N = 40$, $z = 2.5715$, $p < .05$).

Explanation choice justifications. Justifications were coded as in Experiments 1–3, with substantial agreement between the three raters ($k = 0.7484$, $z = 24.538$, $p < 10^{-4}$). Overall, justifications invoked simplicity in 1.6% of cases, probability in 52.8%, and other justifications in 40.7%. The remaining 4.9% of participants who passed other reading checks provided explanations that were designated as misunderstood, and were therefore excluded from other analyses. There were two people who justified

[9] We reanalyzed Experiment 2 using causal strength as a predictor and found largely the same effects. For both ΔP and power-PC, causal strength, condition and the intercept were significant ($df = 288$, $ps < 0.01$). It is worth noting that previous experiments did not manipulate log-support-ratio independently of causal strength; indeed, in all conditions of Experiments 2 and 3, causal strength and log-support-ratio were highly correlated. This results directly from the constraints that we imposed in generating the frequency distributions to represent the different log-support-ratios, namely: having a constant total frequency, a single event in which the root cause occurred and did not in turn cause the proximal disease(s), and (in the diamond structure conditions) holding the conditional probabilities of the diseases to be approximately independent given that the root cause was not present (¬H) so as not to suggest alternative latent common-cause mechanisms for bringing about the proximal diseases. In light of the systematic correspondences and deviations from the predictions of probability matching (derived from the log support ratios), we think it is unlikely that causal strength alone explains our results in Experiments 2–3.

Table 5
Frequency Conditions for Experiment 4

Event type	Frequency		
	Strong	Moderate	Weak
¬H, ¬M, ¬T	17	13	9
¬H, M, T	36	36	36
H, ¬M, ¬T	1	9	21
H, M, T	18	18	18
¬H, M, ¬T	24	22	18
¬H, ¬M, T	24	22	18
H, M, ¬T	0	0	0
H, ¬M, T	0	0	0

Note. "¬" indicates the absence of a disease.

their explanation choice with reference to simplicity; one who chose complete, and one who chose proximal.

Reported frequencies: Bias for complete-choice over proximal-choice. Each individual's bias for evidence consistent with the complete-choice over the proximal-choice was calculated as in Experiments 2 and 3 (see Figure 9). Bias was analyzed in a regression with *causal strength* (Power-PC) and *choosing-complete* as predictors. This analysis revealed a significant coefficient for *causal strength*, $t(114) = 2.844$, $\beta = 11.663$, $p < .01$, as well as for *choosing-complete*, $t(114) = 4.454$, $\beta = 14.034$, $p < 10^{-4}$: participants overestimated the evidence for the complete-choice to a larger degree when the causal strength was greater, and also when they selected the complete-choice. The intercept was not significant, $t(114) = 1.656$, $\beta = 4.637$, $p > 0.1$. A parallel analysis using ΔP values instead of Power-PC yielded equivalent results.

Discussion

Experiment 4 varied the causal strength of the relationship between the candidate root cause in the diamond-structure (i.e., Hummel's disease) and its two potential effects (i.e., Tritchet's and Morad's diseases). As predicted, we found that as causal strength increased, so too did participants' preference for the complete-choice (the root-simpler explanation), even though the support ratio remained constant at 2:1. This finding is consistent with the idea that a preference for root simplicity derives from the goal of efficiently representing exportable causal information, including causes that effectively predict their effects and support maximally effective and efficient interventions.

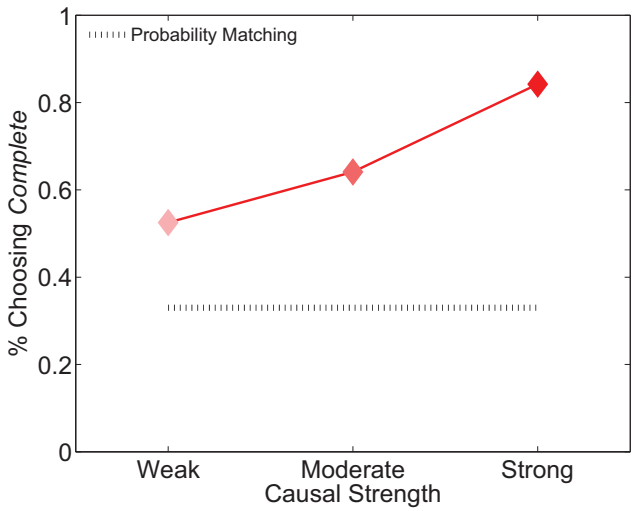

Figure 8. Graph of explanation choices, % of Participants Choosing Complete × Causal Strength.

