tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

Journal of Experimental Psychology: General

0096-3445/12/$12.00 DOI: 10.1037/a0030996

BRIEF REPORT

The Hazards of Explanation: Overgeneralization in the Face of Exceptions

AQ: au

Joseph J. Williams and Tania Lombrozo
University of California at Berkeley

Bob Rehder
New York University

Seeking explanations is central to science, education, and everyday thinking, and prompting learners to explain is often beneficial. Nonetheless, in 2 category learning experiments across artifact and social domains, we demonstrate that the very properties of explanation that support learning can impair learning by fostering overgeneralizations. We find that explaining encourages learners to seek broad patterns, hindering learning when patterns involve exceptions. By revealing how effects of explanation depend on the structure of what is being learned, these experiments simultaneously demonstrate the hazards of explaining and provide evidence for why explaining is so often beneficial. For better or for worse, explaining recruits the remarkable human capacity to seek underlying patterns that go beyond individual observations.

AQ: 1

Keywords: explanation, generalization, overgeneralization, anomalies

People often have the impression of understanding something better after explaining it to someone else, whether it is why a person behaved a certain way, or the solution to a math problem. In fact, prompting students to explain while they study can improve learning (e.g., Fonseca & Chi, 2010), and prompting children to explain can improve generalization to new problems (e.g., Amsterlaw & Wellman, 2006; Siegler, 2002). What is it about explaining that so effectively fosters learning? And if explaining is so beneficial, why don't people spontaneously explain more often?

Educational, developmental, and cognitive psychologists have proposed many answers to the first question. For example, explaining could increase attention, motivation, or engagement (e.g., Chi, 2009; Siegler, 2002); help learners identify and fill gaps in knowledge (e.g., Chi, 2000); or improve learning by facilitating the integration of novel information with prior beliefs (e.g., Chi, de Leeuw, Chiu, & LaVancher, 1994; Lombrozo, 2006; Wellman & Liu, 2007). Given that these processes are demanding, people could fail to explain spontaneously—even when doing so would be beneficial—to avoid what they see as inessential costs in cognitive processing (see, e.g., Fiske & Taylor, 1984; Gigerenzer, 2004).

Despite its appeal, a view of explanation as globally beneficial but inconsistently applied is at best incomplete. For one thing, such a view cannot account for previously documented hazards of explanation. Needham and Begg (1991) found that participants prompted to explain solutions to riddle-like problems outperformed those prompted to memorize the solutions when it came to analogical transfer, but performed more poorly on memory for studied examples. Kuhn and Katz (2009) found that students prompted to explain causal claims were more likely to subsequently justify claims by appeal to potential mechanisms, ignoring relevant evidence from covariation. Finally, Berthold, Röder, Knörzer, Kessler, and Renkl (2011) found that a conceptually oriented explanation prompt improved conceptual learning but impaired procedural learning. These findings suggest that explanation does not have universally beneficial effects, and additionally highlight the need to specify more precisely what explanation directs attention or processing toward and precisely why it does so.

We propose that the very properties of explanation that make it a powerful mechanism for learning under some conditions lead to systematic errors under others. Specifically, we propose that explaining privileges broad generalizations over learning about individual instances, making learners susceptible to erroneous *over*-generalizations. This idea is motivated from unification theories of explanation in philosophy, which propose that a good explanation

Joseph J. Williams and Tania Lombrozo, Department of Psychology, University of California at Berkeley; Bob Rehder, Department of Psychology, New York University.

This research was partially supported by National Science Foundation Grant DRL-1056712 to Tania Lombrozo. Preliminary versions of this work were presented at the annual meeting of the Cognitive Science Society in 2010 (Joseph J. Williams, Tania Lombrozo, & Bob Rehder, 2010) and 2011 (Joseph J. Williams, Tania Lombrozo, & Bob Rehder, 2011). We thank Sam Maldonado and Preeti Talwai for extensive help in designing experiments, data collection, analysis, and feedback throughout this research, as well as Dhruba Banerjee, Hava Edelstein, Jerry Ling, Tim Ko, and Ania Jarocewicz. We also thank Marcia Linn, Jack Gallant, Tom Griffiths, Michael Pacer, Anna Rafferty, Matthew Feinberg, Lauren Barth-Cohen, Kapil Amarnath, Daniel Reinholz, Caren Walker, Peter Epstein, Jamie Dallaire, Adam Anderson, Sam Maldonado, Vanessa Ing, Wendy de Heer, Janell Blunt, Joe Austerweil, David Miller, Chris Holdgraf, and Cathy Chase for comments on prior versions of the article.

