More than just new evidence: How category learning fosters belief revision

Micah B. Goldwater (micah.goldwater@sydney.edu.au)
Monica Bollen (mbol3481@uni.sydney.edu.au)
& Josue Giron (josue.giron@sydney.edu.au)
University of Sydney, School of Psychology, Brennan MacCallum Building A18
Camperdown, NSW 2006 Australia

Abstract

Causal judgments are stubborn. If people learn about two correlated variables B and C, and judge that B causes C, they typically stick to that judgment even when contradictory evidence comes to light. One form of contradictory evidence is that a third variable A causes both B and C, explaining the correlation. This paper extends prior work showing that simply presenting statistical evidence that A is the *common cause* of both B and C does not lead to belief change about B. However, if first subjects learn to categorize phenomena by their underlying causal relationships (i.e., as exemplars of a common cause category), then they can transfer their category knowledge to properly interpret the evidence. They recognize that A is the common cause of B and C and revise their belief about B. These results suggest that teaching abstract causal categories is a promising strategy to help revise false beliefs.

Keywords: belief revision; categories and concepts; analogy; causal learning

Introduction

False beliefs are pervasive, leading to many harmful decisions. For example, Steve Jobs famously delayed surgery for his cancer, and instead chose naturopathic treatments that have no supporting scientific evidence for their efficacy. While it is hard to be certain, there is evidence that decision led to his death (e.g., see Walton, 2011 https://tinyurl.com/ybbc3jnk). A more common example is attributing the fast recovery from a cold to an herbal supplement when in reality recovery would have been just as fast without taking it. Large sample public surveys and laboratory experiments reinforce the notion that false beliefs are at once easily formed, and resistant to change (see Lewandowsky et al., 2012; and Hornsey & Fielding, 2017 for review).

Taylor and Ahn (2012) developed a laboratory paradigm to investigate the mechanisms underlying the real-world problem of stubborn false beliefs. In their paradigm, they demonstrated the phenomenon of "causal imprinting" wherein once a causal judgment was made, it was resistant to change despite new contradictory evidence. They used an observational learning method wherein subjects considered a series of patient medical files to determine if B (the "Burlosis condition") caused C (the "Caprix condition"). In the twenty medical files, the two conditions were highly correlated (see Figure 1, top), and so most subjects endorsed that yes, B does cause C. Then the subjects were told that a new potential cause, A (the "Ablique virus") had been discovered and they should look over the medical files again, now with information about whether or not patients had that virus. The statistics now suggested that the correlation between B and C was due entirely to A; that is when A was present, B and C were also likely to be, and when A was not present, B and C were likely not to be (see Figure 1, bottom). That is, A was the *common cause* of both B and C. Although subjects saw the causal power of A, this did not change their view about B. They thought A could cause B and C, but also that B caused C. On the other hand, if subjects saw the patient files that included A from the beginning, they readily inferred that A caused B and C, and that B did not cause C. Across four experiments Taylor and Ahn ruled out several simple hypotheses about why people failed to revise their beliefs about B, for example not properly recognizing the relationship between the two sets of observations. Rejecting these alternative explanations, they concluded the causal power of B was imprinted on the subjects' minds. What could help change this belief?

BC block contingency

	C	**-C**
B	8	2
-B	2	8

ABC block contingency

	A, C	-A, C	A, -C	-A, -C
B	8	0	1	1
-B	1	1	0	8

Figure 1: Co-occurrence statistics across twenty observational learning trials. Each number refers to the number of trials where some combination of variables was present or not. Negatives refer to "not present." In the contingency table with all three variables, the column labels refer to combinations of A's and C's presence. This shows how the correlation above is explained by A.