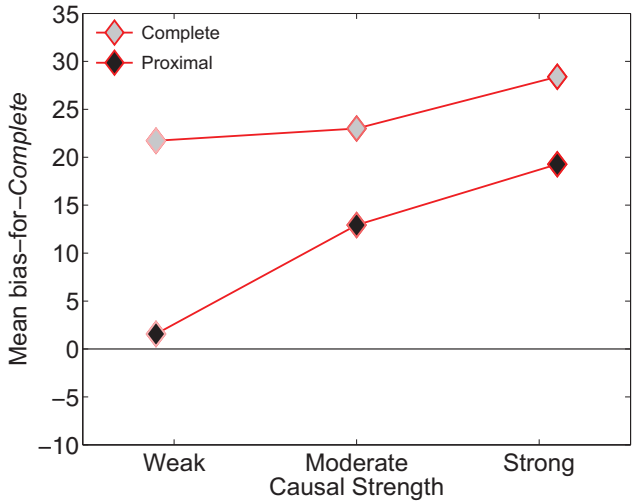

Figure 9. Graph of average bias-for-complete values by causal strength, split by explanation choice.

General Discussion

We began by considering four questions about simplicity in explanations and its role in human cognition:

Q_1: What makes an explanation simple?

Q_2: How are explanations selected when the simplest explanation is not the one best supported by the data?

Q_3: What are the cognitive consequences of a preference for simpler explanations? For example, does the preference bias memory or inference?

Q_4: Why do people prefer simpler explanations?

Our findings from Experiments 1 and 2 suggest an answer to Q_1: people's explanatory preferences correspond to root simplicity (i.e., minimizing the number of *unexplained* causes invoked in an explanation), and not to node simplicity (i.e., minimizing the number of total causes invoked in an explanation). Our findings from Experiment 2 additionally provide a partial answer to Q_2: when participants had access to a sample from the full joint probability distribution over diseases, explanatory preferences were a function of both root simplicity and probability.

Experiments 2 and 3 jointly address Q_3, with findings that suggest an influence of explanation on memory for previous observations. Specifically, participants who chose the simpler explanation when the data did not support this choice systematically misreported their observations: they misestimated the rates at

which disease combinations occurred in a way that made their explanation choice more likely than it truly was. But they only did this when their chosen explanation was the root-simpler option and was not already supported by the data. Experiment 3 went beyond Experiment 2 and demonstrated that choosing a root-simpler explanation (when it was not independently supported by the data) was a causal factor in subsequent memory distortions.

Finally, Experiment 4 explored Q_4, and found that people's explanatory preferences are more responsive to root simplicity when the root causes are strong. We suggested that people's preference for root-simpler explanations derives from the role explanation plays in generating efficient representations of exportable causal information for prediction and control. The stronger a root cause, the more usefully it fulfills this role. Strong root causes can be used to infer their downstream effects with greater certainty, and strong root causes allow larger or more certain effects from a single intervention. Additionally, we found that people's estimation biases were modulated by the strength of the causal relationship, consistent with the idea that these errors are driven by a preference for root simplicity.

Together, our findings present a unified (if complicated) picture of simplicity and its role in human judgment. We find that root simplicity informs explanatory judgments, is systematically combined with probabilistic information, can alter memory for previous observations, and is especially influential in cases involving strong causal relationships. It is interesting to note, however, this consistent role for root simplicity in judgments was not reflected in explicit justifications: participants very rarely invoked simplicity or complexity by name, and the small number of such appeals were not restricted to justifications for explanations that were simpler in terms of root simplicity. This suggests that even though root simplicity influences people's judgments, it may not be what people mean when they explicitly justify an explanation with reference to simplicity.

Relationship to Prior Work

While our findings provide initial answers to Q_{1-4}, they also raise important questions, including their relationship to prior work. For example, we find that simplicity and frequency information jointly influence explanation choices, but how are these two factors combined? Lombrozo (2007) argued that simplicity plays a role in determining the prior probability of a hypothesis, but that frequency information influences how the probability assigned to a hypothesis is updated, with a final decision resulting from probability matching to the resulting posterior distribution. The data from Lombrozo (2007) suggested a prior for simpler explanations ranging from 68% (2007: Experiment 3) to 79% (2007: Experiment 2), and here we find similar results, with priors that range from 68.9% (Experiment 2) to 69.4% (Experiment 3). However, in some cases, we found that participants *underweighted* probability in their final decision (assuming that they were probability matching), potentially because of the format in which the probabilistic information was presented—in Lombrozo (2007), it was also the case that frequency information was weighted less heavily when presented in a series of individual cases as opposed to numerical summary values. It could be that data presented sequentially is treated as involving greater uncertainty than numerical summary values.