Correspondence concerning this article should be addressed to Joseph J. Williams, Department of Psychology, University of California at Berkeley, 3210 Tolman Hall, Berkeley, CA 94720. E-mail: joseph_williams@berkeley.edu

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

is one that subsumes the fact or observation being explained as an instance of a broad pattern or generalization (e.g., Friedman, 1974; Kitcher, 1981, 1989). Explaining should therefore drive people to search for patterns that support satisfying explanations. In line with this prediction, we have found that explaining makes people more likely to discover patterns that account for a broad range of observations, even when alternative patterns are more salient (Williams & Lombrozo, 2010).

Here we test a novel and counterintuitive prediction of this account: that explaining can impair learning by leading to erroneous overgeneralizations when patterns involve exceptions. Such a finding would not only challenge the idea that explanation merely boosts processing or attention, but also shed light on what explanation directs effort and attention toward, and ultimately why explaining is so often beneficial for learning.

Experiment 1

Experiment 1 investigated effects of explanation in learning novel categories. Participants were prompted to explain or "think aloud" while studying 10 labeled exemplars, where the underlying category structure involved a "reliable" pattern without exceptions or a "misleading" pattern with two exceptions. The exemplars additionally involved unique features that supported perfect classification. If explaining encourages learners to discover and privilege broad patterns in the face of exceptions, then participants who explain should fare more poorly than those who think aloud when the pattern is misleading.

The inclusion of a think aloud condition with an identical learning task was crucial to discriminate effects of explanation from previously documented effects of verbalization or intentional learning, which can impair some kinds of category learning, memory, and implicit grammar acquisition (e.g., Ashby & Maddox, 2005; Love, 2002; Mathews et al., 1989; Toth, Reingold, & Jacoby, 1994). Experiment 1 can therefore isolate distinctive contributions of explaining to the specific impairment we predict: ignoring individual instances in favor of generalizations.

Method

Participants. Participants were 240 undergraduates and members of the University of California Berkeley community who participated in exchange for pay or course credit.

Materials and procedure.

Learning phase. Participants learned to classify 10 novel objects (vehicles) into two categories through repeated classification, feedback, and study (see Figure 1). Participants received instructions and then completed a study trial consisting of (a) classifying an unlabeled object as "Dax" or "Kez" based on its description (e.g., blue, lightly insulated, etc.), (b) receiving feedback on category membership, and (c) studying the labeled object. During study, participants in the explain condition were prompted by a sentence on the screen to explain why the item might be a Dax [Kez], whereas those in the think aloud condition were prompted to say out loud whatever they were

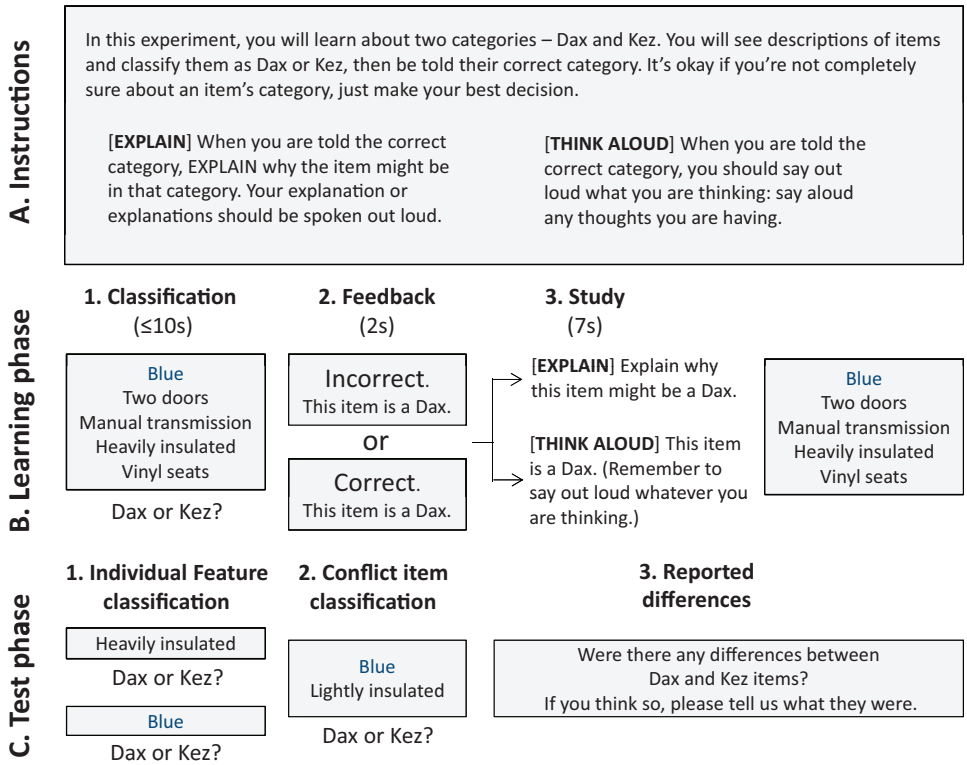

Figure 1. Schematic representation of the learning task from Experiment 1. Steps 1–3 in the learning phase were repeated for each item in each block, with participants repeating blocks until they achieved perfect classification or reached the maximum of 15.