Rottman, Gentner, and Goldwater (2012) showed that science novices are not particularly perceptive of when phenomena share underlying causal relations. The novices were given descriptions of phenomena from a number of domains to sort into categories (e.g., economics, biology), and decided to sort them based on domain, regardless of any differences in their underlying causal structure. On the other hand, science experts saw past the domains, and sorted phenomena via their causal structure. For example, they sorted descriptions of biological and economic phenomena together that each showed a common cause structure

wherein A (an allergic reaction; high unemployment, respectively) caused both B (rash; increased crime) and C (coughing; lower GDP) independently. Another causal category that guided science experts' sorting was a causal chain wherein A (increased oil prices) caused B (increased transportations costs) which in turn directly caused C (increased price of consumer goods). Building on this work, Goldwater and Gentner (2015) taught novices the causal system categories by scaffolding their comparison of phenomena from different domains that shared causal structure. After the comparisons, subjects sorted novel phenomena descriptions via their causal structures in a manner similar to the science experts in Rottman et al. (2012).

The current work examined whether learning the causal system categories as in Goldwater and Gentner (2015) can aid subjects in either preventing or undoing causal imprinting. If subjects had a more general understanding of common causes, they could potentially use that understanding to recognize when they had falsely believed that B caused C, and instead attribute the cause of both B and C to A. In terms of Taylor and Ahn (2012) if they recognized that the relationships among the Ablique Virus and the Burlosis and Caprix conditions was actually an exemplar of a common cause category, they could use their knowledge of the category to help them interpret the statistical pattern in the medical files. To investigate whether this method can aid belief revision, the current work simply combined the category learning task of Goldwater and Gentner (2015) with the observational learning task of Taylor and Ahn (2012).

In Experiment 1, all subjects reviewed patient files with information about the Ablique Virus from the beginning, when prior work showed they could properly infer that A was the common cause of B and C. Before reviewing the patient files some of the subjects received the causal system category training, and some did not. Here, we predicted that the training would have no effect. Our working hypothesis was that learning the causal categories would specifically support belief revision, and when subjects learn about A, B, and C from the start, there will be no need to revise beliefs. This established a baseline for the efficacy of the category training shown in the next two experiments.

In Experiment 2, all subjects first learned about B and C, and then reviewed the files again with information about A. This was the condition from Taylor and Ahn (2012) that showed causal imprinting. Subjects received causal category training, and were given a second opportunity to consider the patient files with information about A, B, and C. Here we expected the causal system category training to reduce the size of the causal imprinting effect. In Experiment 3, we compared the causal category training to a control condition to ensure that the effects of Experiment 2 were not simply due to having an additional opportunity to consider the statistical evidence.

Because the pattern is complex, we briefly summarize the overall rationale for the series of experiments before describing them. We predicted that learning the common cause structure as a category should specifically help belief revision (as in E2 and E3), but have no effect when no revision is needed (when the statistics support a common cause inference from the start in E1). In E2, we contrasted whether category training could prevent causal imprinting with its potential to help undo imprinting. In E3, we compared how the category training helps to undo imprinting to a control condition.

Experiment 1

This experiment had a simple design of two conditions manipulated between-subjects. All subjects reviewed patient files containing information about A, B, and C from the start. Some of the subjects received the causal system category training before reviewing the patient files. Some received no category training.

Methods

Subjects. Sixty-two subjects were recruited from an introductory psychology class at the University of Sydney, and received course credit for their participation.

Materials and Procedure. Thirty-nine of the subjects (randomly assigned) started with the causal system category training task adapted from Goldwater and Gentner (2015). The training focussed on two causal systems such that any application of the category training to the observational learning phase would have to be selective of which category was relevant, and contrasting two categories improves learning of each (e.g., Rohrer & Pashler, 2010). The training task presented two examples each of common cause and causal chain systems along with clear explications of their respective causal structures. Participants were asked why examples of these systems belonged to their respective categories and to identify the elements of one example that corresponded to elements of another example of the same causal system (See Figure 2). Subsequently, participants were provided with four novel examples (two common-cause and two causal chain) and asked which category they belong to and why. Finally, participants were asked to respond to the question: "What are the key differences between common cause systems and causal chain systems?"

