Explaining drives the discovery of real and illusory patterns

Joseph Jay Williams (joseph_williams@berkeley.edu)
Tania Lombrozo (lombrozo@berkeley.edu)
Department of Psychology, University of California, Berkeley

Bob Rehder (bob.rehder@nyu.edu)
Department of Psychology, New York University

Children's and adults' attempts to explain the world around them plays a key role in promoting learning and understanding, but little is known about how and why explaining has this effect. An experiment investigated explaining in the social context of learning to predict and explain individuals' behavior, examining if explaining observations exerts a selective constraint to seek patterns or regularities underlying the observations, regardless of whether such patterns are harmful or helpful for learning. When there were reliable patterns– such as personality types that predict charitable behavior– explaining promoted learning. But when these patterns were misleading, explaining produced an *impairment* whereby participants exhibited less accurate learning and prediction of individuals' behavior. This novel approach of contrasting explanation's positive and negative effects suggests that explanation's benefits are not merely due to increased motivation, attention or time, and that explaining may undermine learning in domains where regularities are absent, spurious, or unreliable.

Keywords: explanation, self-explanation, learning, understanding, generalization, pattern detection, explanation impairment effect

Explanation appears to possess a privileged relationship with learning and understanding. To know a fact without knowing *why* it is true can be deeply unsatisfying, not only to career learners like scientists but also to everyday learners and young children. Engaging in explanation goes beyond rote knowledge to genuine understanding, bringing with it the ability to generalize what is learned to novel situations.

Research in education and cognitive development confirms and sheds light on the close connection between explanation and learning. Educational studies on topics ranging from math and physics to biology and computer programming have found that generating explanations has a powerful impact on learning and generalization (Chi et al, 1994; Renkl, 1997; for a review see Fonseca & Chi, 2010). Even young children exhibit an insatiable desire to request and learn from explanations (Chouinard, 2008; Legare et al, 2009), with prompts to explain accelerating major conceptual transitions in number conservation and theory of mind (Amsterlaw & Wellman, 2006; Siegler, 2002).

The importance of explanation has been recognized in other disciplines as well. In cognitive psychology, explanations are believed to play a central role in the representation of conceptual knowledge, especially knowledge about causal relationships (Carey, 1985; Murphy & Medin, 1985). Research in artificial intelligence on how machines learn has been inspired by a focus on explanation as a process for learning from individual cases (DeJong, 1986; Mitchell et al, 2006). Finally, philosophers of science have attempted to characterize the nature of scientific explanation (Woodward, 2009).

Despite the broad relevance of explanation, little is known about why the process of explaining, in particular, drives effective learning. Previous work has identified explanation's role in revising beliefs and providing metacognitive insight into what is not known (Chi, 2000). Other investigators have proposed that explaining increases motivation and attention (e.g., Siegler, 2002). However, little experimental work has directly investigated and compared alternative theories of the content and consequences of explanation. This leaves important questions unanswered: What is the nature of the cognitive processing invoked by explaining? And why are the relative benefits of explanation greatest in acquiring knowledge that supports generalization?

This paper investigates the hypothesis that the process of explaining drives the explainer to seek general patterns or regularities that can account for or produce whatever observation is the target of explanation. This hypothesis is the central tenet of the *subsumptive constraints* account of explanation (Williams & Lombrozo, 2010a), which is motivated by work in philosophy on pattern subsumption and unification theories of scientific explanation (Kitcher, 1981). Subsumption and unification theories suggest the defining property of an explanation is that it shows how the observation being explained is an instance of (subsumed by) a general pattern or regularity. For example, in answering "Why did that apple fall?" with "Because gravity accelerated it towards the Earth," a hypothetical Newton shows how a particular event is subsumed under a general pattern, in this case a law of gravitation. Furthermore, the greater the number and diversity of observations attributable to a single pattern, the better the explanation.

If learners are sensitive to a subsumptive constraint on explanations, then asking "Why?" should implicitly constrain their thinking, driving them to seek general patterns that underlie what they are trying to explain. And because patterns typically go beyond the idiosyncratic properties of the individual observations being explained, engaging in explanation should generate knowledge that transfers and generalizes to new contexts and problems, such as knowledge about underlying principles, laws, relationships, and causal regularities. The subsumptive constraints account thus sheds light on why explanation promotes learning, and especially generalization.

