Simplicity and Goodness-of-Fit in Explanation: The Case of Intuitive Curve-Fitting

Samuel G. B. Johnson (samuel.johnson@yale.edu)
Department of Psychology, Yale University
2 Hillhouse Ave., New Haven, CT 06520 USA

Andy Jin (ajin@live.unc.edu)
Department of Psychology, University of North Carolina, Chapel Hill
235 E. Cameron Ave., Chapel Hill, NC 27599 USA

Frank C. Keil (frank.keil@yale.edu)
Department of Psychology, Yale University
2 Hillhouse Ave., New Haven, CT 06520 USA

Abstract

Other things being equal, people prefer simpler explanations to more complex ones. However, complex explanations often provide better fits to the observed data, and goodness-of-fit must therefore be traded off against simplicity to arrive at the most likely explanation. In three experiments, we examine how people negotiate this trade-off. As a case study, we investigate laypeople's intuitions about curve-fitting in visually presented graphs, a domain with established quantitative criteria for trading off simplicity and goodness-of-fit. We examine whether people are well-calibrated to normative criteria, or whether they instead have an underfitting or overfitting bias (Experiment 1), we test people's intuitions in cases where simplicity and goodness-of-fit are no longer inversely correlated (Experiment 2), and we directly measure judgments concerning the complexity and goodness-of-fit in a set of curves (Experiment 3). To explain these findings, we posit a new heuristic: That the complexity of an explanation is used to estimate its goodness-of-fit to the data.

Keywords: Explanation; causal reasoning; intuitive statistics; simplicity; diagnostic reasoning.

Introduction

The simplest explanation is often the most likely to be true. Suppose, for example, that a patient has two symptoms: a runny nose (E_1) and a cough (E_2). Three diagnoses are available: influenza (H_1), which explains both symptoms; Hay fever (H_2), which explains only the runny nose; and Strep throat (H_3), which explains only the cough. If these diseases are equally common, then any reasonable doctor would surely posit influenza (H_1) to explain both symptoms, rather than hay fever to explain the runny nose and Strep throat to explain the cough (H_2 & H_3). Laypeople share these intuitions: Other things being equal, adults and children prefer explanations invoking fewer causes and believe they are more likely to be true (Bonawitz & Lombrozo, 2012; Lombrozo, 2007).

These intuitions relate to normative probabilistic reasoning (Lombrozo, 2007). The posterior odds favoring the simple explanation [H_1] over the complex explanation [H_2,H_3] is the product of the prior odds [$P(H_1)/P(H_2,H_3)$], representing the probability of each hypothesis being true before observing the evidence, and the likelihood ratio [$P(E_1,E_2|H_1)/P(E_1,E_2|H_2,H_3)$], representing the probability of the evidence being observed under each hypothesis. If we assume that the simple explanation (influenza) and the complex explanation (hay fever and Strep throat) lead to both symptoms with the same probability, then the likelihood ratio is 1. Lombrozo (2007) noted, however, that the prior odds often favor simple explanations: If H_1, H_2, and H_3 have equal base rates, then $P(H_2,H_3) = [P(H_1)]^2$. Since the complex explanation has a lower prior, the posterior odds favor the simpler explanation.

However, the posterior odds do not *always* favor the simpler explanation, because the evidence is often made more likely (or, put differently, is "better modeled") by a complex explanation than by a simple explanation, leading the likelihood ratio to favor the complex explanation. For example, consider a patient who has a runny nose and a stomach ache. One explanation is that the patient has only a cold, which would make a runny nose very likely and would make a stomach ache plausible (but not especially likely). Alternatively, the patient could have both a cold *and* a stomach virus, which would make both symptoms very likely. The more complex explanation has a lower prior, but a higher likelihood. This is often the case: The more (generative) causes invoked by an explanation, the lower its prior, but the more likely it is to predict the observed effects because each added cause has the potential to produce the observed effects. How do people negotiate this trade-off between simplicity and goodness-of-fit?

