




Cognitive Science 46 (2022) e13169
© 2022 Cognitive Science Society LLC.
ISSN: 1551-6709 online
DOI: 10.1111/cogs.13169

Simplicity as a Cue to Probability: Multiple Roles for Simplicity in Evaluating Explanations

Thalia H. Vrantsidis,^a  Tania Lombrozo^b

^a*University Center for Human Values, Princeton University*

^b*Department of Psychology, Princeton University*

Received 15 March 2022; received in revised form 20 May 2022; accepted 26 May 2022

Abstract

People often face the challenge of evaluating competing explanations. One approach is to assess the explanations' relative probabilities—for example, applying Bayesian inference to compute their posterior probabilities. Another approach is to consider an explanation's qualities or “virtues,” such as its relative simplicity (i.e., the number of unexplained causes it invokes). The current work investigates how these two approaches are related. Study 1 found that simplicity is used to infer the inputs to Bayesian inference (explanations' priors and likelihoods). Studies 1 and 2 found that simplicity is also used as a direct cue to the outputs of Bayesian inference (the posterior probability of an explanation), such that simplicity affects estimates of posterior probability even after controlling for elicited (Study 1) or provided (Study 2) priors and likelihoods, with simplicity having a larger effect in Study 1, where posteriors are more uncertain and difficult to compute. Comparing Studies 1 and 2 also suggested that simplicity plays additional roles unrelated to approximating probabilities, as reflected in simplicity's effect on how “satisfying” (vs. probable) an explanation is, which remained largely unaffected by the difficulty of computing posteriors. Together, these results suggest that the virtue of simplicity is used in multiple ways to approximate probabilities (i.e., serving as a cue to priors, likelihoods, and posteriors) when these probabilities are otherwise uncertain or difficult to compute, but that the influence of simplicity also goes beyond these roles.

Keywords: Simplicity; Complexity; Explanation; Probability; Bayesian inference; Explanatory virtues

Preregistrations, raw data, study materials, and analysis scripts can be found at: https://osf.io/8wync/?view_only. We thank Dr. Sam Johnson for sharing stimuli and data, the Concepts and Cognition lab for valuable feedback on this work, and the Princeton University Center for Human Values and the Program in Cognitive Science at Princeton University for supporting Thalia Vrantsidis.

Correspondence should be sent to Thalia H. Vrantsidis, Department of Psychology, Princeton University, Peretsman Scully Hall, Princeton, NJ 08540, USA. E-mail: tvrantsidis@princeton.edu

1. Introduction

Evaluating explanations plays a key role in human cognition (Keil, 2006; Lombrozo, 2006, 2012, 2016). For instance, people might try to explain their successes or failures, other people's behaviors, or what made them healthy or sick. In most cases, multiple explanations are consistent with available evidence. How do people evaluate these explanations and decide which is best?

One approach is to estimate probabilities. As put by the fictional detective Sherlock Holmes, people could “balance the probabilities and choose the most likely” (Doyle, 1986, p. 30). For instance, suppose someone is comparing two explanations for a pair of symptoms: (1) that a single disease (D_1) caused both symptoms, or (2) that two diseases (D_2 and D_3), each caused one symptom (see Fig. 1A). Choosing the most likely explanation involves comparing the “posterior probabilities” of these explanations given the two symptoms. Unfortunately, such probabilities can be difficult to compute. According to Bayes' rule, posteriors can be computed from *priors* (here, the baserates of having D_1 , or D_2 and D_3) and likelihoods (here, the chance of having the two symptoms if one has D_1 , or D_2 and D_3). Yet, in many cases, people may be uncertain of these values. Furthermore, computing posteriors from these values might be a cognitively challenging and error-prone process (Kahneman, Slovic, Slovic, & Tversky, 1982).

Another approach to evaluating explanations can potentially bypass assessments of probabilities: considering explanations' “virtues,” such as their simplicity, goodness-of-fit, or consistency with existing beliefs (Glymour, 2015; Johnson, Valenti, & Keil, 2019; Lipton, 2004; Lombrozo, 2016; Mackonis, 2013; Thagard, 1978, 1989). For example, people might prefer the one-disease explanation because it is simpler, in that it posits fewer independent, unexplained causes (Pacer & Lombrozo, 2017). Empirical work supports this idea, showing that people do often prefer explanations that are simpler in this sense (Johnson et al., 2019; Lombrozo, 2007; Pacer & Lombrozo, 2017; Read & Marcus-Newhall, 1993; though see, e.g., Zemla, Sloman, Bechlivanidis, & Lagnado, 2017), as well as explanations that possess other explanatory virtues (e.g., Johnson, Johnston, Toig, & Keil, 2014; Khemlani, Sussman, & Oppenheimer, 2011; Read & Marcus-Newhall, 1993; Schupbach, 2011).

Given these two seemingly disparate approaches to evaluating explanations—computing probabilities or weighing virtues—a question that arises is how these might be related. One possibility is that explanatory virtues are valued precisely because they are used to approximate probabilities (however imperfectly) in the face of uncertainty and cognitive limitations (Dellsén, 2018; Henderson, 2014; Johnson et al., 2019; Lipton, 2004; Lombrozo, 2007; Wojtowicz & DeDeo, 2020). Focusing on the virtue of simplicity, the current work investigates whether this might be the case, and if so, how this occurs.