However, a subsumptive constraint also has a hidden danger: What happens if people seek explanations in contexts where underlying patterns do not exist, or are imperfect and misleading? If explanation exerts a subsumptive constraint, it will still drive a search for patterns, and if people "discover" spurious or misleading regularities, explaining will compromise learning. The human preoccupation with explanation offers many opportunities for this constraint to lead people astray. In the context of social interactions, a tendency to explain other people's behavior could drive a search for generalizations even when they are unreliable, at the expense of learning about the individual. For example, instead of simply noting that a friend Anna frequently donates to charities, explaining that behavior might drive the "discovery" of a spurious or misleading generalization that invokes her social group (e.g. female, student) to explain the behavior.

The prediction that explaining can *impair* learning is counterintuitive and stands out against a wealth of evidence demonstrating broad benefits for engaging in explanation (for a rare exception see Kuhn & Katz, 2009). Rather than stemming from a selective constraint to find patterns, an alternative *learning engagement* account of explanation's effects is that they arise through a general increase in engagement with the current learning activity. For example, explaining may help learning because it increases motivation, study time, or attention (for discussion see Siegler, 2002) – factors that are already known to be a powerful force in learning (e.g., Pintrich & Schunk, 2002).

Both the subsumptive constraints account and the learning engagement account predict beneficial effects of explanation in a broad range of contexts, albeit through very different mechanisms. A key divergence is in the untested prediction, generated by the subsumptive constraints account, that explaining will impair learning when patterns are misleading. The experiment reported here tests this prediction in the context of learning about people's behavior, investigating whether explaining interacts with the structure of what is being learned to produce benefits when there are patterns that support learning, but slower and more inaccurate learning when patterns are misleading.

The strategy of investigating explanation by contrasting its *costs* with its benefits may serve the same function as visual illusions. Just as visual illusions illuminate the mechanisms underlying successful perception, *explanation impairment effects* can reveal the mechanisms underlying explanation's profound effects on learning and generalization. Moreover, understanding when the drive to explain leads to false "discoveries" and misleading beliefs is consequential in its own right.

Enhancement and impairment of learning through explaining behavior

Preliminary work on learning novel categories (Williams, Lombrozo, & Rehder, 2010) provided some evidence that explaining drives people to find underlying patterns and impairs learning when the pattern is unreliable. Participants learned about novel categories of vehicles by classifying examples and receiving feedback. Learning about category membership could proceed by using information unique to each example (e.g. color) or a pattern about the vehicles' intended environment (arctic versus jungle) that could be reliable or misleading.[1] While half of participants were prompted to *explain* why an example was in a category, the other half were instructed to *think aloud* to control for the effects of verbalization without exerting the subsumptive constraints of explanation. The experiment found an interaction of explanation with the reliability of the pattern: explaining slowed learning of the novel category when an unreliable pattern was present.

However, this study suffers from an alternative interpretation of the results in terms of implicit task demands: participants may have inferred from the prompt to explain that the experimenter would not ask them to explain unless a pattern was present. On this account, participants' increased attempts to find patterns was due to their beliefs that such patterns existed rather than explaining per se.

Accordingly, a goal of the present work is to establish that it is truly *generating* explanations that drives learners' search for patterns. To this end, we compared an *explain* condition with an *anticipated explanation* control condition in which participants were informed before learning that they would later have to explain. Before and during the learning phase participants therefore believed they would later have to provide explanations to the experimenter. These two conditions are thus equated on the task demand of implying the presence of a pattern while still differing in the extent to which learners generate explanations.

Moreover, the current experiment used social materials (predicting and explaining people's behavior based on descriptions about them) which provides a significant generalization of the previous experiment on artificial category learning. Predictions about behavior differ from artificial category learning concerning vehicles in the beliefs they draw on and the level of prior knowledge available, in the nature and goals of the judgments, and in the degree of personal and emotional relevance. Learning about people's behavior is also an important capacity for navigating the social world and interpersonal relationships. In social psychology, research has examined different kinds of explanations for behavior (Malle, 2004), demonstrated that generating explanations can influence expectations (Wilson & LaFleur, 1995; Sanitioso, Kunda, & Fong, 1990) and even suggested that explanatory considerations play a role in the acquisition, representation, and justification of stereotypes (McGarty, Yzerbyt, & Spears, 2002).