One possibility is that people would require disproportionate evidence to abandon a simpler explanation in favor of a more complex one. Lombrozo (2007) found that people use a *simplicity heuristic*,

assigning higher priors to simpler explanations, with many participants recalling simpler explanations as having higher base rates than warranted by the data. This heuristic led to non-normative explanatory preferences when probabilities were manipulated explicitly: When a simple and complex explanation had equal priors, 80% of participants preferred the simple explanation, and 40% of participants preferred the simple explanation even when it was 10 times *less* likely than the complex explanation.

However, people could also use a *complexity heuristic*, wherein complexity is used to estimate the likelihood or goodness-of-fit. Since simpler explanations (i.e., invoking fewer generative causes) often have higher prior probabilities, and complex explanations (i.e., invoking more generative causes) often have higher likelihoods, this pair of opponent simplicity and complexity heuristics may at times allow for negotiation of the simplicity/goodness-of-fit trade-off in a computationally tractable manner, without the need for explicit probabilistic reasoning. When in operation, a complexity heuristic would be in tension with a simplicity heuristic, potentially leading to no bias or even a complexity bias.

Here, we examined laypeople's intuitions about curve-fitting, a domain with established quantitative criteria for trading off these explanatory virtues (Forster & Sober, 1994). Although curve-fitting and verbal explanation tasks such as diagnostic reasoning are superficially quite different, they share deep formal similarities in that both tasks require trade-offs among explanatory virtues. Intuitions about curve-fitting are therefore a useful test case for explanatory reasoning more broadly—one where the virtues of simplicity and goodness-of-fit are operationalized in an especially direct way.

In these studies, we collected judgments of what family of curve was most appropriate when fitting scatterplot data, comparing those judgments to normative benchmarks (Experiment 1) and to judgments in matched cases where simplicity and goodness-of-fit no longer competed (Experiment 2). We also sought direct evidence for a complexity heuristic by measuring both perceived complexity and perceived goodness-of-fit for a set of curves (Experiment 3). If people use a complexity heuristic, we would expect an illusory increase in perceived goodness-of-fit for more complex curves.

Experiment 1

Consider the task of choosing how complex of a curve to use in fitting a set of datapoints (see Figure 1). It is assumed that these data were produced by both an underlying signal (the same each time a sample is drawn from the population) and random noise (which is different each time a new sample is drawn). Choosing a very complex curve will result in a tight fit to the current set of data points, but such a curve is likely to *overfit*—fitting the noise in addition to the underlying trend, resulting in poor predictive value for a new sample from the same population. In contrast, choosing a very simple curve may result in *underfitting*—failing to take advantage of all the information available in the original dataset for identifying the underlying trend, again leading to poor predictions. This trade-off between simplicity and goodness-of-fit is made by optimizing model selection criteria known as Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which reward models for having good fits to the data and penalize models for using larger numbers of parameters.

