Explanation Constrains Learning, and Prior Knowledge Constrains Explanation

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

Abstract

A great deal of research has demonstrated that learning is influenced by the learner's prior background knowledge (e.g. Murphy, 2002; Keil, 1990), but little is known about the processes by which prior knowledge is deployed. We explore the role of explanation in deploying prior knowledge by examining the joint effects of eliciting explanations and providing prior knowledge in a task where each should aid learning. Three hypotheses are considered: that explanation and prior knowledge have independent and additive effects on learning, that their joint effects on learning are subadditive, and that their effects are superadditive. A category learning experiment finds evidence for a superadditive effect: explaining drives the discovery of regularities, while prior knowledge constrains which regularities learners discover. This is consistent with an account of explanation's effects on learning proposed in Williams & Lombrozo (in press).

Keywords: explanation; self-explanation; learning; prior knowledge; constraints; generalization; category learning

What processes underlie the critical capacity to acquire information and generalize to future situations? The topic of learning is one with a long history in cognitive science and development, and with important practical applications to education. While much research in cognitive science has focused on mechanisms that are independent of the specific knowledge people possess about a domain, studies have repeatedly and reliably demonstrated that prior background knowledge has profound effects on learning. This work suggests that characterizing how prior knowledge influences learning is a key issue for theories of learning.

Effects of prior knowledge have been particularly well characterized in the context of category learning. Prior knowledge that relates the features of a category allows learners to discover an underlying thematic pattern and learn the category more quickly (e.g., Murphy & Allopenna, 1994), and prior knowledge can also influence the construction of features in a way that supports classification (Wisniewski & Medin, 1994). Most broadly, prior knowledge has been seen as helpful because it exerts constraints on the process of knowledge acquisition (Keil, 1990), such as reducing the set of hypotheses learners entertain (Tenenbaum, Griffiths & Kemp, 2006). Most proposed mechanisms for category learning – such as encoding of exemplars, prototype formation, and other associative learning mechanisms – do not capture effects of prior knowledge (see Murphy, 2002), although more recent computational models attempt to incorporate such effects (e.g., Rehder & Murphy, 2003; Tenenbaum et al, 2006).

One possibility is that generating *explanations* plays a role in the effects of prior knowledge on learning. In this paper we consider the relationship between eliciting explanations and effects of prior knowledge. Engaging in explanation during study has been shown to promote learning and generalization in a range of knowledge-rich domains, for both adults (e.g. Chi, et al, 1994) and young children (for a review see Wellman & Liu, 2006). The process of "self-explaining" may be effective in part because explaining integrates new information with prior knowledge (Chi et al, 1994).

Previous work on eliciting explanations has considered the role of prior knowledge in mediating learning gains, but with mixed results. Some studies find that eliciting explanations has the greatest benefit for learners with low levels of prior domain-knowledge (e.g., Renkl et al., 1998), and that self-explanation training may be more useful for learners with low domain knowledge (McNamara, 2004). Other studies have not found a relationship between pre-test performance and the magnitude of post-test gains (e.g. Chi & VanLehn, 1991; Chi et al., 1994; Rittle-Johnson, 2006), although there is suggestive evidence that learners with more background produce higher-quality self-explanations (Renkl, 1997; Best, Ozuru, & McNamara, 2004).

Williams and Lombrozo (in press) propose a *subsumptive constraints* account of the role of explanation in learning that suggests how explanation and prior knowledge might interact to guide learning. The subsumptive constraints account is inspired by theories of explanation in philosophy which propose that explanations show how what is being explained is an instance of (subsumed by) a general pattern. If the explanations learners generate must satisfy this constraint, then attempting to explain should drive learners to discover regularities and underlying principles that are present in the material being explained. In support of this proposal, Williams and Lombrozo (in press) found that participants who explained items' category membership were more likely to discover a subtle regularity underlying category membership than participants who described category items, thought aloud, or engaged in free study.