However, no research has tested whether generating explanations for behavior drives the interpretation of behavior in terms of underlying generalizations, rather than simply learning about person-specific knowledge. If explaining drives a search for patterns that link behavior to general social categories – whether such links are reliable or

[1] These materials were adapted from Kaplan & Murphy (2000).

not – it could play a role in the construction of generalizations represented in stereotypes. For example, instead of simply encoding the fact that a friend Anna frequently donates to charities, one could attempt to explain the behavior by noting that she is an extravert, a member of a particular ethnic group or social class, or a woman. Belief in potential relationships between a social group and behavior–e.g. extraverts frequently donate to charities– could be promoted when these are invoked in explanations– "Anna frequently donates to charities, because… she is extraverted"– either directly or through subsequent biases in confirmation. In sum, if explaining drives a search for patterns, the illusory "discovery" of spurious or misleading generalizations about behavior and social groups could foster erroneous stereotypical beliefs, and impair accurate learning and prediction of people's behavior.

Experiment

In this experiment, participants used descriptions of individuals (e.g. picture, age, personality, major) to predict each individual's behavior (whether they rarely or frequently donated to charities). Accurate learning and prediction of behavior could proceed either through the use of individuating information specific to each person– e.g. Anna, the 29-year old with red hair, frequently donates to charities– or by discovering an abstract, underlying pattern– e.g. individuals with extraverted personality traits, like being friendly or self-assured, frequently donate to charities. The experiment manipulated whether this pattern was *reliable*, meaning that its use led to 100% prediction accuracy, or *misleading,* meaning that its use led to 80% prediction accuracy and 20% errors. For both reliable and misleading patterns, participants were either asked to *explain* why a person engaged in a behavior or they participated in the *anticipated explanation* control condition, in which they were instructed of a future explanation task but not required to perform it during study.

A learning engagement account predicts that explaining will have the general effect of enhancing learning, whether a reliable or misleading pattern is present. For example, participants may be more motivated to make accurate predictions and utilize feedback, and more likely to spend time and attention studying and thinking about the descriptions of people and the behavior they engage in. If a reliable pattern supports prediction then highly engaged explainers may utilize it to learn more quickly than non-explainers, but there is no reason for them to perseverate on a misleading pattern when they can improve performance by learning about individuating information or encoding exceptions to the pattern. In contrast, a subsumptive constraints account predicts an interaction, whereby explaining speeds learning if a reliable pattern is present, but *impairs* learning when the pattern is misleading, generating greater prediction error on our task. This is because attempting to generate explanations should drive learners towards unifying patterns even in the face of errors or exceptions, generating perseveration on imperfect patterns that will slow learning.

The current experiment also bolstered the generality of the findings through a number of changes from the previous study on category learning (Williams et al, 2010). Learning took place for a fixed number of blocks rather than to a learning criterion, control participants were not required to think out loud, and the pattern-related features that provided the exceptions to the misleading pattern were fixed rather than changing from block to block.

Participants Of the 188 participants who participated so far, 76 were UC Berkeley undergraduates who participated in the lab for course credit and 112 were online participants from Amazon Mechanical Turk who received monetary compensation.

Materials Table 1 summarizes the 10 person descriptions that were studied in the *reliable* pattern condition. In the experiment participants used each individual's description to predict whether that individual rarely or frequently donates to charities. Each description consisted of six features. There was one *pattern-related* feature that was unique but an instance of extraversion/introversion (e.g. dominating, cautious), and so discovery of this single generalization (e.g. extraverted people frequently donate to charities, introverts rarely) permitted predictions about all 10 individuals' charitable behavior. There were three *individuating* features that were unique to each person (the person's picture, name and age) and so prediction could proceed by associating these with the individual's charitable behavior, although this required associating individuating features with behavior for all 10 individuals. These features were selected so that no obvious pattern (such as age and gender) was diagnostic of rarely/frequently donating to charity. Two *irrelevant* binary features (e.g. lives on West [East] coast) were not informative about charitable behavior.

In every presentation of a description the picture and name were always listed first while the order of all other features was randomized. To ensure that effects of explanation did not depend on particular prior knowledge, the pairing of charitable behavior (rarely/frequently donates) with the extraverted/introverted pattern was counterbalanced across participants to create the factor *pattern-behavior pairing*: for half extraversion was linked to rarely donating to charities (introversion-frequently) and for half extraversion was linked to frequently donating (introversion-rarely). The materials were identical in the *misleading* pattern condition, except for a critical change: the personality traits of two people (Kevin and Karen) were switched so that the extraverted/introverted pattern now only predicted behavior for 8 of the 10 people and resulted in prediction errors for the other 2 people.