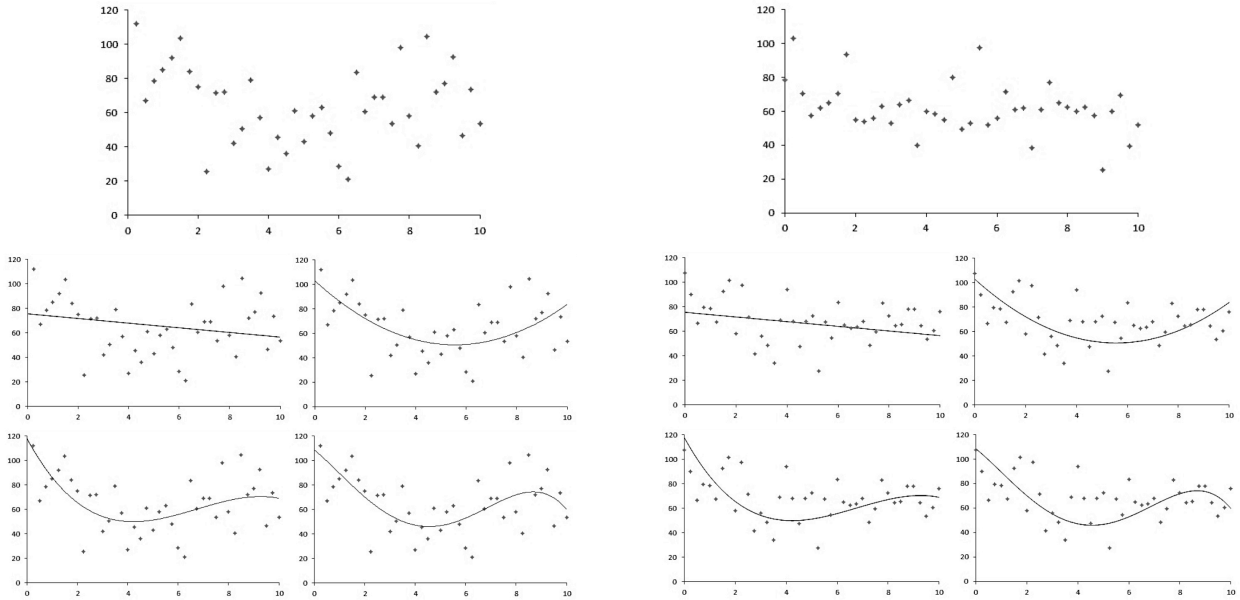
To see whether people make this trade-off in accordance with normative model selection criteria, we generated a set of 16 scatterplots, 8 of which had normatively quadratic fits according to both AIC and BIC, and 8 of which had normatively cubic fits according to both AIC and BIC. Participants were asked to select the best curve for each scatterplot from multiple choice options of degrees 1 through 4 (see Figure 1-A). Two wordings were used to elicit these judgments, between subjects—asking which curve "best represents" the underlying trend, and which curve would "best predict" a different set of data from the same population. The former wording might elicit more intuitive judgments, while the latter wording might make the task clearer to participants.

If people use the complexity of a curve as a cue for estimating its fit to the data, we might find an *overfitting* bias. Alternatively, it is plausible that people would instead have an *underfitting* bias, since simpler explanations are assigned high priors in a biased manner (Lombrozo, 2007). This would also be consistent with Little and Shiffrin's (2009) finding of an underfitting bias in the *generation* of curves for scatterplot data. Although *generating* hypotheses (here, best fit curves) involves computational challenges beyond *evaluating* hypotheses, this result nonetheless motivates the possibility that underfitting would extend to an evaluation task.

Method

Participants We recruited 80 participants from Amazon Mechanical Turk; five were excluded from analysis because they failed a check question (see below).

Materials The materials were 16 datasets displayed in scatterplots, and their best fit curves of degrees one through four (see Figure 1-A). Each scatterplot plotted a dataset including 41 data points, sampled at intervals of 0.25 from 0 to 10 on the x-axis. The y-values were determined by taking the values of second and third degree polynomials and adding Gaussian noise to each data point. Two quadratic functions and two cubic functions were used, and four random datasets were generated from each of these functions—two at relatively high levels of noise, and two at relatively low levels of noise (mean R^2 = .279 vs. .475). In addition, the best fit curves of degrees one through four always differed from each other by at least 7% at one or more points, to ensure that the best fit curves could be discriminated from each other. In addition, the normative complexity of the data was always the same as that of the data-generating



(A) Experiment 1 (B) Experiment 2

Figure 1: Example stimulus items: (A) An item from Experiment 1. (B) The matched version of that item from Experiment 2. The top panels show scatterplots without curves, and the bottom panels show response options (curves of degrees 1–4).

function. For the quadratic functions, the mean R^2 was .276, .390, .411, and .434 for the best fit curves of degrees 1–4, respectively (and the quadratic fit is best according to AIC and BIC), and for the cubic curves, the mean R^2 was .183, .364, .469, and .490 for fits of degrees 1–4 (and the cubic fit is best according to AIC and BIC).

Procedure Participants were instructed that they would see sets of data plotting the relationship between two properties of minerals called 'caltedness' and 'limency', and that each dataset would correspond to a different mineral. Fictitious physical properties were chosen because participants would likely have prior expectations (if any) of a linear relationship, providing a stronger test against the possibility of overfitting.