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

Fn1 thinking.[1] This process was repeated for all 10 items to form a single "block," and participants repeatedly classified blocks until achieving perfect classification or reaching the maximum of 15 blocks.

Each object description included one unique color feature, one feature relevant to a pattern (suitability for hot vs. cold climates), and three features that were not diagnostic of category membership T1 (see Table 1; materials were adapted from Kaplan & Murphy, 2000, by the addition of the unique color features). Participants could thus classify by remembering the 10 unique features (e.g., "The red one is a Dax") or by finding a pattern in the 10 pattern-related features (e.g., "A Dax is a vehicle for warm climates"). We manipulated whether this pattern was reliable (no exceptions, as in Kaplan & Murphy, 2000) or misleading (two exceptions in 10, created by randomly switching two pattern-related features in each block). We chose two exceptions as the most minimal manipulation of pattern reliability.

Postlearning measures. To investigate effects of explanation and pattern reliability on learning about category membership via unique versus pattern-related features, we included several postlearning measures (see Figure 1).

Individual feature classification. Each unique and pattern-related feature was presented individually in a random order. Participants were asked which category an item with that feature would belong to.

Conflict classification. Participants classified 10 "conflict items" that paired a pattern-related feature and a unique feature that corresponded to opposite classifications.

Reported differences. Participants reported differences across categories, typing their responses.

Results

Learning time. A 2 (study condition: explain, think aloud) × 2 (pattern reliability: reliable, misleading) analysis of variance (ANOVA) on the mean number of blocks to reach the learning criterion revealed that participants learned more quickly when the pattern was reliable than misleading, $F(1, 236) = 44.5$, $p < .001$, F2 $\eta_p = .26$ (see Figure 2A). However, this effect was qualified by an interaction between study condition and pattern reliability, $F(1,$ Fn2 $236) = 6.3$, $p < .05$, $\eta_p = .03$.[2] We therefore evaluated effects of explanation separately for each pattern.

When the pattern was reliable, there was a trend for the explain group to learn faster than the think aloud group, $t(118) = 1.43$, $p = .16$, $d = 0.26$. When the pattern was misleading, however, participants in the explain condition were significantly slower to Fn3 reach the learning criterion, $t(118) = 2.1$, $p < .05$, $d = 0.38$.[3] In fact, 52% of participants from the misleading/explain condition never achieved perfect classification—significantly more than the 25% who failed to do so in the misleading/think aloud condition, $\chi^2(1) = 5.4$, $p < .05$.

Individual feature classification. Performance was analyzed with a mixed ANOVA with study condition (2) and pattern reliability (2) as between-subjects factors and feature type (2: unique, pattern related) as a within-subjects factor (see Figure 2B). This analysis revealed an interaction between study condition and feature type, $F(1, 236) = 12.79$, $p < .001$, $\eta_p = .05$: Explaining resulted in fewer errors on pattern-related features, $t(238) = 3.10$, $p < .01$, $d = 0.40$, but more errors on unique features, $t(238) = -2.37$, $p < .05$, $d = -0.31$. This interaction was independently significant when the pattern was reliable (and supported learning) as well as when the pattern was misleading (and hindered learning), ps $< .05$.

There were also main effects of pattern reliability, $F(1, 236) = 7.10$, $p < .01$, $\eta_p = .03$, and feature type, $F(1, 236) = 24.00$, $p < .001$, $\eta_p = .09$, which were superseded by an interaction, $F(1, 236) = 15.04$, $p < .001$, $\eta_p = .06$. Participants in the reliable pattern conditions classified pattern-related features more accurately than those in the misleading pattern conditions, $t(238) = 4.44$, $p < .001$, $d = 0.58$, with a slight trend in the opposite direction for unique features, $t(238) = -1.43$, $p = .15$, $d = -0.19$.

Conflict classification. A 2 (study condition) × 2 (pattern type) ANOVA revealed that a greater proportion of conflict items were classified in line with pattern-related features (as opposed to unique features) when the pattern was reliable rather than misleading, $F(1, 236) = 26.62$, $p < .001$, $\eta_p = .10$, and when participants explained rather than thought aloud, $F(1, 236) = 13.43$, $p < .001$, $\eta_p = .05$. The latter effect was independently significant for each pattern type (ps $< .05$; see Figure 1C).

Reported differences. Participants' typed reports about category differences were independently coded (with 84% agreement) for mention of the hot/cold pattern and/or the unique color features (see Figure 2D). Mention of the pattern was more frequent in the explain than think aloud conditions, whether the pattern was reliable, $\chi^2(1) = 4.04$, $p < .05$, or misleading, $\chi^2(1) = 9.79$, $p < .05$. However, mention of color differences was less frequent in the explain than think aloud conditions: pattern reliable, $\chi^2(1) = 4.82$, $p < .05$; misleading, $\chi^2(1) = 4.48$, $p < .05$.