We consider three possibilities. First, simplicity could be used as a cue to the inputs of Bayesian inference, namely, priors and likelihoods, when these values are uncertain (Johnson et al., 2019; Lombrozo, 2007). For example, people might assign the simple, single-disease explanation a higher prior probability, reflecting the assumption that diseases are rare, so having one disease is generally more likely than having two. This idea has received indirect

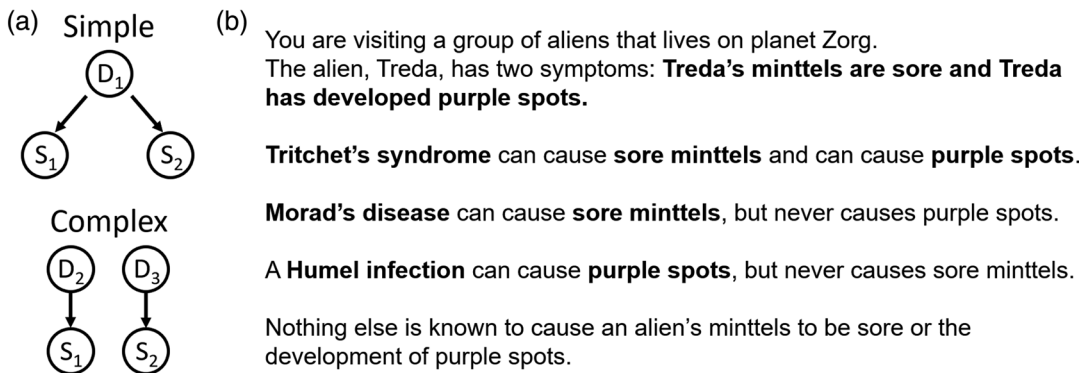


Fig. 1. Example explanations. *Note.* (a) Causal structures for simple and complex explanations. (b) Example scenario wording from Study 1. The “simple” explanation is simpler in that it requires fewer unexplained causes (a single disease vs. two diseases) to explain the same set of symptoms.

support in modeling explanation choices (Lombrozo, 2007) and direct support in tasks that ask participants to assign priors to simple and complex explanations (Johnson et al., 2019). There is also some evidence that people use simplicity to infer likelihoods, though in the opposite direction: assigning higher likelihoods to more complex explanations (Johnson et al., 2019). One goal of the current work is to test whether these types of effects replicate—that is, whether simplicity reliably affects estimates of an explanation’s prior and/or likelihood.

Simplicity might also be used to bypass other challenges faced in evaluating probabilities. Beyond providing cues to the inputs of Bayesian inference, simplicity could be used as a direct cue to the outputs of Bayesian inference (the posterior probability of an explanation), bypassing the need to estimate and then combine priors and likelihoods. A second goal of the current work is to test this previously unexamined possibility. If simplicity is used as a direct cue to posteriors, then we should find that simplicity preferences persist even when controlling for priors and likelihoods. Furthermore, these simplicity preferences might be stronger in cases where posteriors are more uncertain or difficult to compute, such that people have greater need to rely on alternative cues. Indeed, in previous work, effects of simplicity are not reliably found when explanations’ posterior probabilities are explicitly provided or trivial to compute (Bonawitz & Lombrozo, 2012; Lagnado, 1994; Lombrozo, 2007; Shimojo, Miwa, & Terai, 2020).

These two roles for simplicity (as a cue to priors and likelihoods, and as a direct cue to posteriors) correspond to two different ways in which simplicity could be used to approximate an explanation’s probability. However, simpler explanations might additionally or alternatively be preferred for reasons unrelated to probabilistic evaluation. For instance, simpler explanations could be processed more easily, and thus be perceived as more pleasing or plausible (e.g., Greifeneder & Bless, 2017; Scharrer, Bromme, Britt, & Stadler, 2012), or they could offer more cognitively efficient, “compressed” representations that aid in memory or communication (Pacer & Lombrozo, 2017; Wilkenfeld, 2019). Consistent with the

idea that explanation evaluation is not straightforwardly reducible to probabilistic evaluation, some work has found that assessments of explanation quality depart systematically from both perceived and objective probability, and can be affected by different factors (Douven & Schupbach, 2015; Lombrozo, 2007; Pacer, Williams, Chen, Lombrozo, & Griffiths, 2013). This suggests that simplicity might play different roles when it comes to assessing how “satisfying” an explanation is (see Liquin & Lombrozo, 2022) versus how probable it seems, with simpler explanations perhaps being deemed more satisfying, but not necessarily more probable (after accounting for simplicity’s use in inferring priors and likelihoods). Thus, a third goal of the current work is to test whether any simplicity preferences that persist after controlling for priors and likelihoods emerge only for judgments of explanatory satisfaction, or also for judgments of posterior probability.

To help disentangle these alternatives, we report two studies. In both studies, participants were presented with disease scenarios previously found to elicit preferences for simpler explanations (Johnson et al., 2019; Lombrozo, 2007; see Fig. 1B). Participants evaluated the simple and complex explanations by reporting either posterior probabilities or explanatory satisfaction. In Study 1, participants were also asked to estimate both priors and likelihoods; in Study 2, priors and likelihoods were specified in the scenario, with values yoked to the responses from participants in Study 1.

Study 1 addresses the first goal of this work: testing whether simplicity affects estimates of priors and/or likelihoods. Studies 1 and 2 both address the second goal: testing whether simplicity preferences persist when controlling for priors and likelihoods, whether elicited (Study 1) or provided (Study 2). Moreover, by comparing the effect sizes across studies, we can test whether simplicity preferences are stronger when posteriors are more difficult to compute (i.e., in Study 1, where computing posteriors involves the extra step of estimating priors and likelihoods). Finally, studies 1 and 2 both address the third goal of this work: testing whether effects of simplicity depend on whether people are evaluating explanatory satisfaction versus posterior probability.