A second question concerns the way in which explanation affects other judgments, such as probability or frequency estimation. Previous work, reviewed in Koehler (1991), has found that prompting people to explain why something could be the case (e.g., a particular team winning a sports tournament) increases the subjective probability that it is (or will be) the case. Our findings from Experiments 2–4 differ in a number of ways. Most notably, we found the largest explanation-induced changes in frequency estimation when the explanation selected was root-simpler *and* when the data themselves did not favor that explanation. In our experiments, explaining itself was not sufficient to strongly alter estimates.

One explanation for the selectivity of our effect is that estimation biases occur as participants try to reconcile two discrepant sources of evidence for the state of the world: their memory for different kinds of observations and their explanatory commitment. Because these are only likely to conflict when an explanatory preference—such as simplicity—draws people to commit to explanations that mismatch their observations, estimation biases are most likely to arise for participants who choose simple explanations when they're unlikely to be true. Future work could investigate these ideas more directly, with an eye toward isolating the effects of explanation in general from those that arise from specific explanatory preferences.

Our findings are potentially surprising in light of prior research on the causal status effect (e.g., Ahn, 1998; Ahn & Kim, 2000; Ahn, Lassaline, & Dennis, 2000; Murphy & Allopenna, 1994), which finds that people tend to favor causes that are earlier in a causal chain when making decisions about category membership. This might suggest that participants would favor explanations that appeal to "deeper" causes, leading them to favor the complete explanation in our chain-structure—that is, an explanation of the form A→B→C as opposed to B→C. In contrast, we found no preference for A→B→C over

B→C, consistent with the predictions of root simplicity. This apparent mismatch could arise for several reasons. First, the causal status effect is observed when the presence or absence of earlier causes is stipulated, and participants must select a classification. In our task, participants had to infer the presence or absence of an earlier cause in response to observed effects—it's not obvious why the same priority for deeper causes should arise. Second, the causal status effect is itself quite complex, and only arises under particular conditions (e.g., Lombrozo, 2009; Rehder, 2010). In some cases, a *coherence* effect is observed, whereby combinations of features that are consistent with one's beliefs about causal relationships in a given domain provide stronger evidence for membership in a category governed by those causal relationships. Rehder and Kim (2010) found that the relative role of causal status versus coherence was moderated by the strength of causal relationships, with stronger causes leading to *weaker* effects of causal status, in contrast to the stronger effects of root simplicity observed for stronger causes in Experiment 4. It could thus be that a preference for A→B→C over B→C doesn't arise in our task because the nature of the inference differs from that used in causal status classification tasks, or because the conditions under which root simplicity is favored (namely those involving strong root causes) are not those in which a causal status effect is likely to arise.

Limitations and Future Directions

Population and materials. An important limitation to our work stems from the large proportion of participants excluded from analyses, primarily for failing reading comprehension checks. Including such checks is a common practice in research involving data from large online populations (Crump, McDonnell, & Gureckis, 2013; Oppenheimer, Meyvis, & Davidenko, 2009), with difficult questions sometimes eliminating nearly 40% of participants (Downs, Holbrook, Sheng, & Cranor, 2010). These studies focus on individual exclusion criteria, not sets of criteria used simultaneously, and thus it is hard to compare this past work to our overall exclusion rate. However, none of our individual criteria eliminated anywhere near 40% of participants; the greatest percentage of participants eliminated by a single criterion was 26.5%, with most criteria excluding many fewer (see the online supplementary materials, Part B). It is important to note that, our criteria were all established prior to data collection and analyses, so they did not contribute to "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011).

Participant exclusions limit the generalizability of our findings to some extent. For instance, it could be that our results only apply to individuals with particular characteristics (such as high working memory), or that they only apply to individuals when they are engaged in deliberative reasoning. In light of these concerns, we repeated all of our analyses with a relaxed set of exclusion criteria, retaining only those criteria that required reading and following instructions related to the task, and where the response was solicited on the same page as the instructions or other information, thereby minimizing memory demands (see the online supplementary materials, Part D). Doing so resulted in a reduction in exclusion rates of 8%–15% across experiments, and a corresponding increase in sample sizes of 15%–36%. These analyses revealed very similar patterns of significance to those reported here. Most crucially: In Experiment 1, participants selected the complete-choice significantly more often than the proximal-choice in the diamond-structure condition; In Experiment 2, the regression analysis revealed that participants selected the complete-choice in the diamond-structure condition significantly more often than expected on the basis of the frequency information alone; In Experiment 3, regression analyses revealed a significantly greater estimation bias (favoring the complete-choice over the proximal-choice) for participants who explained prior to estimating in conditions with frequency information that did not favor the complete-choice; In Experiment 4, participants were significantly more likely to select the complete-choice as the causal strength of the root cause increased. All additional findings using these relaxed exclusion criteria are reported in the online supplementary materials, Part D.