Discussion

Experiment 1 confirmed our prediction that explaining can impair learning when patterns are misleading, with participants who were prompted to explain requiring more study time to reach the learning criterion than those who thought aloud. The findings additionally shed light on the basis for this impairment: Relative to thinking aloud, explaining improved learning of features that supported patterns but impaired learning of features unique to particular instances.

Experiment 2

Experiment 2 investigated effects of explaining on categorizing people's behavior, an important extension to Experiment

[1] Voice recorders were set up for both groups of participants in both Experiments 1 and 2, but unfortunately the data from all but a handful of these were lost due to a computer error.

[2] To address concerns about nonnormality, we repeated this analysis with a nonparametric test. We sorted the number of blocks to learning into five bins of three block increments and performed an ordinal regression with study condition and pattern reliability as factors. This analysis also revealed a significant interaction.

[3] To address concerns about nonnormality, all t tests reported in this experiment were checked with nonparametric Mann–Whitney U tests, which supported the same conclusions.

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

Table 1
Stimuli From the Reliable Pattern Condition in Experiment 1

Unique features	Pattern-related features	Irrelevant features		
Color	Cold/warm climate	Transmission	Seat covers	Doors
	Dax			
Blue	Drives on glaciers	Manual	Cloth	Two
Silver	Made in Norway	Automatic	Vinyl	Two
Purple	Used in mountain climbing	Automatic	Vinyl	Four
Red	Heavily insulated	Manual	Vinyl	Four
Yellow	Has treads	Manual	Cloth	Two
	Kez			
Cyan	Drives in jungles	Manual	Vinyl	Four
Magenta	Made in Africa	Manual	Cloth	Four
Olive	Has wheels	Automatic	Cloth	Two
Maroon	Lightly insulated	Manual	Vinyl	Two
Lime	Used on safaris	Automatic	Vinyl	Two

Note. In the misleading pattern conditions, a pair of pattern-related features was randomly switched across the Dax and Kez categories in each block.

1 and past work on explanation in categorizing novel objects (Williams & Lombrozo, 2010). Because behavior is arguably explained in terms of unique features (Malle, 2011; Master, Markman, & Dweck, 2012) more readily than is the category membership of objects, finding an impairment here would bolster the evidence that explaining drives people toward broad patterns at the expense of idiosyncratic particulars.

Experiment 2 also went beyond Experiment 1 in comparing effects of generating explanations during study to a control condition in which participants anticipated having to later gen-