For each dataset, participants were shown a scatterplot (e.g., the top panel of Figure 1-A) and told that "The following scatter plot shows multiple measurements of caltedness and limency for a sample of the mineral [*mineral name*]. Each measurement is affected by both the inherent relationship between caltedness and limency in [*mineral name*], as well as by random errors such as variability from sample to sample and imprecision in measuring equipment," where a different mineral name was given for each dataset. On the next page, participants were presented with four multiple choice options (e.g., the bottom panel of Figure 1-B), each an image of the dataset (displayed in blue) with a best fit curve (displayed in black) overlaid, of degrees one through four. Between-subjects, participants were told either to select the option "that you believe best represents the relationship between caltedness and limency for [*mineral name*]" in the *represents* condition, or "that you believe would best predict the relationship between caltedness and limency for a different sample of [*mineral name*]" in the *predicts* condition. The datasets were presented in a random order, and the response options were randomized for each item.

After the main task, participants completed a check question (a dataset for which the quadratic fit was clearly optimal). Five participants were excluded from data analysis because they answered this question incorrectly.

Results and Discussion

The multiple choice options corresponded to the best fit curves for each dataset for degrees one through four, and therefore formed an interval scale from 1 to 4. Unbiased performance on this task would yield a score of 2.5 on this scale (the midpoint), because half of the curves were of degree 2 and half were of degree 3. Instead, participants selected curves with mean degree 2.72 (SD = 0.70), which is significantly greater than 2.5, $t(74) = 2.74$, $p = .008$, $d = 0.32$. A mixed-model ANOVA with degree (quadratic or cubic) and noise (high or low) as within-subjects factors and wording (represents or predicts) as a between-subjects factor revealed only a main effect of degree, $F(1,73) = 23.81$, $p < .001$, $\eta_p^2 = .25$, because participants fit higher degree curves to the normatively cubic ($M = 2.85$, $SD = 0.76$) than to the normatively quadratic ($M = 2.59$, $SD = 0.72$) datasets. In particular, there was no main effect of wording, $F(1,73) < 0.01$, $p = .97$, $\eta_p^2 < .01$, indicating that participants overfitted equally regardless of whether they were selecting the curve that best represented the data or the curve that would best predict a different set of data.

Could these effects can be accounted for by a response strategy such as random responding, centering, or

response variation? Although there appears to have been some regression to the mean (hence the slight underfitting for the cubic curves), these strategies would not lead to *biased* responses above the midpoint. Since the normative responses were degrees 2 and 3, and response options were curves of degrees 1–4, these strategies would lead to an unbiased mean near the midpoint of the scale.

This overall overfitting bias is surprising in light of an *underfitting* bias in curve generation (Little & Shiffrin, 2009) and the tendency to overestimate the priors of simple explanations (Lombrozo, 2007). This bias may occur because people use the complexity of a curve as a cue to goodness-of-fit, compensating for the simplicity heuristic that would push people toward underfitting. However, an alternative explanation is that the simplicity heuristic—that is, using simplicity to estimate prior probability—is not at work in this task, and people would choose relatively complex curves regardless of how well they fit the data. We test this possibility in Experiment 2.

Experiment 2

It is a mathematical truism that the *best* fit curve of a higher-degree polynomial family will be a better fit to the data than the best fit curve of a lower-degree polynomial family. However, it is not the case that *any* curve of higher degree will be a better fit than any curve of lower degree. By randomly perturbing the datasets used in Experiment 1, we generated a new set of scatterplots for which the curves used in Experiment 1 no longer exhibited the characteristic property that more complex curves are better fits to the data (see Figure 1-B). If people use simplicity as a proxy for prior probability (Lombrozo, 2007) in this task, then they should no longer choose complex curves for these scatterplots. In contrast, if people choose relatively complex curves for any dataset (regardless of actual fit), then the results of Experiment 2 would be similar to Experiment 1.

Method

Participants We recruited 80 participants from Amazon Mechanical Turk; 15 were excluded because they failed the check question used in Experiment 1.

Materials The curves were identical to those used in Experiment 1, but the datasets were perturbed randomly so that the quartic curves were no longer the best fits. Instead, for each dataset, the linear curve was a slightly better fit than the quadratic curve, and so forth, subject to the constraint that R^2 for the linear curve be no more than .03 greater than R^2 for the quartic curve (see Figure 1-B).