All studies were fully preregistered, including the hypotheses, design, analysis plan, sample size, and exclusion criteria. Any departures from the preregistered analysis plan are noted. Preregistrations as well as access to raw data, study materials, and analysis scripts can be found at: https://osf.io/8wync/?view_only. All experiments were approved by the Princeton University Research Ethics Board.

2. Study 1

Study 1 examined: (1) whether simplicity is used as a cue to priors and likelihoods, and therefore affects estimates of these values, (2) whether simplicity is used as a *direct* cue in explanation evaluations, and therefore affects these evaluations after controlling for priors and likelihoods, and (3) whether direct effects of simplicity on explanation evaluation (if found) are restricted to judgments of explanatory satisfaction, or found for judgments of posterior probability as well.

2.1. Methods

2.1.1. Participants

Participants were 233 adults recruited from the United States through Prolific (age: $M = 28$, $SD = 9$; gender: 190 women, 34 men, 9 additional/multiple responses). Participants were excluded from the task if they failed to correctly answer all of the scenario comprehension questions by their second attempt, or if they failed to pass additional attention check questions.

2.1.2. Materials and procedures

Participants completed three trials, which each involved reading and answering questions about a scenario. (We illustrate throughout with one example.) Each trial began by presenting a disease scenario like our opening example (see Fig. 1B), in which a pair of symptoms could be explained by either a simple explanation (that Treda has Tritchet's syndrome) or a complex explanation (that Treda has both Morad's disease and a Humel infection). The simple explanation thus uses fewer unexplained causes (diseases) to explain the same set of symptoms.

After reading the scenario (which remained visible throughout the trial), participants completed three comprehension questions to ensure they knew which diseases caused which symptoms. If participants made any errors, they were given a second chance to answer these questions.

Participants then evaluated each explanation. Participants assigned to rate posterior probabilities were asked: "Please estimate the probability of Treda having each disease or combination of diseases: [Tritchet's syndrome/Morad's disease and a Humel infection]." Participants randomly assigned to rate satisfaction were asked: "How satisfying is each explanation for Treda's symptoms? [Treda has Tritchet's syndrome/Treda has Morad's disease and a Humel infection]." Ratings were made on 100-point scales, ranging from 0% to 100% chance, or "Not all satisfying" to "Extremely satisfying," respectively.

Finally, participants estimated priors and likelihoods for each explanation. Priors were elicited by asking: "Suppose a random alien was selected from planet Zorg. The alien may or may not be sick, and may or may not have any symptoms; you have no knowledge either way. How likely do you think this random alien would be to have: [Tritchet's syndrome/Morad's disease and a Humel infection]." Likelihoods were elicited by asking: "Imagine an alien who has [Tritchet's syndrome/both Morad's disease and a Humel infection]. How likely is it that this alien would have both feverish muffs and wrinkled ears?" For exploratory purposes, participants also rated the chance that a random alien from the planet would have this pair of symptoms (i.e., the probability of the evidence, in Bayesian terms). All ratings were made on 100-point scales ranging from 0% to 100% chance.

The three scenarios used varied only in the names introduced for the diseases, symptoms, aliens, and planets. Because the scenarios were otherwise matched, variation in participants' priors and likelihoods should only reflect noise in their judgments. Accounting for this noise using structural equation modeling ensured that any simplicity effects that persist when controlling for priors and likelihoods are not merely caused by noisy measurements of these values (Buttrick, Axt, Ebersole, & Huband, 2020).

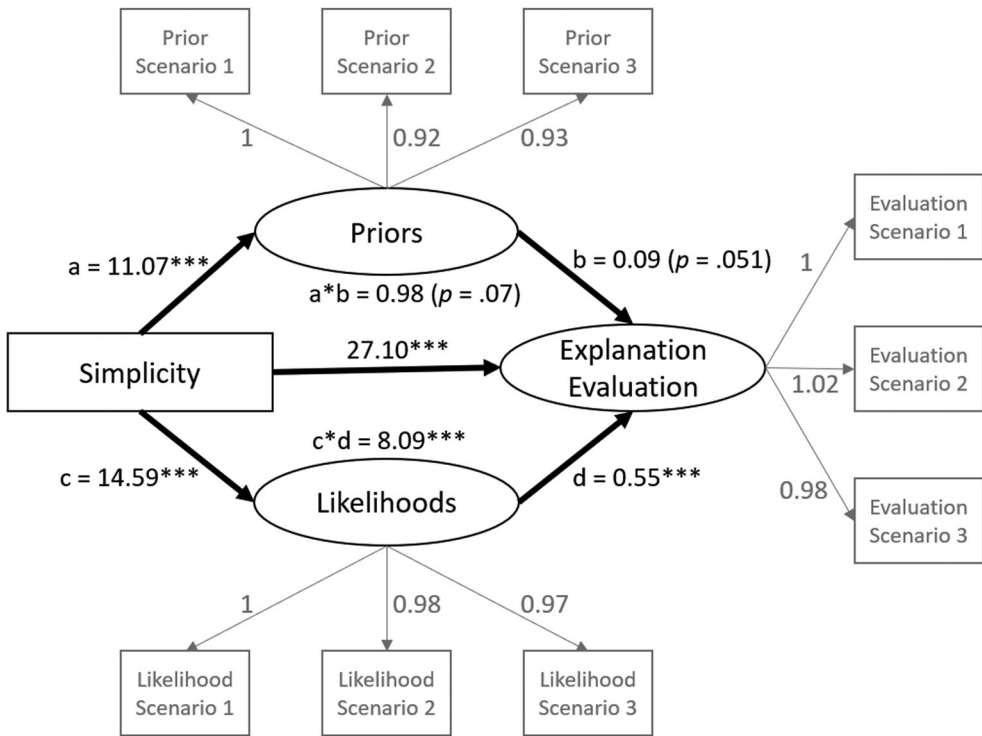


Fig. 2. Study 1 structural equation model results. *Note.* Unstandardized coefficients shown. *** indicates $p < .001$.