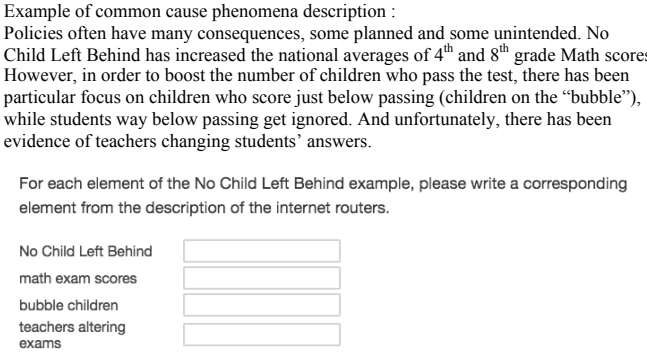

Figure 2. Example phenomena description from category training task (top), and correspondence task to scaffold a

structural comparison (bottom; see Goldwater & Gentner 2015 for complete description and evidence of effectiveness).

All subjects completed the observational learning task, which started with a practice round based on Dennis and Ahn (2001; also used by Taylor & Ahn, 2012) For space we have cut out detail here, but to summarize: Subjects were asked to determine whether ingesting an exotic plant causes people to become sick by observing several cases describing whether or not an exotic plant was ingested and whether or not that person became sick. After two blocks, one suggesting a causal relationship, and one that did not, participants were provided with a summary of the cases and an explanation demonstrating how to infer the strength of the causal relationship between events based on the co-occurrence statistics and how to score the strength of the causal relationship on a scale from 0-100. They were given four examples to calibrate their use of the scale (again based on Taylor & Ahn 2012). 0 was labelled "not a cause" and the example given was "the degree to which one rain drop causes a rise in the stock market." 30 was labelled "a weak cause" and the example was "the degree to which getting wet causes a cold." 70 was labelled "a strong cause" and the example was "the degree to which being exposed to a virus causes a cold." 100 was labelled "a very strong cause" and the example was "the degree to which rain causes the ground to be wet." Subjects could refer to these guidelines for each rating they made throughout the experiment.

Next all subjects moved onto the primary observational learning task wherein subjects were told:

"Now, imagine that you are a research assistant at the Garvan institute of medical research. Scientists have recently discovered three new medical conditions and are trying to understand if there is a relationship between them. Blood was taken from 20 patients and it was found that three events were common among the samples. These events were

1. "Has Burlosis condition" or "No Burlosis condition"
2. "Has Caprix Condition" or "No Caprix condition."
3. "Has Ablique virus" or "No Ablique virus."

…From the preliminary research scientists have discovered that the 'Caprix' condition can NOT cause the 'Burlosis' condition or the 'Ablique' virus. For this reason, your job as a research assistant will be to observe 20 patients to determine whether:

1. The 'Burlosis' condition causes the 'Caprix' condition
2. The 'Ablique' virus causes the 'Burlosis' condition
3. The 'Ablique' virus causes the 'Caprix' condition"

Then the subjects reviewed the 20 patient files at their own pace, and after completion they rated the three potential causal relationships from 0 to 100.

Results

Figure 3 shows the causal strength ratings for B causes C, and the average of A causes B and A causes C. Using the average of both A ratings allows for an easier visual comparison between A's perceived causal strength (the true cause), and B's (the false cause). The two A ratings never significantly differed from each other across the three experiments.

To simplify the statistical analyses for all three experiments, we subtract the B causal strength rating from the average A rating. A value of zero means they see A and B as equally causally strong. A positive value means A is seen as stronger than B. A negative value means B is seen as stronger than A. We note the higher the positive value, the more consistent the ratings are with a common cause relationship among the variables.