Procedure All participants were instructed that they would observe descriptions of people and should learn (in

preparation for future testing) which ones rarely and frequently donate to charities. On each learning trial participants had 10 seconds to judge from an individual's description whether the person rarely or frequently donates to charities. They then saw the person's actual behavior and studied it along with the description for a further 10 seconds. In the *explain* condition participants were instructed that once they saw the person's actual behavior they should explain out loud why the person rarely (frequently) donates to charities. Participants in the lab were recorded using a voice recorder while those online were not. In the *anticipated explanation* condition participants were informed that they would later be asked to explain why each person rarely or frequently donates to charities, but were free to study as they chose. All participants were therefore aware that the experimenters expected them to be able to explain, but only the *explain* condition was required to do so *during* learning.

Unique features			**Pattern related features**	**Irrelevant features**	
Picture	**Name**	**Age**	**Personality**	**"living on the"**	**"a graduate of a"**
Rarely donates to charities (Frequently donates to charities)					
	Anna	37	dominating	East coast	science major
	Joseph	42	friendly	West coast	humanities major
	Sarah	28	boastful	West coast	science major
	Jessica	32	self-assured	East coast	science major
	Kevin	24	energetic (quiet)	West coast	humanities major
Frequently donates to charities (Rarely donates to charities)					
	Steven	30	cautious	East coast	science major
	Josh	26	discreet	West coast	humanities major
	Laura	23	studious	West coast	science major
	Janet	45	self-conscious	West coast	humanities major
	Karen	39	quiet (energetic)	East coast	science major

Table 1: Person descriptions in the experiment consist of individuating features, pattern-related features (extraverts/ introverts), and irrelevant features. In the misleading pattern condition, the personality features of the 5th and 10th people (Kevin and Karen) were switched.

After visual and audio presentation of the instructions, participants had practice trials on 6 descriptions and then read and listened to the instructions again. Participants then encountered four blocks of the 10 person descriptions (a total of 40 presentations) in which predicted behavior and studied feedback.

To assess what knowledge was acquired during learning, participants were presented with subsets of features from the person descriptions along with novel personality features and were asked to indicate whether a person with those features would rarely or frequently donate to charities. They also rated confidence in their judgment on a seven-point scale. There were four kinds of judgments, concerning pattern-related, transfer-pattern, and individuating features, as well as conflict items. These different items were all randomly interleaved. Knowledge about the relationship between the pattern and charitable behavior was assessed by presenting the 10 studied *pattern-related* personality features (e.g. talkative), as well as 8 novel *transfer pattern* personality features, which were associated with extraversion/introversion but not previously presented (e.g. talkative, reserved). Learning a link between an individual and their behavior was measured in predictions about the 10 studied triples of *individuating* features (picture, name, age). The 6 *conflict* items measured participants' preferred basis for prediction, by pitting novel pattern-related personality features against studied triples of individuating features to give opposite judgments.

To examine the use of the pattern-related and individuating features in generalization, participants made predictions about how likely (on a scale from 0 to 100) individuals were to engage in novel charitable behaviors (giving old clothes to the Salvation Army, supporting taxes that increase welfare programs, giving money to homeless people). Specifically, 6 *transfer pattern generalization* judgments used single novel personality features related to extraversion/introversion, and 4 *conflict generalization* judgments pitted novel personality features against studied triples of individuating features. In closing participants were asked to report what differences they observed between people who rarely and frequently donated to charities.

Results

To examine effects of pattern reliability on learning, the prediction errors during learning for person descriptions 5 and 10 were analyzed, as they were consistent with the pattern in the reliable pattern condition, but inconsistent with it when the pattern was misleading. This prediction error was subjected to a 2 (Block: 1st, 2nd, 3rd or 4th) x 2 (study condition: explain vs. anticipated explanation) x 2 (pattern type: reliable vs. misleading) x 2 (pattern-behavior pairing) x 2 (participation context: lab vs. online) mixed effects ANOVA. A significant block x study condition x pattern type interaction, $F(3, 172) = 3.91$, $p < 0.01$, revealed that the effect of explaining changed over time. In the misleading pattern condition, the degree to which explaining increased errors changed with additional exposure– the impairment was mitigated over time.