Across participants, we counterbalanced the type of explanation evaluation judgment (posteriors and satisfaction), the order in which priors and likelihoods were elicited, and whether the simple or complex explanation was described and rated first. The three scenarios were presented in a random order for each participant.

2.2. Results

All analyses were performed using R (v. 4.1.2; R Core Team, 2021) with the following packages: lmerTest (v. 3.1-3; Kuznetsova, Brockhoff, & Christensen, 2017), lavaan (v. 0.6-9; Rosseel, 2012), and metafor (v. 3.0-2; Viechtbauer, 2010). Regressions were fit as multilevel models with random intercepts for participants. For analyses involving simplicity or evaluation type, these variables were coded as follows: simplicity (0.5 = simple, -0.5 = complex); evaluation type (0.5 = posterior, -0.5 = satisfaction).

Preliminary analyses confirmed that participants showed simplicity preferences. Specifically, explanation evaluations were predicted from explanation type (simple and complex) and its interaction with evaluation type (posteriors and satisfaction). Simpler explanations were preferred ($B = 36.14, p < .001$), and this held for both evaluation types (posteriors: $B = 35.22, p < .001$; satisfaction: $B = 37.06, p < .001$, interaction: $B = -1.84, p = .44$), thus

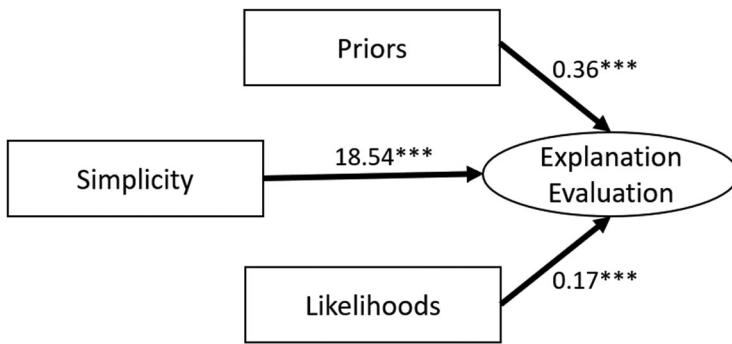


Fig. 3. Study 2 regression results. *Note.* Paths indicate unstandardized coefficients, controlling for all predictors shown as well as interactions with evaluation type. *** indicates $p < .001$.

replicating simplicity preferences found previously with similar scenarios (Johnson et al., 2019; Lombrozo, 2007).

To test whether simplicity affected the priors and likelihoods assigned to an explanation, we ran regressions predicting either priors or likelihoods from explanation type (simple and complex). In line with previous work (Johnson et al., 2019), participants assigned higher prior probabilities to simpler explanations ($M = 33.17$, $SD = 26.97$) than to complex ones ($M = 22.58$, $SD = 23.15$; $B = 10.59$, $p < .001$). Participants also assigned higher likelihoods to simple explanations ($M = 80.54$, $SD = 22.06$) than to complex ones ($M = 66.14$, $SD = 30.36$; $B = 14.40$, $p < .001$). This likelihood effect goes in the opposite direction of previous work (Johnson et al., 2019). Despite this reversal, these results support the idea that people use simplicity as a cue to both priors and likelihoods when these values are uncertain, and the direction of both effects could contribute to preferences for simpler explanations.

To test whether simplicity influences explanation evaluation beyond its role in estimating priors and likelihoods, we examined whether simplicity preferences persist after accounting for simplicity's influence on priors and likelihoods, using the structural equation model specified in Fig. 2. This model showed good fit ($SRMR = 0.04$, $CFI = 0.99$, $RMSEA = 0.04$). In the fitted model, the direct effect of simplicity on explanation evaluation was significant ($B = 27.01$, $p < .001$), indicating that simplicity preferences persisted after controlling for priors and likelihoods.¹ Furthermore, secondary analyses separately analyzing the two evaluation types showed that this simplicity preference occurred for both posteriors ($B = 27.63$, $p < .001$) and satisfaction ($B = 27.67$, $p < .001$), with a meta-analytic comparison indicating no significant difference between these effects ($B = -0.04$, $p = .99$); see Fig. 3. Thus, simplicity influenced explanation evaluations beyond its effects on priors and likelihoods, and this effect was not restricted to satisfaction judgments.

2.3. Discussion

Study 1 found that simplicity was used as a cue to the inputs of Bayesian inference: simpler explanations were assigned both higher priors and higher likelihoods. While the former effect is consistent with prior work (Johnson et al., 2019), the latter effect is not: Johnson et al.

(2019) found that simpler explanations were assigned *lower* likelihoods. This inconsistency is addressed in a supplementary study, which replicates both the current effect and Johnson et al.'s effect, and suggests that these differences are due to different independence assumptions implied by the scenarios used in each case (see Supplementary Materials and General discussion).

Study 1 went beyond previous research by showing that effects of simplicity on explanation evaluation persisted after controlling for priors and likelihoods, and that this occurred for both posterior and satisfaction judgments. One interpretation of this result is that simplicity is used as a direct cue to posteriors, potentially bypassing the need to estimate and combine priors and likelihoods, with satisfaction judgments then based on these estimated posteriors. Alternatively, simpler explanations could be found satisfying for reasons unrelated to perceived probability, with posterior probability judgments made either independently, or based on this sense of satisfaction. Either way, the results show that the effect of simplicity on explanation evaluations is not fully explained by simplicity's effect on priors and likelihoods.