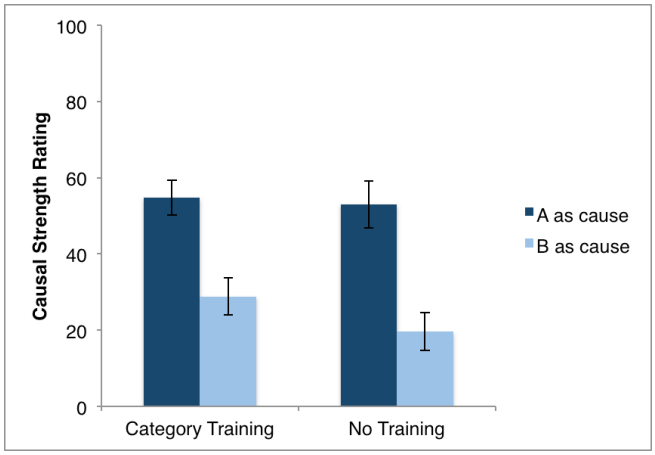

Figure 3. Mean (and standard error) causal strength ratings for Experiment 1.

There was no difference between the two conditions, as the subjects who received the category training (M = 25.97) and the subjects who did not (M = 33.50) both rated A as having greater causal strength than B, t (60) = .68, p =.50. Consistent with predictions, and replicating Taylor and Ahn (2012), when all three variables are introduced together, subjects can clearly infer that A is the common cause of B and C, and B is not a direct cause of C.

Experiment 2

The second experiment now tested whether causal system category learning could prevent or undo causal imprinting. Here, the observational learning task induced causal imprinting in the same manner as Taylor & Ahn (2012) by first having subjects review patient files with information only about B and C (the BC block), and then following up with the block with information about A, B, and C (ABC block) from Experiment 1. This experiment added a second ABC block identical to the first to give subjects a second time to reflect on the statistical pattern and potentially change their causal strength ratings. This order of BC block, then ABC block, then another ABC block was identical for all subjects.

In addition, all subjects received the causal system category learning task from Experiment 1, however the timing of the causal system category learning task was manipulated between-subjects (see Figure 4). Subjects either

received the category training right from the start, or received it in between the two ABC blocks. If there were differences between the two conditions' causal strength ratings after the first ABC block, this would suggest that the causal category training could mitigate the size of causal imprinting from the start. If the second ABC rating differed from the first, this would suggest that the causal category training could aid belief revision and undo some of the effects of causal imprinting. We predicted that the category training would elicit an effect of increasing A's causal strength rating relative to B's, but we were agnostic to an effect of when this occurred.

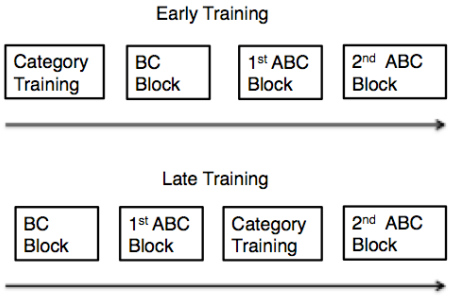

Figure 4. Task sequence for Experiment 2

Methods

Subjects. 151 subjects were recruited from an introductory psychology class at the University of Sydney, and received course credit for their participation.

Materials and Procedure. The materials and procedure were largely identical to Experiment 1. The primary difference is that at the start of the primary observational learning task, the instructions only mentioned the Burlosis and Caprix conditions. Similar to the previous instructions, the subjects were told that the earlier research ruled out that the Caprix condition could cause the Burlosis condition, but they were to consider whether Burlosis could cause Caprix. They then reviewed the twenty medical files without any information about the Ablique virues. To be consistent with Taylor and Ahn (2012), subjects did not rate the strength of B causes C after the BC block (their pilot work showed that rating B's causal strength before the ABC block has no effect on the causal imprinting shown after the ABC block).