The subsumptive constraints account suggests two ways in which explanation and prior knowledge could interact. First, explanations could determine *which prior knowledge* is deployed. According to the subsumptive constraints account, learners should invoke beliefs that demonstrate how what is being explained can be subsumed under general patterns. Second, the account suggests that prior knowledge could provide a source of constraint on *which subsuming generalizations* are considered explanatory. Consider the task of learning about the categories "psychology lecturer" and "psychology student" from the limited observation of a single lecture. The underlying bases for the categories could be that a psychology student is seated while a psychology lecturer is standing, but this generalization seems like an implausible basis – and a poor explanation – for category membership. Distinguishing law-like generalizations from accidental generalizations is notoriously difficult (for discussion in philosophy see Caroll, 2008; and in psychology, Kalish, 2002), but prior knowledge may provide one source of constraint on which patterns are seen as explanatory, therefore determining which patterns participants are more likely to discover and employ in seeking explanations.

To investigate the relationship between explanation and prior knowledge, we restrict our focus to cases where explanation and prior knowledge would be expected to help learning, and consider whether their joint effects on learning are independent and additive, subadditive (less than the sum of their independent effects), or superadditive (greater than the sum of their independent effects).[1]

The proposed experiment uses a category-learning task in which there are patterns underlying category membership, and an explanation manipulation (explain vs. free study) is crossed with a prior knowledge manipulation (knowledge relevant to an underlying pattern is provided vs. no additional knowledge). The experiment aims to discriminate three alternative hypotheses about the joint effects of explanation and prior knowledge on learning.

One possibility is that explanation and prior knowledge have independent and additive effects. This hypothesis is a sensible default in the absence of evidence that eliciting explanations and prior knowledge interact, and no specific accounts have been proposed as to how prior knowledge might be deployed through explaining. Independent effects of explanations and prior knowledge would be likely if explaining helps learning through mechanisms that do not interact with those by which prior knowledge plays a role. For example, explaining might increase attention and motivation, while prior knowledge might independently constrain the hypotheses under consideration.

A second possibility is that prior knowledge and explanation have subadditive benefits. This could occur if the effects of explanation and prior knowledge are achieved through common mechanisms. For example, prompts to explain and the provision of prior knowledge may both guide learners to seek meaningful regularities in category structure. Explaining when prior knowledge is already available may therefore have little benefit above simply possessing prior knowledge.

A final possibility is a superadditive effect of explanation and prior knowledge, such that explanation and prior knowledge interact in a way that produces a learning benefit that exceeds either of their independent effects. This could occur if explanations deploy prior knowledge that might otherwise be inert, or if prior knowledge influences the generation of explanations in a way that fosters more effective learning. The subsumptive constraints account suggests one way this might work: attempting to generate explanations (e.g. for category membership) could invoke prior beliefs in order to supply candidate subsuming patterns, and prior beliefs could simultaneously constrain which candidate subsuming regularities are deemed explanatory.

Experiment

There are many ways that prior knowledge could impact learning, and accordingly a multitude of ways in which prior knowledge could be manipulated. In this experiment, we provide category labels intended to activate prior knowledge relevant to which features might underlie membership.

We used eight category items, shown in Figure 1. There were two rules that could be used to categorize: an *antenna rule* (shorter left vs shorter right antenna) and a *foot rule* (pointy vs flat feet). The *prior knowledge* variable was operationalized by providing uninformative category labels that were neutral with respect to the two rules (*low* prior knowledge condition: items labeled as Glorp and Drent robots) versus labels that could be related to the foot rule (*high* prior knowledge condition: labeled as Outdoor and Indoor robots). The motivation for these rules was that participants' knowledge might account for Outdoor robots having pointy fleet and Indoor robots having flat feet, but not for why Outdoor or Indoor robots would have shorter left or right antennae.[2]

While all participants were informed that they would later be tested on their ability to categorize robots, those in the *explain* condition were prompted to explain the category membership of the Glorp and Drent (or Indoor & Outdoor) robots, while those in the *free study* condition were allowed to study the robots without specific prompts, yielding a *task* variable with two levels (explain vs. free study).