3. Study 2

Study 2 had two aims. The first was to conceptually replicate two key findings from Study 1: that simplicity preferences persist after controlling for priors and likelihoods, and that this occurs for judgments of posteriors, not just satisfaction. In Study 2, this was tested in a case where participants had no need to estimate priors and likelihoods, because these values were explicitly provided. Continuing to find simplicity preferences here (after controlling for priors and likelihoods) would provide converging evidence that simplicity influences explanation evaluation beyond its role in estimating these values.

The second aim of Study 2 was to test whether a direct role for simplicity in explanation evaluations emerges at least partly in response to the challenge of computing posteriors from estimated priors and likelihoods. If people rely on simplicity to bypass the computational challenge of estimating and combining priors and likelihoods, then simplicity preferences (after controlling for priors and likelihoods) should be larger in Study 1 than in Study 2. This is because participants in Study 2 were provided with precise values for priors and likelihoods, thus reducing uncertainty associated with these values and removing a step in the process of computing posteriors. Finding larger simplicity preferences in Study 1 would, therefore, suggest that reliance on simplicity is moderated by the demands of probabilistic computation, with higher demands resulting in greater reliance on simplicity. Alternatively, failing to find this difference between studies might indicate that a direct use of simplicity in evaluating explanations is unrelated to the demands of probabilistic computation, and instead occurs for other reasons (e.g., perhaps because simpler explanations are more aesthetically pleasing, or easier to process, and judged more probable on this basis).

3.1. Methods

3.1.1. Participants

Participants were 233 adults recruited from the United States through Prolific (age: $M = 32$, $SD = 12$; gender: 170 women, 53 men, 3 additional/multiple responses). Exclusion criteria

were the same as in Study 1. Due to technical errors, an additional 29 participants completed the study. These participants were excluded prior to analysis to conform to the preregistered sample size.

3.1.2. *Materials and procedure*

The methods for Study 2 were the same as Study 1, with the following exception: participants were provided with numerical values for the priors and likelihoods of each explanation (as well as the probability of the evidence), rather than having to estimate these values themselves. The numerical values for these quantities were based on participants' responses from Study 1, with each participant in Study 2 matched to a unique participant in Study 1, so that the values provided in each scenario corresponded to the matched Study 1 participant's responses to that scenario. This allowed for a more closely matched comparison of simplicity effects across studies 1 and 2.²

In each scenario, values for the priors and likelihoods were provided after describing each disease or pair of diseases. In the example from Study 1, this information was provided as follows: "If an alien has [Tritchet's syndrome/both Morad's disease and a Humel infection], [likelihood]% of the time they will have both sore minttels and purple spots. About [prior]% of the aliens on Zorg have [Tritchet's syndrome/both Morad's disease and a Humel infection]." The probability of the evidence was provided by stating that "[probability of evidence]% of aliens on Zorg have both sore minttels and purple spots."

The counterbalanced factors (i.e., the type of explanation evaluation, and the order of the simple vs. complex explanations) were set to correspond to the matched participant from Study 1. The three scenarios were presented in a random order.

3.2. *Results*

A preliminary analysis as in Study 1 confirmed that participants again tended to show simplicity preferences in these scenarios ($B = 24.77, p < .001$), for both posteriors ($B = 21.40, p < .001$) and satisfaction ($B = 28.14, p < .001$; interaction: $B = -6.74, p = .003$). However, this preference could be driven either by simplicity itself, or by differences in the priors and likelihoods provided for simple versus complex explanations.

The primary analysis distinguished these possibilities by testing whether simplicity predicted explanation evaluations after controlling for priors and likelihoods (priors and likelihoods mean-centered, all predictors interacted with evaluation type). As shown in Fig. 3, participants still preferred simpler explanations ($B = 18.54, p < .001$) even when accounting for effects of priors and likelihoods (priors: $B = 0.36, p < .001$; likelihoods: $B = 0.17, p < .001$).³ Furthermore, this held for both types of explanation evaluations (posteriors: $B = 13.33, p < .001$; satisfaction: $B = 23.67, p < .001$), though effects on satisfaction were significantly larger ($B = -9.96, p < .001$); see Fig. 4. Like Study 1, this suggests that simplicity can provide direct cues to explanation quality, and that this holds for evaluations of posterior probability as well as satisfaction.

To test whether this simplicity preference was larger in Study 1 than in Study 2, meta-analytic tests compared the size of this simplicity effect to the size of the simplicity direct

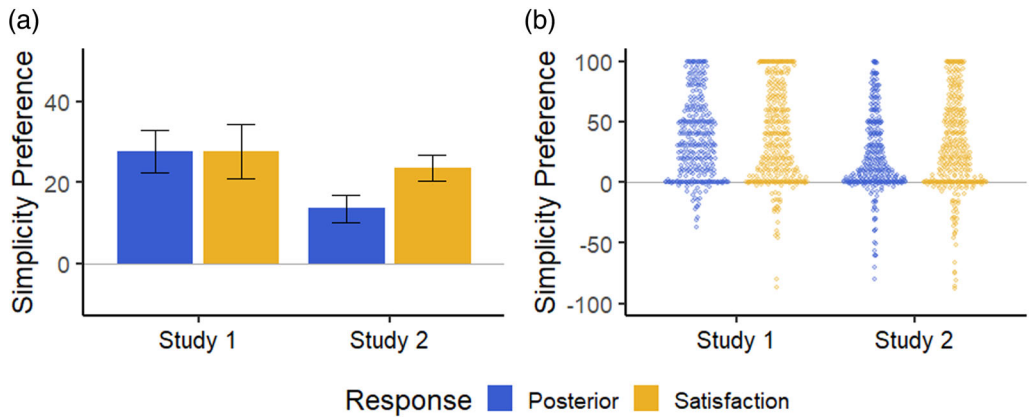


Fig. 4. Direct effects of simplicity on explanation evaluations in studies 1 and 2. *Note.* (a) Simplicity's effect on explanation evaluations after controlling for priors and likelihoods. Estimates indicate average difference in evaluation of simple versus complex explanations, with higher positive values indicating stronger simplicity preferences. 95% CIs shown. (b) Difference in evaluation of simple versus complex explanations for each trial, with higher positive values indicating stronger simplicity preferences, and lower negative values indicating stronger complexity preferences. Raw difference scores are shown, therefore data reflect simplicity preferences without controlling for priors and likelihoods.