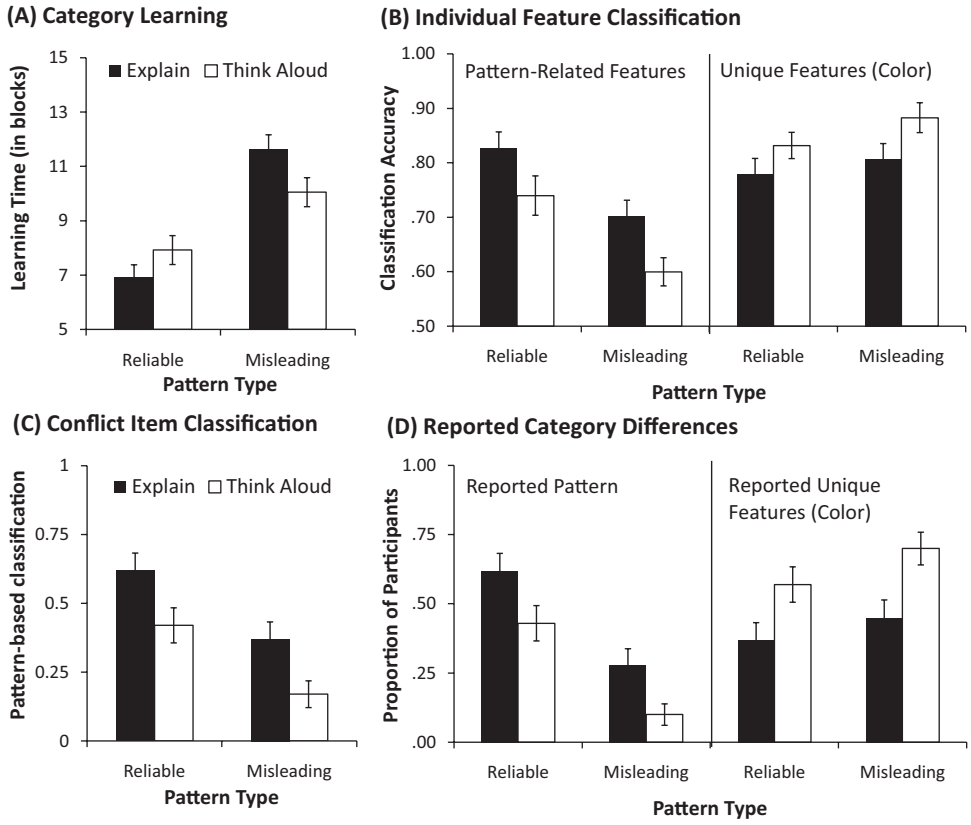

Figure 2. (A) Average learning time from Experiment 1 as a function of study condition and pattern type. (B) Average accuracy in classifying individual features after training. (C) Proportion of pattern-consistent classifications for conflict items after training. (D) Explicitly reported differences across categories. Error bars correspond to 1 standard error of the mean in each direction.

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

erate explanations. Expectations about the viability of explanations were thus matched while manipulating the degree to which participants explained during study. In addition, learning was evaluated by examining errors during a learning session of fixed time, avoiding the variability generated by learning to criterion.

Method

Participants. Participants were 164 undergraduates and members of the University of California Berkeley community who participated in exchange for pay or course credit.

Materials and procedure.

Learning phase. Participants were instructed to learn whether each of 10 hypothetical individuals rarely or frequently donated to charities (see Figure 3). Explain participants were told to explain each individual's behavior during study, whereas control participants were told that they would explain each individual's behavior at a later point. Participants then read a description of each individual and had 10 s to judge whether the individual rarely or frequently donated before receiving the correct answer and studying the individual for another 10 s. This process was repeated five times for each of the 10 individuals, with learning assessed by the proportion of errors over these trials.

Each description included the person's (ostensible) picture, name, and age, as well as a personality descriptor (e.g., friendly) and two additional features, college major and geographic location, which were not correlated with behavior (see Table 2). The unique picture and name supported perfect predictions concerning behavior (e.g., "Laura rarely donates"), and the age and personality features conformed to patterns that correlated with behavior (e.g., "Younger people rarely [frequently] donate," "People with extraverted traits rarely [frequently] donate").[4] Using two patterns (age and personality) created a more complex learning context than in Experiment 1, which involved a single pattern. In the reliable patterns conditions, the patterns involving age and personality correlated perfectly with behavior. In the misleading patterns conditions, each pattern involved two exceptions.

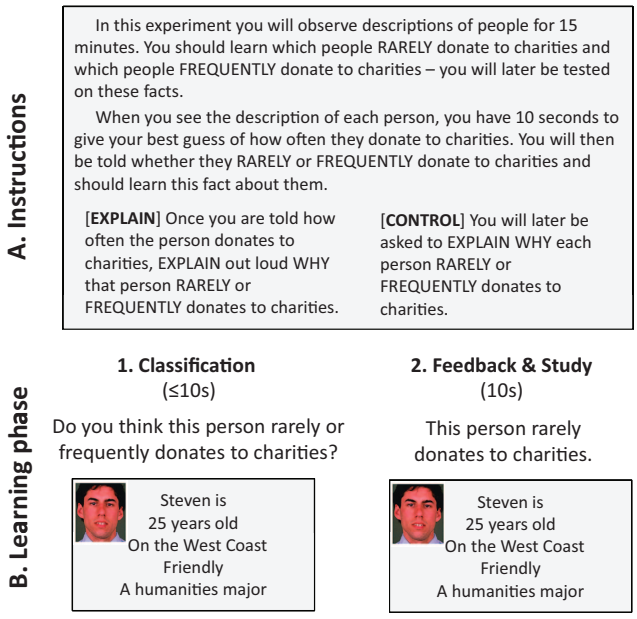

Figure 3. Schematic representation of the learning task from Experiment 2. Steps 1–2 were repeated for each of the 10 items in each block, and each block was repeated five times.

Results

Learning error was computed as the proportion of errors made during the learning phase, and was analyzed with an ANOVA that included study condition (2: explain, control) and pattern reliability (2: reliable, misleading) as between-subjects factors. This analysis revealed more errors in the misleading patterns conditions than the reliable patterns conditions, $F(1, 160) = 32.41$, $p < .001$, $\eta_p = .25$, as well as the critical interaction between study condition and pattern reliability, $F(1, 160) = 4.63$, $p < .05$, $\eta_p = .03$ (see Figure 4A). Explaining had no significant effect on errors when the patterns were reliable, $t(75) = 0.35$, $p = .73$, $d = 0.08$, but increased errors when the patterns were misleading, $t(85) = 2.50$, $p < .05$, $d = 0.54$.