Procedure The procedure was the same as Experiment 1.

Results and Discussion

Even though the curves used in Experiment 2 were identical to those in Experiment 1, participants selected curves of mean degree 2.12 (SD = 0.74), which is significantly *below* the midpoint of the scale, $t(64) = -4.17$, $p < .001$, $d = -0.52$. A two-way ANOVA with experiment (Exp. 1 or Exp. 2) and wording (represents or predicts) as between-subjects factors revealed only a main effect of experiment, $F(1,136) = 24.15$, $p < .001$, $\eta_p^2 = .15$. There was no main effect of wording, $F(1,136) = 0.18$, $p = .67$, $\eta_p^2 < .01$, nor an interaction of wording with experiment, $F(1,136) = 0.22$, $p = .64$, $\eta_p^2 < .01$, showing that these effects did not depend on whether participants were selecting the curve that best represented the data or would best predict a different set of data.

These results are consistent with Lombrozo's (2007) finding that simplicity is used as a proxy for prior probability. This experiment also acts as a control for Experiment 1, showing that the overfitting found in Experiment 1 occurred as a consequence of the increasing values of R^2 associated with increasingly complex curves—when R^2 was approximately constant between curves, people tended to choose relatively simple curves.

Nonetheless, participants still chose curves considerably above the floor of the scale—the mean of 2.12 indicates that participants chose quadratic curves on average. Although scale-use strategies (e.g., centering or response varying) may at least partially account for this result, one might nonetheless expect participants to more frequently choose linear curves, since they were both the simplest *and* the best-fitting curves for each scatterplot. One possibility is that people's estimates of goodness-of-fit were distorted by the complexity of the curves. If people used complexity to estimate goodness-of-fit, then they might choose more complex curves simply because they believed they *were* better fits to the data. This possibility was tested in Experiment 3.

Experiments 3A and 3B

To test directly whether complexity is used as a heuristic for estimating goodness-of-fit, participants in Experiment 3A provided their estimates of how well the curves fit the data for the 64 curves used as response options in Experiment 2. Because perceived complexity is not necessarily a linear function of the curve's degree, we collected perceived complexity judgments in Experiment 3B. We expected that both actual goodness-of-fit (as measured by R^2) and perceived complexity would independently influence goodness-of-fit judgments.

Method

Participants We recruited 120 participants from Amazon Mechanical Turk for Experiment 3A, and 40 participants for Experiment 3B. (A greater number of participants were necessary for Experiment 3A due to lower between-items variance.) Seven participants from Experiment 3A and zero participants from Experiment 3B were excluded because they failed check questions (see below).

Materials The materials for Experiment 3A were the 64 scatterplots and curves of degrees 1 through 4 used as response options in Experiment 2 (see Figure 1-B). Therefore, the R^2 values were slightly higher for the linear than for the quartic curves (see Table 1). These curves

Table 1: Means and SDs in Experiments 3A and 3B.

Degree	**1**	**2**	**3**	**4**
Goodness-of-Fit	51.47	54.02	54.06	52.48
Exp. 3A	(19.55)	(19.17)	(19.18)	(19.18)
Complexity	9.90	51.64	63.94	71.40
Exp. 3B	(1.22)	(3.57)	(7.36)	(5.04)
Actual R^2	.523	.515	.507	.499
Arclength	30.36	53.01	63.16	68.19

also varied in noise, since higher and lower noise datasets were used in Experiments 1 and 2. The materials for Experiment 3B were the same 64 curves, but with the datapoints omitted.

Procedure For Experiment 3A, participants were instructed that for each graph, they would "judge how closely the black line fits the blue data points" and were given examples of two identical quadratic curves, one with data that it fit poorly and the other with data that it fit well. They were then shown each of the 64 scatterplots with best fit lines, and rated "how closely you think the black line fits the blue data points" on a scale from 0 ("Very poor fit") to 100 ("Very close fit"). After the main task, participants responded to two check questions with linear curves that were, respectively, very close and very poor fits. Participants who rated the close fitting curve no more than 20 points higher than the loose fitting curve were excluded from analysis.