The two (*Task*: Explanation vs. Free Study) x two (*Prior knowledge*: Low vs. High) design therefore allowed for a test of whether the joint effect of explanation and prior knowledge on learning a basis for categorization is independent and additive, subadditive, or superadditive.

[1] Whether explanation and prior knowledge help or hurt learning depends on the nature of what is being learned. Prior beliefs about a domain may be incorrect, or explaining may drive learners to unreliable patterns (Williams & Lombrozo, in press; Williams, Lombrozo, & Rehder, in press). In this paper we do not aim to investigate interactions of explanation and prior knowledge in settings where either will individually impair learning. In many real-world cases and educational contexts, both explaining and prior knowledge would be expected to benefit learning – for example, if there are regularities to discover and prior knowledge is correct – and this is the kind of setting we explore.

[2] Participants could have drawn on prior knowledge to explain why antenna length was related to being Outdoor/Indoor, or have had beliefs that conflicted with, for example, Outdoor robots having pointy feet, but the significant difference between conditions suggests this was not true for the majority of participants.

Participants

Two hundred and forty (60 in each condition) UC Berkeley students participated for course credit or monetary reimbursement (161 in the lab, 79 online).

Materials

The task involved *study items*, *test items*, and *transfer items*.

Study items. There were two categories of alien robots; the image participants saw in the *high prior knowledge* condition is displayed in Figure 1. The category labels were chosen based on whether the condition was *low* or *high* prior knowledge: the robots were labeled as *Glorps* and *Drents i*n the low prior knowledge condition, and as *Indoor* and *Outdoor* robots in the high prior knowledge condition.

Each robot was composed of six elements: left color (blue, green, red, yellow), right color (brown, cyan, grey, pink), body shape (square, circular), left antenna length (short, long), right antenna length (short, long), and foot shape (eight different geometric shapes). Color and body shape were uncorrelated with category membership: every right and left color occurred exactly once per category, and each category had two robots with square bodies and two with circular bodies. All four Outdoor (Glorp) robots had a shorter left antenna and all four Indoor (Drent) robots had a shorter right antenna. Although each robot had a unique geometric shape for feet, there was a subtle regularity across categories: all four Outdoor (Glorp) robots had pointy feet while all four Indoor (Drent) robots had flat feet. For simplicity, from this point on we refer to the robots in each category by their high prior knowledge label (*Outdoor/Indoor* robots).

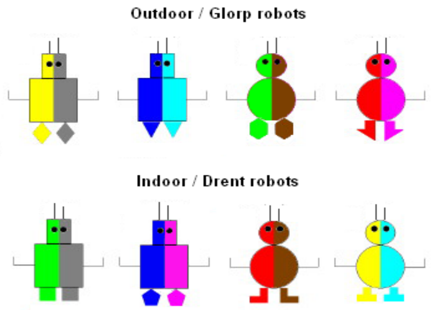

Figure 1: Study items.

This category structure supported at least three distinct bases for categorization. First, participants might not draw any generalizations about category membership, and instead categorize new items on the basis of their similarity to individual study items, where similarity is measured by tallying the number of shared features across items. We call this *item similarity*. Alternatively, participants could notice the antenna feature (Outdoor robots had shorter left antennas, Indoor robots shorter right antennas) and use it as a categorization rule: this is termed the *antenna rule*. Finally, participants could discover that although each robot had a unique geometric shape for feet, there was a subtle regularity termed the *foot rule*: Outdoor robots had pointy feet and Indoor robots had flat feet.