While these additional analyses attenuate concerns about the characteristics of our sample, it remains a possibility that a variety of state or trait variables moderate the observed effects. For instance, there is evidence that people with lower levels of education are more likely to endorse the idea that complex problems can have simple solutions (e.g., van Prooijen, 2017), as well as evidence that individuals vary in their "attributional complexity," which itself correlates with need for cognition (Fletcher, Danilovics, Fernandez, Peterson, & Reeder, 1986). As in these cases, we expect that our findings succeed in identifying a dimension that governs intuitive judgments of complexity (namely the root simplicity of causal explanations), but that individual or cultural factors could influence the relative importance of this dimension, both in absolute terms and in relation to alternative bases for evaluating competing explanations.

Individuating causes and explanations. Thus far, our discussions and analyses have evaluated simplicity with respect to causes that were already individuated, and without assessing the "complexity" of the individual causes themselves. However, both of these assumptions deserve critical scrutiny: simplicity may interact with the individuation of causes, and some causes may be more "complex" than others.

These issues potentially arise in Experiment 4, where we suggested that invoking a small number of "strong" root causes allows for more efficient prediction and intervention via more efficient representations of causal systems. If root causes are deterministic causes of their children (which was not the case in any of our studies), then an observation of the root cause is formally equivalent to observing the cause's children. Faced with that situation, people may reindividuate the causes, representing the deterministic root and its children as a single entity—perhaps as a single cause with more complex internal structure. We expect further study about the role of variable individuation, internal complexity, and its relation to preferences for simpler explanations to prove fruitful.

A related concern is how to individuate explanations themselves—for example, how to determine "how far back" in the causal chain to go, and when doing generates a "new" explanation. Prior work by Thagard (1989) and Read and Marcus-Newhall (1993) suggests that explanations that are themselves explained should be favored, and Read and Marcus-Newhall (1993, Study 2) find this to be the case (but see Preston & Epley, 2005). In our analysis, if a cause explains another cause, by definition it becomes part of the total explanation under evaluation. As a result, although we contrast explained and unexplained *causes*, all of our *explanations* are themselves "unexplained." However, there is evidence that people do not always favor or generate explanations that include every cause in a causal system (Pacer et al., 2013), and doing so presumably becomes increasingly cumbersome the more complex the causal is. On the other hand, choosing a subset of causes introduces a computational explosion in the number of potential explanations (Shimony, 1991). Identifying the "boundaries" of explanations and treating some explanations as explanations of other explanations may be an excellent approach to addressing this problem. Further work looking at how root simplicity interacts with this process of "chunking" a causal system into individually coherent explanations could prove valuable (see also Johnson & Ahn, 2015).

Formal metrics of simplicity. Simplicity has received many formal treatments over the years, and a full story about explanation will assuredly have at least some formal elements. How do our findings relate to these formal approaches, and might our method be adapted to testing formal metrics more directly?

Several metrics consider the number of parameters included in a model and assign models with fewer parameters a higher probability (e.g., Jeffreys, 1998; Jeffreys & Berger, 1992; Popper, 1959; Akaike, 1974; or see Baker, 2010). Our findings are difficult to reconcile with these accounts without modification. First, such models assume that simplicity is valuable only instrumentally, as

a cue to probability, whereas our results are consistent with a stronger role for simplicity, as participants continued to favor simpler explanations even when evidence unambiguously favored an alternative. Second, such metrics have typically been concerned only with the *number* of parameters required, not the *values* of those parameters, which Experiment 4 suggests can also moderate preferences for simplicity. Although recent modifications to such metrics have considered the values of parameters, these accounts penalize for large coefficients, that is, stronger relationships (Fan & Li, 2001), which is the opposite of what we found in Experiment 4, in which stronger causal relationships between the variables resulted in including a higher count of variables.