For Experiment 3B, participants were instructed that for each graph, they would "judge how complex the line is" and told that "relatively simple curves can be described in a small amount of information, while relatively complex curves require more information to describe." Participants were given examples of a less complex (linear) curve and a more complex (quadratic) curve. They were then shown each of the 64 best fit lines without the datapoints, and rated "how complex you think the line is" on a scale from 0 ("Very simple") to 100 ("Very complex"). After the main task, participants responded to two check questions with curves of degrees 1 and 5. Every participant rated the curve of degree 5 more complex than the curve of degree 1 by at least 20 points, so no participants were excluded from analysis.

Results and Discussion

Participants in Experiment 3A based their goodness-of-fit ratings both on the noisiness of the data and on the complexity of the curve (see Table 1 for item means and SDs). A repeated measures ANOVA was conducted with degree (linear, quadratic, cubic, or quartic) and noise (high or low) as within-subjects variables. There was a main effect of noise, $F(1,112) = 938.96$, $p < .001$, $\eta_p^2 = .89$, because the high noise curves were judged looser fits than the low noise curves ($M = 35.35$, $SD = 13.14$ vs. $M = 70.66$, $SD = 9.99$). More interestingly, however, there was also a main effect of degree, $F(3,336) = 9.68$, $p < .001$, $\eta_p^2 = .08$, with linear curves judged the loosest fit even though they actually had the highest R^2 values (albeit only slightly). Quadratic and cubic curves were judged the best fits, and quartic curves somewhat worse. Possible reasons for this non-monotonic pattern are discussed below.

Participants in Experiment 3B used degree to judge complexity, $F(3,117) = 297.29$, $p < .001$, $\eta_p^2 = .88$, giving higher complexity ratings with increasing degree (see Table 1). However, degree had diminishing returns on perceived complexity, resulting in a large boost between linear and quadratic curves, but a more modest boost between cubic and quartic curves.

These results are not what would be expected if participants were judging goodness-of-fit based solely on R^2. One possibility is that the nonmonotonic pattern in Experiment 3A occurred because participants used both perceived complexity and the actual R^2 to estimate goodness-of-fit. According to this explanation, since there was a large increase in perceived complexity between the linear and quadratic curves but only a small decrease in actual R^2, participants would judge the quadratic curves better fits than the linear curves. In contrast, there was a much smaller increase in perceived complexity between the cubic and quartic curves but the same decrease in actual R^2. This would lead to the overall decrease in perceived goodness-of-fit found in Experiment 3A.

As a first test of this possibility, we computed for each of the 16 datasets the partial correlation between simplicity ratings (Exp. 3A) and goodness-of-fit ratings (Exp. 3B) across degrees, controlling for actual R^2. The mean partial correlation for each dataset was $r = .54$, which is significantly different from 0 in a one-sample *t*-test, $t(15) = 3.66$, $p = .002$, $d = 0.92$. This shows that when R^2 is held constant, more complex curves tend to be rated better fits to the data.

We followed up with a more fine-grained path analysis (Kline, 1998) with the 64 scatterplots as the units of analysis, and mean ratings of goodness-of-fit (from Exp. 3A) and complexity (Exp. 3B) as endogenous (dependent) variables. To explore other factors that might affect complexity ratings, we computed each curve's arclength and used this as an exogenous (predictor) variable, along with actual R^2 and each curve's degree.

As shown in the path diagram (Figure 2), degree ($\beta = .69$, $p < .001$) and arclength ($\beta = .33$, $p < .001$) both contributed to perceived complexity, with curves of higher degree and greater arclength receiving higher complexity ratings. Most importantly, perceived complexity was a significant predictor of goodness-of-fit ratings ($\beta = .57$, $p = .005$) after all other variables were taken into account, showing that people used complexity as a proxy for goodness-of-fit. The actual R^2 also contributed to goodness-of-fit ratings ($\beta = .77$, $p < .001$). Curiously, controlling for the other variables, curves of greater arclength were actually perceived as *worse* fits to the data ($\beta = -.76$, $p < .001$). This may have occurred because the relative density of the datapoints compared to the length of the curve is lower for longer curves. The