Test probe items. Three types of test item were constructed by taking novel combinations of the features used for the study items. Each type yielded a categorization judgment (of Outdoor/Indoor) that was diagnostic of one basis for categorization (item similarity, antenna rule, foot rule), by pitting that basis for categorization against the other two. For example, categorizing a yellow/gray robot with a shorter right antenna and pointy feet as an Indoor robot would suggest a participant relied on the antenna rule. We call these item similarity probes (three items), antenna rule probes (three items), and foot rule probes (four items). There was one extra item for which all three bases gave the same response.

Transfer Items. These four items used completely novel foot shapes to distinguish participants who genuinely drew an abstract generalization concerning “pointy” versus “flat” feet from those who simply recognized the importance of particular foot shapes. For each item, the foot rule was pitted against item similarity and the antenna rule.

Procedure

The task involved a study phase, a categorization phase, and additional measures designed to probe what participants had learned about the categories.

Study phase. Participants were instructed that they would be looking at two types of robots on the planet Zarn: Outdoor (Glorp) and Indoor (Drent) robots, with labels chosen based on being in the *high* or *low* prior knowledge condition. They were also informed that they would later be tested on their ability to remember the robots they had seen, and their ability to decide whether robots were Outdoor (Glorp) or Indoor (Drent) robots.

After advancing the instruction screen they saw a color image displaying the eight study items in a scrambled order, with each robot numbered 1 through 8 and category membership clearly indicated for each robot (the actual image for the high prior knowledge condition is shown in Figure 1). In both conditions participants were informed that they were seeing eight robots on ZARN and that the picture would be onscreen for two minutes. Participants in the *explain* condition were told “Explain why robots 1, 2, 3 & 4 might be Outdoor (Glorp) robots, and explain why robots 5, 6, 7 & 8 might be Indoor (Drent) robots.”[3] Participants typed their explanations into a box onscreen. Those in the *free study* condition were told “Robots 1, 2, 3 & 4 are Outdoor robots, and robots 5, 6, 7 & 8 are Indoor robots.” The image was onscreen for exactly two minutes and then the screen automatically advanced.

Categorization phase. The eleven test items were presented in random order, followed by the four transfer items in random order, with participants categorizing each robot as Outdoor (Glorp) or Indoor (Drent).

[3] In all quoted prompts, the alternative labels (Glorp/Drent instead of Outdoor/Indoor) are displayed in parentheses, but only one set of labels was actually displayed.

Probability of pattern. To assess participants' belief about the presence of a defining feature or rule, they were asked: "What do you think the chances are that there is one single feature that underlies whether a robot is Outdoor (Glorp) or Indoor (Drent) - a single feature that could be used to classify ALL robots?"

Category differences. Participants were explicitly asked "Were there any noticeable differences between Outdoor (Glorp) and Indoor (Drent) robots? If you think there were, please be SPECIFIC about what you thought the differences were."

Ranking of question informativeness.[4]

Features used for categorization. Participants were asked which features they used in categorizing robots. There was a separate line to enter features of Outdoor (Glorp) robots and features of Indoor (Drent) robots. [5]

Antenna Informativeness. Participants were asked if they could tell whether a robot was Outdoor (Glorp) or Indoor (Drent) by looking at its antenna, and if they could, to state what the difference was.

Antenna classification.[4]

Explanation self-report. All participants were asked if they were trying to explain the category membership of robots while the image of all 8 robots was onscreen.

Previous exposure. Participants were asked if they had seen the robots before, or already done an experiment using the materials.[6]

Foot informativeness. Participants were asked if they could tell what category a robot belonged to by looking at its feet, and if they could, to state what the difference was.

Results

In the interests of space, we do not report all dependent measures, especially as many support the same conclusions.