To examine whether errors in the explain condition resulted from overgeneralizing patterns to exceptions, we performed an additional ANOVA on errors in the misleading patterns condition that considered two additional within-subjects factors: item type, where items were identified as pattern-consistent if the individual's behavior conformed to the age and personality patterns, and as exceptions if not (see Figure 4B), and learning block (see Figure 4C). The inclusion of item type and learning block did not alter the key finding that explaining increased errors relative to control, $F(1, 83) = 5.12$, $p < .05$, $\eta_p = .06$. However, there was additionally a main effect of learning block, $F(4, 80) = 37.42$, $p < .001$, with errors decreasing over time; a main effect of item type, $F(1, 83) = 16.45$, $p < .001$, $\eta_p = .32$, with more errors for exceptions; and a (marginal) interaction between study condition and item type, $F(1, 83) = 3.82$, $p = .054$, $\eta_p = .04$: Participants in the explain condition made significantly more errors than those in the control condition for exception items, $F(1, 83) = 5.12$, $p < .05$, $\eta_p = .06$ (this difference was significant by Block 2, $p < .05$, and was still significant in Block 5, $p < .05$), with a similar but nonsignificant trend for pattern-consistent items, $F(1, 83) = 1.44$, $p = .23$, $\eta_p = .02$. We speculate that participants prompted to explain did not outperform control participants on pattern-consistent items because they may have switched between patterns (age vs. personality) in the face of exceptions instead of abandoning patterns altogether.

Discussion

Experiment 2 replicated the key prediction that explaining can impair learning when patterns are misleading, extending the finding from Experiment 1 to a novel domain, a new measure of learning, and a different control condition. The analysis of error types suggests that explaining impaired learning through rapidly formed overgeneralizations that persisted in the face of repeated counterevidence.

[4] The direction of the association between age, personality, and behavior was counterbalanced across conditions, but this manipulation had no effect and is not discussed further.

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

Table 2
Stimuli From Experiment 2

Unique features		Pattern-related features		Irrelevant features	
Picture	Name	Age	Personality	A graduate of a	Living on the
Rarely donates to charities (frequently donates to charities)					
	Laura	30	Dominating	Science major	East Coast
	Steven	25	Friendly	Humanities major	West Coast
	Jessica	32	Boastful	Science major	West Coast
	Janet	26	Self-assured	Science major	East Coast
	Kevin	23	Energetic	Humanities major	West Coast
Frequently donates to charities (rarely donates to charities)					
	Joseph	37	Cautious	Science major	East Coast
	Josh	47	Discreet	Humanities major	West Coast
	Karen	39	Studious	Science major	West Coast
	Anna	45	Self-conscious	Humanities major	West Coast
	Sarah	42	Quiet	Science major	East Coast

Note. In the misleading patterns conditions, the age and picture of the fourth and eighth individuals were switched, as were the personality features of the fifth and 10th individuals.

General Discussion

Our findings reveal the double-edged nature of explanation. Although explaining can be beneficial, it can also make learners vulnerable to overgeneralizations when categorizing artifacts (Experiment 1) or behaviors (Experiment 2), and for measures of both learning speed (Experiment 1) and learning accuracy (Experiment 2). In Experiment 1, explainers more accurately learned pattern-related features and less accurately learned unique features than those who thought aloud. In Experiment 2, explainers were especially inaccurate when it came to categorizing exceptions, and this effect emerged early in learning and persisted throughout training. These findings suggest that explainers focused on features that supported patterns at the expense of idiosyncratic information about individual items, and that they perseverated in seeking or applying broad patterns despite evidence against their generality.

Just as visual illusions shed light on the mechanisms by which visual perception is so often accurate, our findings shed light on why explanation is so often beneficial. In particular, our findings suggest that explaining "why?" recruits evaluative criteria for what constitutes a good explanation, directing learners to seek broad patterns that can accommodate what is being explained (see also Lombrozo, 2012; Williams & Lombrozo, 2010). Explaining can therefore support the remarkable human capacity to discover patterns and construct generalizations from sparse observations, but this capacity has associated risks: disregarded exceptions and overgeneralization.

Our findings also rule out the idea that explaining simply increases attention or processing to yield global improvements, and can help make sense of the seemingly disjointed set of negative effects of explanation reviewed in the introduction. For example, relative to instructions to remember examples, explanation promotes analogical transfer at the expense of memory (Needham & Begg, 1991), and this effect makes sense if explaining highlights broad patterns over individual exam-