On penalizing large parameters, strong causes, and sparsity. Some techniques within machine learning include explicit or implicit commitments to simplicity, expressed in terms of the structure of parameters, their strength, or sometimes both. For instance, many traditional techniques that involve fitting matrices of parameters to data, such as Principle Components Analysis (PCA) or Latent Semantic Analysis (LSA), implicitly rely on a penalty that emerges from the structure or the rank of the factored predictive matrix. These techniques penalize matrices with larger absolute numbers of terms included in the final predictive matrix (Srebro & Shraibman, 2005). Modern machine learning methods such as deep learning often employ large sets of parameters without issue, but run into issues when faced with large parameter values, which can result in unstable algorithmic behavior. For cases like these, penalizing the strength of the parameters can be a computational necessity. Regularization techniques apply penalties to model scores based on both (or either) the number and the size of parameters. By doing so, modelers can reduce generalization error by considering both structure and strength (in the vein of our discussion in Experiment 4).[10]

Simplicity also arises in machine learning discussions of sparse coding and sparse representations. Sparse representations are weight settings for which (given a particular input) only a small number of units are expected to have nonzero activations or contributions to the explanation of some learning example. Or, more strongly, "In a good sparse model, there should only be a few highly activated units for any data case" (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Small numbers of large activations are arguably consistent with our findings regarding root simplicity. However, there is some controversy over whether people's priors over causal structures in fact favor a small number of strong causes. On the one hand, Lu et al. (2008) and Powell, Merrick, Lu, and Holyoak (2016) argue that people's priors are in fact sparse (giving few causes weight at all) and strong (giving those causes that have weights large weights). On the other hand, empirical estimates of prior distributions support a prior that favors strong causes, but not sparsity among causes (Yeung & Griffiths, 2015). Importantly, though, the cases examined in Yeung and Griffiths (2015) involved relatively few causal variables (two: a potential cause and a background cause). When people confront a large number of candidate causes, let alone the number of parameters that can be present in deep learning models (often $k > 10^6$), resource limitations could impose a constraint favoring node-style simplicity (or some other way to reduce the number of variables). This suggests that priors favoring sparse representations may manifest only in cases where many variables are at play. Indeed, when more variables exist as potential competing causes, as in Powell et al. (2016), models with priors giving more weight to sparse parameter assignments perform better at modeling human causal strength judgments.[11] Further work is needed to explore the relationship between the preferences *against* strong links prevalent in machine learning and preferences *for* strong links found in causal induction and explanation.

Kolmogorov complexity and algorithmic information theory. Another approach that is closer in spirit to root simplicity is that exemplified by Kolmogorov Complexity (Kolmogorov, 1965) in the field of algorithmic information theory (Solomonoff, 1960), according to which simplicity corresponds to the length of code required to encode a program that generates an object in a universal descriptor language (where the canonical example of such a programming language would be that of a Universal Turing Machine). Or (to put it far too simply) the easier it is to compress, the simpler it is. This approach has been advocated most prominently in psychology by Chater (together with computer scientist Vitányi), who has suggested that this notion of simplicity offers a unifying principle for understanding all of cognition (Chater, 1999; Chater & Vitányi, 2003).

While there are clear connections between formal notions of compression (such as Kolmogorov Complexity) and our suggestions in Experiment 4, our own proposal was concerned with efficiently representing a particular kind of information: that which would best support prediction and intervention, and perhaps communication in causal settings. However, Kolmogorov Complexity is an information-theoretic account devoid of causal or interventional information. If we are correct in suggesting that causal information of this sort is relevant for explanation (see also, Pacer et al., 2013), then alternatives to Kolmogorov Complexity that represent causal information (such as *causal information flow*, see Ay & Polani, 2008) may need to be developed to fully describe these relationships. Exploring the connections between Kolmogorov Complexity and this causally defined notion of information is a promising direction for future work.

Beyond simplicity: Other explanatory virtues. Many other explanatory virtues can (and should) be explored to develop a full picture of human explanatory judgments. These include concerns about unification and explanatory scope (Khemlani, Sussman, & Oppenheimer, 2011; Kitcher, 1989), explanatory power (Schupbach, 2011; Schupbach & Sprenger, 2011), subsumption (Williams & Lombrozo, 2010, 2013; Williams, Lombrozo, & Rehder, 2013), interactions between different "levels" of explanation and general concerns about granularity (Anderson, 1990; Marr, 1982; Rottman & Keil, 2011), and several others. Many of these explanatory features have not been analyzed in the context of a computational theory of explanation, but we suggest that the paradigms

[10] Another modern technique used in deep learning is called "dropout" and involves excluding a randomly selected set of parameters from the model on different training examples. This technique does not concern itself with the simplicity of the final model per se, but rather the "class" of models that are effectively being trained via this sampling procedure. In practice, dropout does seem to result in models that are "simpler" in both strength and structure: it acts as a strength-based regularizer (Wager, Wang, & Liang, 2013; Erhan et al., 2010) and also encourages sparse representations (Srivastava et al., 2014).