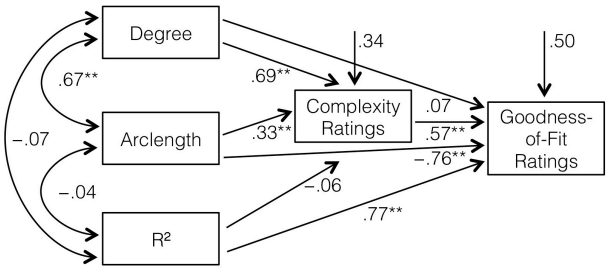


Figure 2: Path diagram for Experiment 3.
Paths marked with ** are significant at $p < .01$.

negative association between arclength and perceived fit speaks against the possibility that the relationship between perceived complexity and fit occurred because participants were judging absolute distance to the curve rather than vertical distance (measured by R^2). If that were the case, one would expect arclength to be positively associated with fit ratings, since longer curves have more variability along the y-axis and hence more opportunities for R^2 and absolute distance to differ.

These results support our hypothesis that complex explanations create an illusion of goodness-of-fit, at least in intuitive curve-fitting. Increases in perceived complexity were accompanied by increases in perceived goodness-of-fit, even when factors such as degree, arclength, and actual goodness-of-fit are controlled for.

General Discussion

In three experiments, we investigated how people negotiate the trade-off between simplicity and goodness-of-fit in explanations by testing intuitions about curve-fitting. In Experiment 1, participants showed an overall overfitting bias—a surprising result in light of Lombrozo's (2007) finding that people assign higher prior probabilities to simple explanations in a biased manner. In Experiment 2, we showed that when simplicity and goodness-of-fit no longer compete, people prefer simpler curves, consistent with Lombrozo (2007). To explain these findings, we hypothesized that people use a *complexity* heuristic, in which complexity is used to estimate goodness-of-fit. In Experiment 3, we provided direct evidence for this claim, with more complex curves judged better fits to the data, after factors such as the actual goodness-of-fit are taken into account.

Overall, these findings support an *opponent heuristic* account of simplicity preferences, with simplicity used in two opposing ways when evaluating explanations. First, people appear to assign higher prior probabilities [P(H)] to simpler explanations. Lombrozo (2007) showed this directly, with participants systematically overestimating the recalled base rates of simple explanations. Similarly, participants in our Experiment 2 showed a simplicity preference when goodness-of-fit was held constant. Second, people appear to assign higher likelihoods [P($E|H$)] to complex explanations. We provided direct evidence for this claim in Experiment 3, where judgments of how well a curve fit a set of datapoints were influenced by the perceived complexity of the curve.

Why might people use a complexity heuristic? One possibility is that people are sensitive to the normative structure of the environment, where more complex explanations (in the sense of having more generative causes) often provide better fits to the data. Alternatively, this heuristic may involve pragmatic or pedagogical inferences. Since more complex explanations are selected from a larger hypothesis space compared to simpler explanations, people may infer that more complex explanations are unlikely to be selected unless there was a reason for selecting that one in particular—in particular, that it had a better fit to the data. Converging evidence from other tasks such as verbal diagnostic reasoning can help to tease apart these possible accounts.

Evidence from other tasks can also help to clarify the relationship between these findings and other lines of research. For instance, a complexity heuristic could in part explain why people seem to prefer verbal explanations involving more technical vocabulary (Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). Another direction for future work is to clarify the relationship between cognitive and perceptual explanation. The results of Experiment 2 suggest a visual simplicity heuristic complementing Lombrozo's (2007) results from verbal tasks. This raises the possibility that explanatory heuristics may be used widely across cognition, and perhaps even perception.

Acknowledgments

This research was supported by a grant from the National Institutes of Health to F.C. Keil. We thank the members of the Cognition & Development Lab for helpful feedback.

References

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

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*, 1–35.

Kline, R.B. (1998). *Principles and practice of structural equation modeling* (1st Ed.). New York: Guilford Press.

Little, D.R., & Shiffrin, R.M. (2009). Simplicity bias in the estimation of causal functions. *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1157–1162). Austin, TX: Cognitive Science Society.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257.

Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E., & Gray, J.R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*, 470–477.