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

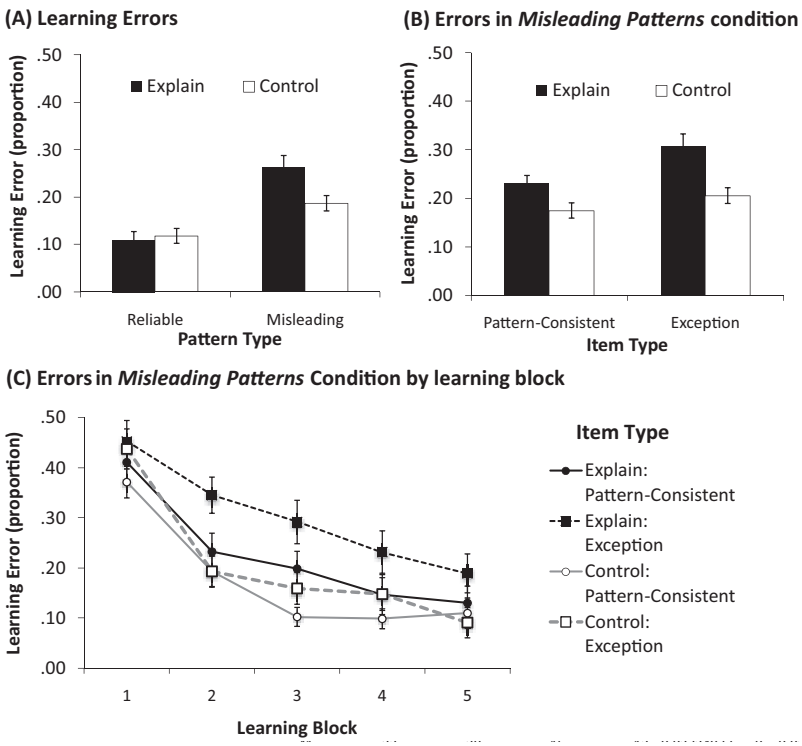

Figure 4. (A) Average proportion of errors during learning for Experiment 2 as a function of study condition and pattern type. (B) Errors during learning for the misleading patterns condition as a function of study condition and item type. (C) Errors from the misleading patterns condition for pattern-consistent and exception items as a function of learning block number. Error bars correspond to 1 standard error of the mean in each direction.

ples. Similarly, explaining could encourage learners to invoke causal mechanisms over particular observations in justifying causal claims (Kuhn & Katz, 2009) because doing so relates what is being explained to broader regularities. Finally, explaining could encourage learners to draw generalizations over conceptual rather than procedural aspects of a domain when the content of what they are explaining is conceptual (Berthold et al., 2011).

Our finding that explanation can promote erroneous overgeneralizations goes beyond these previous results in isolating a prediction of our account, but additionally suggests that effects of explanation are not a simple consequence of directing limited attention or processing to the target of a particular study prompt (e.g., memory, causal mechanisms, or conceptual knowledge) at the expense of alternatives (e.g., relational structure, covariation, or procedural knowledge). Nothing about the content of our explanation prompts highlighted broad patterns over properties of individuals. If anything, a prompt to explain the category membership or behavior of a specific instance could have directed attention or processing to the particulars of that instance. Instead, the results support our proposal that explaining "why?" is inherently linked to patterns, with the content of the question potentially affecting the nature of the patterns considered (e.g., whether they involve conceptual or procedural regularities; see also Williams & Lombrozo, 2013).

Our account also predicts conditions under which explanation should have minimal effects. For example, explanation prompts could have no effect when learners can successfully identify and apply broad generalizations without explaining (e.g., because they receive rich and effective instruction) or when learners lack the requisite knowledge to generate reasonable hypotheses about underlying patterns (for related discussion, see Matthews & Rittle-Johnson, 2009; Rittle-Johnson, 2006).

In their search for patterns, participants who explained could have recruited a host of well-documented processes, including verbal reasoning (e.g., Meissner & Memon, 2002; Schooler, 2002) and analogical comparison (Gentner, 2010), or triggered a more explicit and deliberative (Mathews et al., 1989), analytic and rule-based (Ashby & Maddox, 2005; Shanks & St. John, 1994), or intentional (Dienes et al., 1991; Reber, 1989) mode of learning. Given our closely matched study conditions, it is likely that these processes and strategies were also triggered in participants in control conditions, if to a lesser degree. Our account is not in conflict with these views concerning cognitive mechanisms or architectures, but instead suggests that if explanation did recruit these mechanisms or systems, it was in the service of finding broad patterns, and it is this feature of explanation that explains our results.

Although we specifically designed conditions conducive to an explanation impairment using feature lists in a laboratory context, a range of real-world situations involve similarly sparse observations and unreliable patterns. For example, explaining a single (potentially unrepresentative) observation can generate the kinds of beliefs that underlie stereotypes (Risen, Gilovich, & Dunning,

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

2007), and trying to explain chance events could reinforce superstitious beliefs or conspiracy theories. Future work can investigate these hazards of explanation, and additionally aim to reconcile them with cases where explaining exceptions is useful in discovering novel regularities, as when anomalies presage scientific theory change (Chinn & Brewer, 1993; Kuhn, 1962) or guide children's causal learning (Legare, Gelman, & Wellman, 2010).