Each of the three kinds of test probe items pitted one basis for categorization against the other two, so participants' patterns of categorization over the full set was used to determine whether their basis for categorization was most consistent with 'item similarity', the 'antenna rule', or the 'foot rule', with ties coded as 'other'. The proportion of participants using each basis is shown in Table 1, as a function of condition. In addition to examining the basis participants' *used*, direct measures of *antenna rule discovery* and *foot rule discovery* were also coded from participants' responses to questions about whether they could classify robots based only on antenna or feet. These generally mirrored the findings on rule use. Figure 2 shows the proportion of participants who discovered the foot and antenna rules and Figure 3 shows the proportion that discovered *a* rule (antenna or foot), as a function of condition.

A log-linear analysis on *task* (explain vs. free study), *prior knowledge* (low vs. high), and *foot rule use* (used vs. did not use foot rule, as computed from inferred basis) revealed a significant three-way interaction, $\chi2$ (1) = 7.27, p < 0.01, while that for *foot rule discovery* was marginal, $\chi2$ (1) = 3.16, p = 0.08. Explanation and prior knowledge had a joint, superadditive effect on use of the foot rule. This interaction was driven by privileged use of the foot rule by participants who explained *and* had high prior knowledge (the explain-high PK condition): the combination of explaining and relevant prior knowledge exceeded the effects of each factor on its own. In fact, in the absence of explaining (i.e., the free study conditions) prior knowledge did not have an effect on foot rule use, $\chi2$ (1) = 0.06, p = 0.81.

	Foot Rule	Antenna Rule	Item Similarity	Other
Explain- Low PK	0.32	0.60	0.05	0.03
Explain- High PK	0.67	0.25	0.06	0.02
Free Study- Low PK	0.35	0.22	0.38	0.05
Free Study- High PK	0.35	0.20	0.40	0.05

Table 1: Proportion of participants using each basis for categorization, by condition.

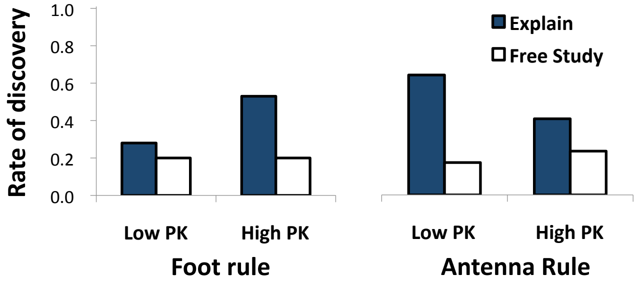

Figure 2: Proportion of participants who discovered the foot and antenna rules, by condition.

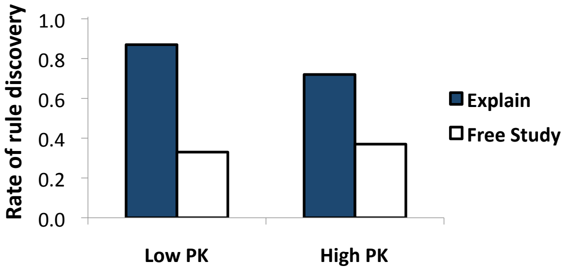

Figure 3: Proportion of participants who discovered a rule (antenna or foot), by condition.

There was also a three-way interaction between task, prior knowledge and both *antenna rule use*, $\chi2$ (1) = 5.48, p < 0.05, and *antenna rule discovery*, $\chi2$ (1) = 5.40, p < 0.05, driven by the explain-low PK condition. Overall, use of *a*

[4] This question asked participants to rank how informative different questions would be about membership, but is redundant with other reported measures and so omitted to save space.

[5] Some participants' categorization responses were reverse coded, if their explicit reports about the differences between categories or features used to categorize revealed they had reversed category labels, such as stating that outdoor robots had flat feet when in fact the opposite was true.

[6] Those who indicated previous participation were excluded.

rule (either antenna or foot) was higher for explainers (interaction between task and whether a rule was used*,* $\chi 2$ (1) = 42.76, $p < 0.001$, while reliance on *item similarity* was higher in the free study condition (interaction of task and item similarity use, $\chi 2$ (1) = 41.90, $p < 0.001$). Interestingly, overall rule discovery was actually higher in the explain-low PK than explain-high PK condition, $\chi 2$ (1) = 4.09, $p < 0.05$.