[11] In Powell et al. (2016) priors with only a preference for strong causes failed to account for the competition between candidate causes.

developed here will adapt well to broader exploration, including to cases of noncausal explanation.

Is root simplicity a kind of simplicity? Recognizing the multiplicity of explanatory virtues raises a question about whether the virtue we have identified is best characterized as *simplicity*, especially in light of the small proportion of participants who explicitly identified simplicity as a basis for their choice of the root-simpler explanation. Could it be that root simplicity in fact reflects some other virtue, or a combination of other virtues? To some extent this question is empirical, but to a large extent it is a question of nomenclature for the scientific community to resolve. In our view, the similarities between root simplicity and other notions of simplicity that have arisen in philosophy, science, and computer science are more striking than the divergences. These other notions differ not only from the everyday usage of the term "simplicity," but also from each other. In naming the virtue that we identify root *simplicity*, we contribute to the conversation around an evolving term of art within psychological theory; a term which may ultimately diverge from folk usage. Such divergence is not unusual: psychology is a technical science, and this is reflected in divergences between how terms are used in psychological theorizing and in everyday speech, such as "prototypicality," "depression," or "memory."

Balancing conflicting virtues. Recognizing the plurality of explanatory virtues also raises questions about how decisions are made when different virtues conflict. For instance, the simplest explanation may not be the broadest or most fruitful. It should be possible to construct cases in which people favor a root complex explanation over a root simple one. Our account predicts that this should occur when other explanatory virtues favor the root complex explanation, thereby "outweighing" the influence of root simplicity. However, it is unclear how people combine the influences of diverse, potentially conflicting virtues into singular judgments.

One proposal, developed by Thagard (1989, 2004), is that the best explanation is chosen after a process of constraint satisfaction involving multiple virtues. This is a valuable approach that addresses the challenge of balancing virtues through its implementation in a neural-network-like architecture for described dependencies between propositions. Thagard's theory is explicitly posed in nonprobabilistic terms. Thagard writes that his account of coherence "contrasts markedly with probabilistic accounts" (Thagard, 1989), and emphasizes the categorical nature of judgments: propositions are either accepted or not accepted, not assigned an intermediate value that can be interpreted as a subjective probability (Thagard, 2004).

In contrast, our own approach is grounded in probabilistic theories of causal learning and explanation. We show that people are able to learn about causal relationships from frequency data, and that frequency data affects their explanations. Moreover, we show (in Experiment 4) that these probabilistic relations can alter their reliance on root simplicity. A constraint satisfaction approach to resolving conflicting virtues could potentially accommodate these results, but to do so it would need to be expanded to reason about probabilistic relations, and to define explanatory virtues in a commensurate representational framework. Developing such a framework is a valuable goal for future research, but we expect that doing so will be more viable once additional virtues have been empirically vetted (through the kind of experimentation we report here for the virtue of simplicity). With identified virtues in hand, it will be valuable to revisit the question of how they trade off in complex, real-world cases.

Conclusion

By using methods drawn from philosophical, psychological, and statistical toolboxes, we suggest that

1. *root* simplicity is a better predictor of human behavior than *node* simplicity;
2. simplicity trades-off with probability in choosing explanations;
3. choosing simple explanations can alter memory;
4. and the value of *root* simplicity increases with causal strength.

We unify our findings with a theory of explanation as a process with very specific aims: to inform information-rich representations of causal systems, exportable to other situations in which these representations improve prediction and intervention.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition, 69,* 135–178. http://dx.doi.org/10.1016/S0010-0277(98)00063-8

Ahn, W., & Kim, N. S. (2000). The causal status effect in categorization: An overview. In D. L. Medin (Ed.), *Psychology of learning and motivation* (Vol. 40, pp. 23–65). New York, NY: Academic Press.

Ahn, W., & Kim, N. S. (2008). Causal theories of mental disorder concepts. *Psychological Science Agenda, 22,* 3–8.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41,* 361–416. http://dx.doi.org/10.1006/cogp.2000.0741

Ahn, W. K., Novick, L. R., & Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review, 10,* 746–752. http://dx.doi.org/10.3758/BF03196541

Ahn, W. K., Proctor, C. C., & Flanagan, E. H. (2009). Mental health clinicians' beliefs about the biological, psychological, and environmental bases of mental disorders. *Cognitive Science, 33,* 147–182. http://dx.doi.org/10.1111/j.1551-6709.2009.01008.x

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80,* 123–129. http://dx.doi.org/10.1016/S0165-1765(03)00032-6

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, 19,* 716–723.