effect in study 1.⁴ These effect sizes were predicted from study number (Study 1 = 0.5 and Study 2 = -0.5) and its interaction with evaluation type. Simplicity preferences were larger in Study 1 than in Study 2 ($B = 9.11, p < .001$), driven primarily by a reduced effect on posteriors ($B = 14.07, p < .001$), rather than satisfaction ($B = 4.16, p = .28$; interaction: $B = 9.91, p = .049$); see Fig. 4. The reduced effect on posteriors is consistent with the idea that people used simplicity as a direct cue to posteriors, and did so to a greater extent when it was harder to compute posteriors because priors and likelihoods were unspecified.

3.3. Discussion

Study 2 found that participants use simplicity as a direct cue in evaluating explanations, even when priors and likelihoods are provided. This offers additional evidence that the effect of simplicity on explanation evaluation is not reducible to its role in estimating priors and likelihoods. Instead, simplicity influences both judgments of explanatory satisfaction and estimates of posterior probability above and beyond these effects.

Study 2 also found a smaller effect of simplicity on estimates of posterior probability than that observed in Study 1. This reduced effect on posteriors suggests that participants relied more heavily on simplicity as a direct cue to posteriors when posteriors were more difficult to compute. In this case, posteriors were more difficult to compute in Study 1 (vs. 2) due to the uncertainty associated with estimated priors and likelihoods and the effort required to estimate these values. Notably, the findings were different when it came to explanatory satisfaction: for this measure, simplicity had comparable effects across studies 1 and 2. The fact

that computational demands moderated the influence of simplicity on judgments of posterior probability, but not judgments of explanatory satisfaction, suggests that there might be separate mechanisms operating in each case. Specifically, simplicity's additional role in satisfaction judgments might stem from valuing simplicity for reasons unrelated to probabilistic computation (e.g., finding simpler explanations more aesthetically pleasing or more efficient to process).⁵

The comparison between studies 1 and 2 additionally helps rule out the possibility that the direct effect of simplicity observed in Study 1 stemmed from systematic errors in the measurement of subjective priors and likelihoods, and thus a failure to effectively control for simplicity's effect through these values. Were this the case, we would expect the direct effect of simplicity on both posteriors and explanatory satisfaction to be substantially smaller in Study 2, given that priors and likelihoods were stipulated rather than measured. However, the fact that the direct effect of simplicity was comparable across studies for satisfaction judgments (and persisted to some extent even for posterior judgments) helps substantiate the measures of priors and likelihoods used in Study 1, and suggests that the direct effects observed in Study 1 were not primarily due to measurement errors.

4. General discussion

Across two studies, we identify three distinct roles for simplicity in the evaluation of explanations, where we define one explanation as simpler than another if it invokes fewer unexplained causes (Pacer & Lombrozo, 2017). First, an explanation's simplicity is used as a cue to that explanation's prior probability and to the likelihood with which it can produce the evidence, with simpler explanations assigned higher priors and likelihoods (Study 1). Second, an explanation's simplicity is used as a direct cue to its posterior probability as well as how satisfying it is, in that effects of simplicity on these judgments are not simply a consequence of effects of simplicity on priors and likelihoods (Studies 1 and 2). These results suggest that simplicity is used to estimate both the inputs (priors and likelihoods) and outputs (posteriors) of Bayesian inference when these values are uncertain, with a larger role for simplicity in estimating posteriors when this computation is more uncertain or complex (Study 1 vs. Study 2). Finally, the fact that simplicity's effect on satisfaction judgments remained relatively unchanged across studies suggests that simplicity plays additional roles unrelated to explicitly probabilistic considerations.

The fact that people use simplicity to approximate the inputs and outputs of Bayesian inference does not address the question of whether this role for simplicity in fact leads to more accurate estimates of probability. In practice, the reliability of simplicity as a cue to priors, likelihoods, and/or posteriors is likely to vary dramatically across contexts, as is the reliability and availability of alternative ways to infer these probabilities. Thus, if simplicity contributes to the accuracy of probability estimates, people's use of simplicity should vary with such contextual features. Some evidence suggests that this is the case. For example, previous research has found that simplicity has a greater effect on likelihoods when they are more difficult to estimate (because they are stochastic vs. deterministic), and that simplicity preferences are

stronger in physical versus social domains, plausibly tracking real differences in the structure of these domains (Johnson et al., 2019). The current work expands on this, finding that reliance on simplicity when estimating posteriors is moderated by the demands of probabilistic computation (in study 1 vs. study 2), with a larger role for simplicity under conditions of greater uncertainty and computational complexity—that is, when the process of computing posteriors from priors and likelihoods may be more inaccurate or unreliable. Speculatively, this use of simplicity could improve the accuracy of posterior estimates by bypassing or compensating for errors in these intermediate computations. Together, these findings suggest that people may use simplicity in a relatively sophisticated way that could help minimize inaccuracies under some conditions, rather than using it as a more inflexible and thus less-optimal heuristic strategy.