To examine the initial effects of explanation, Figure 1 shows prediction error for the first two blocks as a function of study condition and pattern type. The results confirm the predictions of a subsumptive constraints account. The key predicted interaction of explanation with the reliability of the pattern was revealed by a significant two way interaction of study condition and pattern type in a 2 (study condition: explain vs. anticipated explanation) x 2 (pattern type: reliable vs. misleading) x 2 (pattern-behavior pairing) x 2 (participation context: lab vs. online) ANOVA, $F(1, 172) =$

5.12, $p < 0.05$. When the pattern was misleading, explaining tended to impair learning about the exceptions to the pattern ($t(89) = 1.74$, $p = 0.085$), providing evidence against an account of explanation's effects in terms of learning engagement.

The ANOVA also revealed a significant main effect of pattern type, $F(1, 172) = 38.66$, $p < 0.001$, and a four way interaction of study type, pattern type, pattern-behavior pairing and participation context, $F(1, 172) = 5.89$, $p < 0.05$. This appeared to be due to differential effects of explaining on lab and online participants when learning about the pattern-behavior pairings. This could be due to differences in prior knowledge (e.g. concerning charitable behavior and extraversion/introversion) between the undergraduate students and online population.

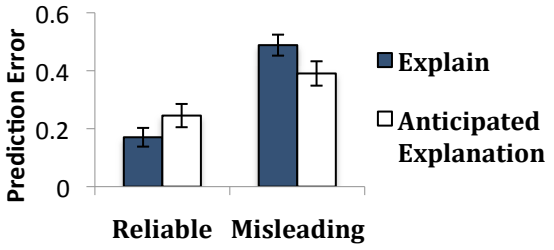

Figure 1: Proportion of prediction errors during first two blocks of learning the descriptions inconsistent with the pattern, as a function of study condition and pattern type.

As might be anticipated from the changing effects of explanation, the post-learning measures did not reveal significant effects of explanation, and are not reported in the interest of space. The extensive prediction, feedback, and study exposure of over 40 presentations may have mitigated the effects of explanation by the end of learning. In real-world contexts explaining may be more likely to foster persistent and misleading generalizations than in a laboratory task, as salient feedback on mistaken beliefs is typically less available.

Discussion

The current experiment provides evidence that engaging in explanation invokes a subsumptive constraint: Asking "why?" selectively constrains learners to find general regularities that underlie or produce the observations targeted by explanation, going beyond the individuating features of specific instances to underlying generalizations. In this experiment, explaining drove the discovery of a pattern that linked charitable behavior to having an extraverted versus introverted personality. When reliable patterns are present, explanation's selective constraint to find patterns can drive the discovery of accurate generalizations. But when patterns are misleading or spurious, attempts to explain still invoke the constraint to find patterns, which can then drive the illusory "discovery" of generalizations that are in fact unreliable and misleading, and thus impair learning and prediction.

The experiment was designed to address whether explanation's effects might be due to an implicit task demand, an alternative interpretation of a previous study (Williams et al, 2010), whereby participants may infer from the fact that the experimenter expects them to explain that there are patterns present, and so make a conscious decision to seek these patterns. This experiment provides evidence against this possibility: even when participants in both the explain and anticipated explanation conditions were informed that the experimenter later expected them to provide explanations, generating explanations during study enhanced and impaired learning through increased pattern seeking.

The subsumptive constraints account explains why explanation has a distinctive and profound impact on learning, generalization, and conceptual representation. For both scientists and everyday learners, the drive to *explain* rather than merely know or describe fosters true understanding: discovery of the general principles and laws that underlie particular observations. The patterns and regularities discovered through explaining have relevance beyond particular learning contexts and support future predictions, reasoning, and problem-solving in novel contexts. A subsumptive constraint similarly clarifies why explanations play a key role in the representation of conceptual knowledge. While the storage of facts, observations, and instances in memory is important for representing concepts, the distinctive contribution of explanations is that they capture unifying generalizations and regularities which foster a coherent understanding and provide the basis to flexibly deploy conceptual knowledge in new situations.

The reported explanation impairment effect shows that explaining does not impact learning merely by increasing learning engagement or boosting cognitive processing. We expect that multiple mechanisms play a role in explanation's effects and would not argue that this rules out an effect of learning engagement. However, if *increased* processing does not completely account for the current effects, a more fruitful question may concern the *nature* of processing. Increased attention and motivation could enhance memory for details, encoding of examples, prediction accuracy, or discovery of patterns. What does explaining selectively increase attention to and what exactly does it motivate people to learn?