Anderson, J. R. (Ed.). (1990). *The adaptive character of thought*. New York, NY: Psychology Press.

Ay, N., & Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems, 11,* 17–41. http://dx.doi.org/10.1142/S0219525908001465

Baker, A. (2010). Simplicity. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Palo Alto, CA: Stanford University. Retrieved from http://plato.stanford.edu/archives/fall2013/entries/simplicity/

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology, 48,* 1156–1164. http://dx.doi.org/10.1037/a0026471

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103,* 566–581. http://dx.doi.org/10.1037/0033-295X.103.3.566

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology, 52A,* 273–302. http://dx.doi.org/10.1080/713755819

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7,* 19–22. http://dx.doi.org/10.1016/S1364-6613(02)00005-0

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. http://dx.doi.org/10.1037/0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58,* 545–567. http://dx.doi.org/10.1037/0022-3514.58.4.545

Clark, R. (2001). Information theory, complexity, and linguistic descriptions. In S. Bertolo (Ed.), *Parametric linguistics and learnability* (pp. 126–171). New York, NY: Cambridge University Press.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE, 8,* e57410. http://dx.doi.org/10.1371/journal.pone.0057410

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010, April). Are your participants gaming the system? Screening Mechanical Turk workers. In W. Mackay (Chair), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2399–2402). New York, NY: Association for Computing Machinery.

Dweck, C. S. (2008). Can personality be changed? The role of beliefs in personality and change. *Current Directions in Psychological Science, 17,* 391–394. http://dx.doi.org/10.1111/j.1467-8721.2008.00612.x

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines, 21,* 389–410. http://dx.doi.org/10.1007/s11023-011-9241-3

Erhan, D., Bengio, Y., Courville, A., Manzagol, P. A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research, 11,* 625–660.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96,* 1348–1360. http://dx.doi.org/10.1198/016214501753382273

Feldman, J. (2000, October 5). Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630–633. http://dx.doi.org/10.1038/35036586

Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review, 116,* 875–887. http://dx.doi.org/10.1037/a0017144

Fletcher, G. J., Danilovics, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual differences measure. *Journal of Personality and Social Psychology, 51,* 875–884. http://dx.doi.org/10.1037/0022-3514.51.4.875

Forster, M., & Sober, E. (1994). "How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions". *British Journal for the Philosophy of Science, 45,* 1–36.

Frances, A. J., & Egger, H. L. (1999). Whither psychiatric diagnosis. *The Australian and New Zealand Journal of Psychiatry, 33,* 161–165. http://dx.doi.org/10.1046/j.1440-1614.1999.00534.x

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy, 71,* 5–19. http://dx.doi.org/10.2307/2024924

Gopnik, A. (2000). Explanation as orgasm and the drive for causal understanding: The evolution, function and phenomenology of the theory-formation system. In F. Keil & R. Wilson (Eds.), *Cognition and explanation* (pp. 299–323). Cambridge, Mass: MIT Press.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511814563

Jeffreys, H. (1998). *The theory of probability*. Oxford, United Kingdom: Oxford University Press.

Jeffreys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80,* 64–72.

Johnson, S. G., & Ahn, W. K. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science, 39,* 1468–1503. http://dx.doi.org/10.1111/cogs.12213

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57,* 227–254.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28,* 107–128. http://dx.doi.org/10.1037/h0034225

Kelly, K. T. (2007). How simplicity helps you find the truth without pointing at it. *Induction, Algorithmic Learning Theory, and Philosophy, 9,* 111–143.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition, 39,* 527–535. http://dx.doi.org/10.3758/s13421-010-0028-1

Kim, N. S., & Ahn, W. K. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General, 131,* 451–476. http://dx.doi.org/10.1037/0096-3445.131.4.451

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. *Scientific explanation, 13,* 410–505.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110,* 499–519. http://dx.doi.org/10.1037/0033-2909.110.3.499

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission, 1,* 1–7.

Lagnado, D. (1994). *The psychology of explanation: A Bayesian approach*. (Unpublished master's thesis). University of Birmingham, Birmingham, England.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10,* 464–470. http://dx.doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232–257. http://dx.doi.org/10.1016/j.cogpsych.2006.09.006

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition, 110,* 248–253. http://dx.doi.org/10.1016/j.cognition.2008.10.007

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.) *Oxford handbook of thinking and reasoning* (pp. 260–276). New York, NY: Oxford University Press.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences, 20,* 748–759. http://dx.doi.org/10.1016/j.tics.2016.08.001

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99,* 167–204. http://dx.doi.org/10.1016/j.cognition.2004.12.009

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115,* 955–984. http://dx.doi.org/10.1037/a0013256

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. *Theories in social psychology,* 72–95.