Finally, we should note that ignoring exceptions may sometimes help learning. For example, when there is substantial variability in observations, exceptions could erroneously lead learners away from noisy but reliable patterns. Moreover, there are many settings, such as those in mathematics and science education, where explaining has proven beneficial, for which the benefits of erring on the side of overgeneralization can outweigh minor costs. Providing a unified account of the positive and negative effects of explanation can not only help avoid the hazards of explanation, but also maximize and extend its benefits, whether in everyday, educational, or scientific contexts.

References

Amsterlaw, J. A., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development, 7,* 139–172. doi:10.1207/s15327647jcd0702_1

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178. doi:10.1146/annurev.psych.56.091103.070217

Berthold, K., Röder, H., Knörzer, D., Kessler, W., & Renkl, A. (2011). The double-edged effects of explanation prompts. *Computers in Human Behavior, 27,* 69–75. doi:10.1016/j.chb.2010.05.025

Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology: Vol. 5. Education design and cognitive sciences* (pp. 161–238). Mahwah, NJ: Erlbaum.

Chi, M. T. H. (2009). Active–constructive–interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1,* 73–105. doi:10.1111/j.1756-8765.2008.01005.x

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

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63,* 1–49. doi:10.3102/00346543063001001

Dienes, Z., Broadbent, D., & Berry, D. C. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 875–887. doi:10.1037/0278-7393.17.5.875

Fiske, S. T., & Taylor, S. E. (1984). *Social cognition*. New York, NY: Random House.

Fonseca, B., & Chi, M. T. H. (2010). The self-explanation effect: A constructive learning activity. In R. Mayer & P. Alexander (Eds.), *The handbook of research on learning and instruction* (pp. 270–321). New York, NY: Routledge.

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

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34,* 752–775. doi:10.1111/j.1551-6709.2010.01114.x

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62–88). Malden, MA: Blackwell. doi:10.1002/9780470752937.ch4

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 829–846. doi:10.1037/0278-7393.26.4.829

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48,* 507–531. doi:10.1086/289019

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. C. Salmon (Eds.), *Minnesota Studies in the Philosophy of Science, Vol. 13. Scientific explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology, 103,* 386–394. doi:10.1016/j.jecp.2009.03.003

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development, 81,* 929–944. doi:10.1111/j.1467-8624.2010.01443.x

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

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

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9,* 829–835. doi:10.3758/BF03196342

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In D. Chadee (Ed.), *Theories in social psychology* (pp. 72–95). Malden, MA: Wiley-Blackwell.

Master, A., Markman, E. M., & Dweck, C. S. (2012). Thinking in categories or along a continuum: Consequences for children's social judgments. *Child Development, 83,* 1145–1163. doi:10.1111/j.1467-8624.2012.01774.x

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1083–1100. doi:10.1037/0278-7393.15.6.1083

Matthews, P. G., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104,* 1–21. doi:10.1016/j.jecp.2008.08.004

Meissner, C. A., & Memon, A. (2002). Verbal overshadowing: A special issue exploring theoretical and applied issues. *Applied Cognitive Psychology, 16,* 869–872. doi:10.1002/acp.928

Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory & Cognition, 19,* 543–557. doi:10.3758/BF03197150

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219–235. doi:10.1037/0096-3445.118.3.219

Risen, J. L., Gilovich, T., & Dunning, D. (2007). One-shot illusory correlations and stereotype formation. *Personality and Social Psychology Bulletin, 33,* 1492–1502. doi:10.1177/0146167207305862

Rittle-Johnson, B. (2006). Promoting transfer: The effects of direct instruction and self-explanation. *Child Development, 77,* 1–15. doi:10.1111/j.1467-8624.2006.00852.x

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology, 16,* 989–997. doi:10.1002/acp.930

tapraid5/zfr-xge/zfr-xge/zfr00113/zfr2387d13z | xppws | S=1 | 12/5/12 | 19:43 | Art: 2012-0649

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367–395. doi:10.1017/S0140525X00035032

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, NY: Cambridge University Press. doi:10.1017/CBO9780511489709.002

Toth, J. P., Reingold, E. M., & Jacoby, L. L. (1994). Toward a redefinition of implicit memory: Process dissociations following elaborative processing and self-generation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 290–303. doi:10.1037/0278-7393.20.2.290

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 261–279). Oxford, England: Oxford University Press. doi:10.1093/acprof:oso/9780195176803.003.0017

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

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

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* (pp. 2906–2911). Austin, TX: Cognitive Science Society.

Williams, J. J., Lombrozo, T., & Rehder, B. (2011). Explaining drives the discovery of real and illusory patterns. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1352–1357). Austin, TX: Cognitive Science Society.

Received March 20, 2012
Revision received October 19, 2012
Accepted October 24, 2012 ■

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES 1

AQau—Please confirm the given-names and surnames are identified properly by the colors.
■= Given-Name, ■= Surname
The colors are for proofing purposes only. The colors will not appear online or in print.

AQ1—Author: Please provide from three to five keywords or phrases.