Discussion

In the context of category learning, we found that explanation and prior knowledge interacted, producing an effect on the discovery of a regularity related to prior knowledge that surpassed the independent effects of explanation or prior knowledge alone. This finding challenges the possibility that explaining and prior knowledge influence learning independently. Since a subadditive effect was not found, it also provides evidence against the hypothesis that explanation and prior knowledge draw on the same mechanisms or resources in promoting learning. The best explanation for the current findings is that explanation and prior knowledge influence learning by neither independent nor identical means, but have an interactive relationship.

This relationship can be understood in terms of the subsumptive constraints account of explanation and learning (Williams & Lombrozo, in press). If explaining exerts the constraint that learners generate explanations that show how what is being explained is subsumed by a general pattern, prior knowledge can provide constraints on which patterns support reasonable explanations. In the current experiment, explaining why items were Outdoor and Indoor robots drew on prior knowledge that constrained learners to explain membership in terms of the foot rule rather than a rule concerning antenna length. Not all subsuming patterns are equally explanatory; patterns must also make sense in light of prior knowledge.

An alternative account could instead implicate attentional mechanisms: Explaining promotes attention to items while prior knowledge exerts constraints on which item features are the focus of this attention, leading to an interactive effect on discovery of the foot rule. However, prior knowledge did not focus attention on the foot rule in the free study conditions. Moreover, Williams et al (in press) provide evidence that explaining can actually *impair* learning, suggesting that its effects go beyond increasing attention to exerting subsumptive constraints. If explaining influences attention, the evidence suggests it is not a generalized attentional boost to encode item details or monitor more information, but through constraints to attend to underlying patterns, which we would endorse as consistent with the subsumptive constraints account.

While we report a superadditive effect of explanation and prior knowledge, there are likely contexts in which different kinds of interactions would obtain. For example, it is known that the learning benefits of explanations (Williams et al, in press) and of prior knowledge (Wattenmaker et al, 1986) depend on the relationship between the constraints imposed by explanation or prior knowledge and the structure of the material being learned. If explanation exerts inappropriate constraints or prior knowledge is incorrect, their joint effects will be markedly different. Also, in cases where explanation automatically recruits prior knowledge or prior knowledge produces spontaneous explanation, their joint effect may appear to be independent or subadditive. The goal in the current work was to take a first and necessarily circumscribed step towards the ambitious goal of understanding the interactions between explanation and prior knowledge in learning.

Despite these limitations, the findings have implications for education and suggest interesting directions for applied research. Providing evidence that explaining invokes and is influenced by prior knowledge helps to explain why it has such powerful effects on learning. Explaining drives the discovery of regularities *and* guides learners to interpret what they are learning in terms of what they already know: an activity students may not engage in spontaneously even if they possess relevant prior knowledge.

If explaining promotes consistency with prior knowledge, its benefits may depend on having acquired correct and useful prior knowledge. Learning strategies that focus on acquiring background knowledge may be a necessary precursor to activities that involve explanation, and failures of explanation may suggest the need to develop background knowledge. The dangers inherent in incorrect prior knowledge are also brought into clear relief: effects of explaining may be reduced by incorrect or inappropriate prior knowledge, and may even be harmful. Examining the relationship between explanation and prior knowledge might therefore be one way to understand robust misconceptions and difficulties with conceptual change.

The current findings speak to the possibility that explanation is a mechanism by which prior knowledge is brought to bear in learning. In this experimental context, simply providing prior knowledge was insufficient to support learning: the high and low prior knowledge free study conditions did not differ in rule discovery. It may be that when learners explain and must satisfy subsumptive constraints, prior knowledge is accessed and deployed to inform which patterns are subsuming, so that explaining is a mechanism by which prior knowledge influences learning. Further research could explore what kinds of prior knowledge explaining might deploy, such as logical or causal inferences versus information stored in memory. Another issue concerns the amount of prior knowledge necessary for these interactive effects. The current experiments compared just two levels of prior knowledge, although prior knowledge spans a much broader continuum.