Two discrepancies between our findings and prior work point to additional ways in which the role of simplicity may be appropriately moderated by contextual factors. First, we found that simpler explanations were assigned higher likelihoods (study 1), whereas Johnson et al. (2019) found the reverse. In a supplementary study (see Supplementary Materials), we were able to replicate both patterns of results, and moreover found that differences stemmed in part from different assumptions across scenario wording concerning the conditional independence of effects (e.g., how likely symptoms are to co-occur when caused by one vs. two diseases). Notably, these different assumptions shifted explanation evaluations in directions consistent with the rules of probability, again suggesting that reliance on simplicity may reflect more sophisticated probabilistic inferences, and does not reflect the operation of a strictly inflexible heuristic.

More broadly, while our findings of simplicity preferences replicate previous studies using these types of scenarios (Johnson et al., 2019; Lombrozo, 2007), they are discrepant with findings that, under some conditions, *complex* explanations seem to be preferred (e.g., Lim & Oppenheimer, 2020; Liquin & Lombrozo, 2022; Zemla et al., 2017). At least some of these cases may reflect appropriate sensitivity to different probabilistic assumptions (e.g., stemming from different causal structures), in ways that may promote more accurate probabilistic inferences (see Liquin & Lombrozo, 2022; Zemla et al., 2017). However, it remains an open question whether all such cases can be explained in this way, or whether some complexity preferences might reflect the use of more inflexible heuristics (Johnson et al., 2014, 2019; Lim & Oppenheimer, 2020), or be consequences of other confounding factors.

Beyond a role in estimating probabilities, a preference for simpler explanations could have other beneficial (or detrimental) effects. Indeed, a preference for simpler explanations has already been shown to have both beneficial and detrimental consequences for learning (see, e.g., Lombrozo, 2016; Walker, Bonawitz, & Lombrozo, 2017; Wilkenfeld & Lombrozo, 2015; Williams, Lombrozo, & Rehder, 2013). The practice of favoring simpler explanations could also have other cognitive benefits, for example, resulting in less cognitively demanding processing, more compressed representations, and a greater sense of understanding (e.g., Pacer & Lombrozo, 2017; Scharrer et al., 2012; Wilkenfeld, 2019). It has also been argued that favoring more “virtuous” explanations (e.g., those that are simpler, or have greater explanatory power) might be beneficial for reasons related to prediction accuracy, but not because doing so approximates Bayesian inference (which minimizes long-term inaccuracy), and instead

because doing so deviates from Bayesian inference in ways that will often minimize shorter-term inaccuracy, and thus support faster convergence to the truth (Douven, 2022; Douven, 2021; Douven, 2020; see also Pacer et al., 2013). More fully characterizing how explanatory virtues can and should inform reasoning—probabilistic or otherwise—thus remains an important question for future research.

Future work in this area would also benefit from considering explanatory virtues beyond simplicity (e.g., Blanchard, 2018; Douven & Schupbach, 2015; Khemlani et al., 2011; Wojtowicz & DeDeo, 2020), as well as other notions of simplicity (e.g., Blanchard, Lombrozo, & Nichols, 2018). One benefit of doing so is that alternative metrics for simplicity, often focusing on data compression and algorithmic complexity, have been invoked in other cognitive domains, from perception to language (e.g., Chater & Vitányi, 2003; Chekaf, Cowan, & Mathy, 2016; Gauvrit, Zenil, Delahaye, & Soler-Toscano, 2014; Mathy & Feldman, 2012; Soler-Toscano, Zenil, Delahaye, & Gauvrit, 2014). Linking simplicity in explanation evaluation to these bodies of work could offer a richer and more unified picture of inductive inference and its role in human cognition.

In sum, the current work addressed the question of how people evaluate explanations, and, in particular, how assessing explanations' probabilities relates to the explanatory virtue of simplicity. We found that simplicity plays multiple roles in these evaluations: it is used as a cue to the inputs and outputs of Bayesian inference, which may help address the uncertainty and cognitive limitations associated with evaluating probabilities. Furthermore, simplicity's role may extend beyond evaluating probabilities, guiding people toward explanations that could be valuable for a variety of other reasons.

Notes

- 1 Secondary analyses which included the interaction of the priors and likelihoods in this model (using the product indicator approach with double mean-centering; Lin, Wen, Marsh, & Lin, 2010) did not produce a significant interaction ($B = -0.00, p = .14$), nor did this meaningfully alter the direct effect of simplicity ($B = 27.16, p < .001$).
- 2 Using participant responses from Study 1 for these values meant that participants in Study 2 occasionally saw scenarios with values that were inconsistent with the scenario description or otherwise extremely implausible. Therefore, additional exploratory analyses of Study 2 (as well as study 1, for comparison) were performed after excluding participants where any of the following were true in any of the three trials: the likelihood of either explanation producing the symptoms was rated as 0, the prior probability of either explanation was rated as 0 or 100, or the probability of the evidence was rated as 0. All results replicated both the direction and significance of the reported effects.
- 3 A secondary analysis that allowed priors and likelihoods to interact in this model did not produce a significant interaction ($B = 0.00, p = .37$), and did not meaningfully alter the effect of simplicity ($B = 18.57, p < .001$).
- 4 This was preregistered as a secondary analysis, but was used here as it is more clearly interpretable than the main preregistered test of this question. The results replicated with both analyses.