The impairments observed when explaining in the presence of misleading patterns raise pressing issues and questions. Since explaining the behavior of others drives the discovery of misleading generalizations rather than simply learning about the behavior of particular individuals, engaging in explanation may be a mechanism for forming stereotypical beliefs. Explaining may promote beliefs that link behavior to social groups (e.g., introverts rarely donate to charities) or even produce novel causal generalizations (e.g., extraverts are generous because they like to interact

with people). Given the ubiquity of explanations, the finding that explaining encourages people to seek patterns raises concerns about the illusory discovery of spurious or misleading generalizations in a broad range of domains: detecting illusory correlations, student misconceptions in science education, the formation of conspiracy theories, and overgeneralization from small samples.

The finding that explaining "why?" drives learners towards underlying patterns can provide guidance on educational uses of explanation. The reported impairments caution that prompts to explain can be counterproductive (see also Kuhn & Katz, 2009) if students construct spurious patterns or identify misleading regularities. More successful explanations may be scaffolded by supplying prior knowledge that elucidates how observations are instances of a generalization, or by structuring the to-be-explained observations to highlight underlying principles. The current findings also raise the possibility that explaining "why?" plays a unique role. Many previous studies have examined spontaneous explanation while thinking aloud, or prompted explanations for the meaning of a sentence or paragraph (for a review see Fonseca & Chi, 2010), so that "explaining" refers to a heterogeneous collection of activities. While the current experiment suggests that explaining "why?" may highlight underlying principles, laws, and patterns (see also Renkl, 1997), self-explaining a sentence or procedure, explaining "what" a concept is, or "how" a process occurs may construct different kinds of knowledge or have differential effects on processes like metacognition.

Given the benefits and costs of the subsumptive constraints on explanation, examining children's development of a sensitivity to this constraint and how it aids or restricts their learning will be informative. Evidence for a subsumptive constraint on explanation raises the question of what the relationship is between explaining and other cognitive processes such as comparison and analogy, which also promote discovery of abstract generalizations. Further work will more precisely characterize the nature of the subsumptive constraint, such as how prior knowledge informs which patterns are judged to be subsuming and explanatory (Williams & Lombrozo, 2010b). The counterintuitive but revealing strategy of examining both the beneficial and harmful effects of explanation can shed light on these and other issues.

References

Amsterlaw, J., & Wellman, H. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development, 7,* 139-172.

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: Bradford Books, MIT Press.

Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Chouinard, M. (2007). Children's questions: a mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, 72, 1-57.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine learning*, *1*(2), 145–176.

Fonseca, B.A. & Chi, M.T.H. (2010). Instruction based on self-explanation. In Mayer, R. & Alexander, P. (Eds.), The Handbook of Research on Learning and Instruction. Routledge Press.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 829-846.

Kitcher, P. (1981). Explanatory Unification. Philosophy of Science, 48, 507-31.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? Journal of Experimental Child Psychology, 103, 386–394.

Legare, C.H., Gelman, S.A., & Wellman, H.M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development, 81*, 92-944.

Malle, B. F. (2004). How the mind explains behavior: Folk explanations, meaning, and social interaction. The MIT Press.

McGarty, C., Yzerbyt, V., & Spears, R. (2002). Stereotypes as explanations: The formation of meaningful beliefs about social groups. Cambridge University Press.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine learning*, *1*(1), 47–80.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. Psychological Review, 92, 289-316.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Merrill Upper Saddle River, NJ.

Renkl, A. (1997). Learning from Worked-Out Examples: A Study on Individual Differences. *Cognitive Science*, *21*(1), 1–29.

Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology*, *59*(2), 229–241.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge University.

Williams, J. J., & Lombrozo, T. (2010a). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science, 34,* 776-806.

Williams, J.J., & Lombrozo, T. (2010b). Explanation constrains learning, and prior knowledge constrains explanation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2912-2917). Austin, TX: Cognitive Science Society.

Williams, J. J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Wilson, T. D., & LaFleur, S. J. (1995). Knowing What You'll Do: Effects of Analyzing Reasons on Self-Prediction. *Journal of Personality and Social Psychology*, *68*(1), 21–35.