If explaining deploys prior knowledge in learning, it may be that spontaneously explaining category membership plays a role in knowledge effects on category learning. This possibility is bolstered by demonstrations that explaining increases use of features that are unified by prior knowledge into thematic patterns (Chin-Parker et al, 2006; Williams et al, in press). Moreover, Wisniewski & Medin (1994) reported that activating prior knowledge through meaningful category labels drove the construction of novel and abstract features. The effects they report may in fact be best understood in terms of an interaction between prior knowledge and explanations for category membership, which the subsumptive constraints account can help explain.

Explanation's effects on category learning warrant an examination of the relationship between explanation-based learning and existing models of category learning. While the subsumptive constraints account aligns naturally with rule-based models (e.g. Nosofsky et al, 1994), the reported interaction shows how both our account and rule-based models need to be extended to account for effects of prior knowledge on *which* rules count as good bases for category membership. More broadly, while representations such as exemplars play one role in learning about a category, the effect of explanation may be to construct more abstract representations that are consistent with general prior knowledge about a category, such as its origin or function.

The current work suggests a number of future directions. Do different types of prior knowledge differentially support learning, such as prior knowledge about causal mechanisms vs. functions? When does prior knowledge help because it supplies candidate patterns that can subsume observations, versus help because it informs which patterns are subsuming? Given that subsumption and consistency with prior knowledge both constrain learning, how do they trade off? These and further questions await future research.

Acknowledgments

We thank Ania Jarosewicz and Preeti Talwai for collecting data and providing other assistance and feedback on this project. This work was partially supported by the McDonnell Foundation Collaborative Initiative on Causal Learning, and JJW was supported by an NSERC post-graduate scholarship.

References

Best, R., Ozuru, Y., & McNamara, D.S. (2004). Self-explaining science texts: strategies, knowledge, and reading skill. *Proceedings of the 6th international conference on learning sciences,* 89-96.

Carroll, J. W. (2008). Laws of nature, in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, URL = <http://plato.stanford.edu/archives/fall2008/entries/laws-of-nature/>.

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

Chi, M., & VanLehn, K. (1991). The content of physics self-explanations. *Journal of the Learning Sciences, 1,* 69-105.

Chin-Parker, S., Hernandez, O., & Matens, M. (2006). Explanation in category learning. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1098–1103*)*. Mahwah, NJ: Erlbaum.

Kalish, C. W. (2002). Gold, Jade, and Emeruby: The value of naturalness for theories of concepts and categories. *Journal of Theoretical and Philosophical Psychology, 22*, 45-56.

Keil, F. C. (1990). Constraints on constraints: Surveying the epigenetic landscape. *Cognitive Science*, *14*(1), 135–168.

McNamara, D.S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*, 1-30.

Murphy, G. L. (2002). *The big book of concepts*. The MIT Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–53.

Rehder, B., & Murphy, G. L. (2003). A Knowledge-Resonance (KRES) model of category learning. Psychonomic Bulletin & Review, 10, 759-784.

Renkl, A. (1997). Learning from worked-out examples: a study of individual differences. *Cognitive Science, 21,* 1-29.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23,* 90-108.

Rittle-Johnson, B. (2006). Promoting transfer: effects of self-explanation and direct instruction. *Child Development, 77*, 1-15.

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.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309-318.

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. *Causal learning: psychology, philosophy, and computation*, 261–279.

Williams, J. J., & Lombrozo, T. (in press). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*.

Williams, J. J., Lombrozo, T., & Rehder, B. (in press). 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

Wisniewski, E. J. & Medin, D. L. (1994). On the interaction of theory and data in concept learning. Cognitive Science, 18, 221-281.