Marr, D. (1982). *Vision*. Cambridge: MIT Press.

Monterosso, J., Royzman, E. B., & Schwartz, B. (2005). Explaining away responsibility: Effects of scientific explanation on perceived culpability. *Ethics & Behavior, 15,* 139–158. http://dx.doi.org/10.1207/s15327019eb1502_4

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 904–919. http://dx.doi.org/10.1037/0278-7393.20.4.904

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009

Pacer, M., Williams, J., Xi, C., Lombrozo, T., & Griffiths, T. L. (2013). Evaluating computational models of explanation using human judgments. In A. Nicholson & P. Smyth (Chairs), *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. New York, NY: Association of Computing Machinery.

Popper, K. (1959). *The logic of scientific discovery*. New York, NY: Routledge.

Powell, D., Merrick, M. A., Lu, H., & Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive Psychology, 86,* 62–86. http://dx.doi.org/10.1016/j.cogpsych.2016.02.001

Preston, J., & Epley, N. (2005). Explanations versus applications: The explanatory power of valuable beliefs. *Psychological Science, 16,* 826–832. http://dx.doi.org/10.1111/j.1467-9280.2005.01621.x

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology, 65,* 429–447. http://dx.doi.org/10.1037/0022-3514.65.3.429

Rehder, B. (2010). Causal-based categorization: A review. *Psychology of Learning and Motivation, 52,* 39–116. http://dx.doi.org/10.1016/S0079-7421(10)52002-4

Rehder, B., & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1171–1206. http://dx.doi.org/10.1037/a0019765

Rottman, B. M., & Keil, F. C. (2011). What matters in scientific explanations: Effects of elaboration and content. *Cognition, 121,* 324–337. http://dx.doi.org/10.1016/j.cognition.2011.08.009

Salmon, W. C. (1989). *Four decades of scientific explanation*. Pittsburg, PA: University of Pittsburgh Press.

Schupbach, J. N. (2011). Comparing probabilistic measures of explanatory power. *Philosophy of Science, 78,* 813–829. http://dx.doi.org/10.1086/662278

Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power*. *Philosophy of Science, 78,* 105–127. http://dx.doi.org/10.1086/658111

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Sherman, S. J., Skov, R. B., Hervitz, E. F., & Stock, C. B. (1981). The effects of explaining hypothetical future events: From possibility to probability to actuality and beyond. *Journal of Experimental Social Psychology, 17,* 142–158. http://dx.doi.org/10.1016/0022-1031(81)90011-1

Shimony, S. E. (1991). Explanation, irrelevance and statistical independence. *In the Proceedings of the ninth National conference on Artificial intelligence*. 1, 482–487.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Sober, E. (2006). Parsimony. In S. Sarkar & J. Pfeifer (Eds.), *The philosophy of science: An encyclopedia* (Vol. 2). New York, NY: Routledge.

Solomonoff, R. (1960). *A preliminary report on a general theory of inductive inference*. Cambridge, MA: Zator. Retrieved from http://raysolomonoff.com/publications/rayfeb60.pdf

Srebro, N., & Shraibman, A. (2005). Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory,* 545–560.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15,* 1929–1958.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–641.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12,* 435–467. http://dx.doi.org/10.1017/S0140525X00057046

Thagard, P. (2004). Causal inference in legal decision making: Explanatory coherence vs. Bayesian networks. *Applied Artificial Intelligence, 18*(3–4), 231–249. http://dx.doi.org/10.1080/08839510490279861

Van Prooijen, J. W. (2017). Why education predicts decreased belief in conspiracy theories. *Applied Cognitive Psychology, 31,* 50–58.

Wager, S., Wang, S., & Liang, P. S. (2013, December). *Dropout training as adaptive regularization*. Paper presented at the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science, 34,* 776–806. http://dx.doi.org/10.1111/j.1551-6709.2010.01113.x

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology, 66,* 55–84. http://dx.doi.org/10.1016/j.cogpsych.2012.09.002

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General, 142,* 1006–1014. http://dx.doi.org/10.1037/a0030996

Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology, 76,* 1–29. http://dx.doi.org/10.1016/j.cogpsych.2014.11.001

Received November 19, 2015
Revision received March 27, 2017
Accepted March 30, 2017 ■