5 Further evidence of potential dissociations between judgments of explanatory satisfaction and posterior probabilities comes from secondary analyses comparing the influence of prior beliefs on these judgments. In both studies, prior beliefs had a larger effect on posteriors compared to satisfaction judgments, and, in study 1, priors did not significantly affect satisfaction judgments at all (thus accounting for the overall nonsignificant effect of priors on explanation evaluations shown in Fig. 2). Study 1: posterior: $B = 0.17$, $p < .001$; satisfaction: $B = -0.01$, $p = .88$; interaction: $B = 0.18$, $p = .049$; study 2: posterior: $B = 0.54$, $p < .001$; satisfaction: $B = 0.18$, $p < .001$; interaction: $B = 0.33$, $p < .001$. (Likelihood effects did not differ across evaluation type in either study.) This change in the use of priors further supports the idea that explanatory satisfaction may rely less heavily on the inputs to Bayesian inference, and may instead rely more heavily on other factors.

Open Research Badges



This article has earned Open Data, Open Materials badges, and pre-registered. Data are available at https://osf.io/8wync/?view_only, materials are available at https://osf.io/8wync/?view_only and pre-registered are available at https://osf.io/8wync/?view_only.

References

- Blanchard, T. (2018). Bayesianism and explanatory unification: A compatibilist account. *Philosophy of Science*, 85(4), 682–703.
- Blanchard, T., Lombrozo, T., & Nichols, S. (2018). Bayesian Occam's razor is a razor of the people. *Cognitive Science*, 42(4), 1345–1359.
- Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, 48(4), 1156.
- Buttrick, N., Axt, J., Ebersole, C. R., & Huband, J. (2020). Re-assessing the incremental predictive validity of implicit association tests. *Journal of Experimental Social Psychology*, 88, 103941.
- Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1), 19–22.
- Chekaf, M., Cowan, N., & Mathy, F. (2016). Chunk formation in immediate memory and how it relates to data compression. *Cognition*, 155, 96–107.
- Dellsén, F. (2018). The heuristic conception of inference to the best explanation. *Philosophical Studies*, 175(7), 1745–1766.
- Douven, I. (2020). The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science Part A*, 79, 1–14.
- Douven, I. (2021). How explanation guides belief change. *Trends in cognitive sciences*, 25(10), 829–830.
- Douven, I. (2022). *The art of abduction*. MIT Press.
- Douven, I., & Schupbach, J. N. (2015). Probabilistic alternatives to Bayesianism: The case of explanationism. *Frontiers in Psychology*, 6, 459.
- Doyle, A. C. (1986). *Sherlock Holmes: The complete novels and stories* (Vol. 2). Bantam Press.
- Gauvrit, N., Zenil, H., Delahaye, J.-P., & Soler-Toscano, F. (2014). Algorithmic complexity for short binary strings applied to psychology: A primer. *Behavior Research Methods*, 46(3), 732–744.

- Glymour, C. (2015). Probability and the explanatory virtues. *British Journal for the Philosophy of Science*, 66(3), 591–604.
- Greifeneder, R., & Bless, H. (2017). The interplay of cognition and feelings: Fluency. In R. Greifeneder, H. Bless, & K. Fiedler (Eds.), *Social cognition* (2nd ed., pp. 145–164). Psychology Press.
- Henderson, L. (2014). Bayesianism and inference to the best explanation. *British Journal for the Philosophy of Science*, 65(4), 687–715.
- Johnson, S., Johnston, A., Toig, A., & Keil, F. (2014). Explanatory scope informs causal strength inferences. *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Johnson, S., Valenti, J. J., & Keil, F. C. (2019). Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cognitive Psychology*, 113, 101222.
- Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
- Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57(1), 227–254.
- 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(3), 527–535.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(1), 1–26.
- Lagnado, D. (1994). *The psychology of explanation: A Bayesian approach*. Masters Thesis. Schools of Psychology and Computer Science, University of Birmingham.
- Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PLoS One*, 15(4), e0230929.
- Lin, G.-C., Wen, Z., Marsh, H. W., & Lin, H.-S. (2010). Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies. *Structural Equation Modeling*, 17(3), 374–391.
- Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). Oxford University Press.
- Liquin, E. G., & Lombrozo, T. (2022). Motivated to learn: An account of explanatory satisfaction. *Cognitive Psychology*, 132, 101453.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464–470.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3), 232–257.
- Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford University Press.
- Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10), 748–759.
- Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, 190(6), 975–995.
- Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, 122(3), 346–362.
- Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, 146(12), 1761.
- Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. (2013). Evaluating computational models of explanation using human judgments. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*.
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3), 429.
- Rosseeil, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, 48(2), 1–36.
- Scharrer, L., Bromme, R., Britt, M. A., & Stadler, M. (2012). The seduction of easiness: How science depictions influence laypeople's reliance on their own evaluation of scientific information. *Learning and Instruction*, 22(3), 231–243.

- Schupbach, J. N. (2011). Comparing probabilistic measures of explanatory power. *Philosophy of Science*, 78(5), 813–829.
- Shimojo, A., Miwa, K., & Terai, H. (2020). How does explanatory virtue determine probability estimation?— Empirical discussion on effect of instruction. *Frontiers in Psychology*, 11, 3444.
- Soler-Toscano, F., Zenil, H., Delahaye, J.-P., & Gauvrit, N. (2014). Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines. *PLoS One*, 9(5), e96223.
- Thagard, P. (1978). The best explanation: Criteria for theory choice. *Journal of Philosophy*, 75(2), 76–92.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12(3), 435–467.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48.
- Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children’s preference for simpler hypotheses. *Psychonomic Bulletin & Review*, 24(5), 1538–1547.
- Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies*, 176(10), 2807–2831.
- Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education*, 24(9), 1059–1077.
- Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, 142(4), 1006.
- Wojtowicz, Z., & DeDeo, S. (2020). From probability to consilience: How explanatory values implement Bayesian reasoning. *Trends in Cognitive Sciences*, 24(12), 981–993.
- Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, 24(5), 1488–1500.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.