After the BC block, the instructions told the subjects that a new discovery had been made, the Ablique Virus, and that they were to re-examine the same patient files now with information about who had or did not have the Ablique Virus. From there, the instructions proceeded identically to the ABC block from Experiment 1. Another novel part of the procedure was the second ABC block wherein subjects were instructed to consider the ABC block one more time and were then given a second opportunity to rate the causal strength of the three candidate causal relationships.

The final novel aspect of the procedure was that approximately half (n=76) of the subjects received the category training task before the observational learning task, while the other subjects (n=75) received the training in between the two ABC block (randomly assigned; Figure 4).

Results

Figure 5 shows the full pattern of causal strength ratings. Subjects viewed B's causal strength as quite close to A's on the first rating, replicating Taylor and Ahn's (2012) causal imprinting effect. The second rating however shows belief revision, with a reduction in B relative to A.

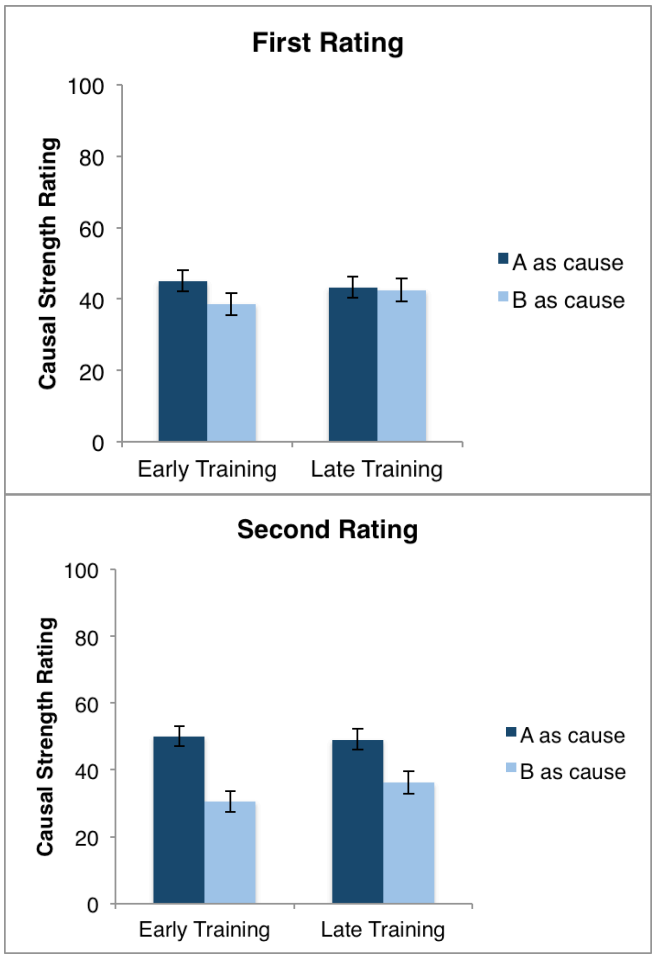

Figure 5. Mean (and standard error) causal strength ratings for Experiment 2.

For the statistical analyses we again use the summary strength score that subtracts B's causal strength rating from the average of both of A's ratings. Recall high positive scores show subjects recognized that A is the greater cause of B and C than B is the cause of C. However scores close to or below 0 show causal imprinting. They view B's causal strength as similar or greater to A's.

We analyzed the data with a 2 (Early Category Training vs. Late Category Training; between-subjects) X 2 (1st causal strength rating vs. 2nd causal strength rating; within-subjects) mixed-measures ANOVA. There was no main effect of the timing of category training, as both the first rating (Early Training, M = 6.40; Late Training, M = 0.81),

and the second rating (Early Training, M = 19.49; Late Training, M = 12.85), showed similar scores between the two conditions, $F(1,149) = 1.46$, $p = .23$, $\eta^2_p = .010$. However there was a main effect of rating, as subjects' rating at the second rating were higher in both training timing conditions than at the first, $F(1,149) = 16.58$, $p < .001$, $\eta^2_p = .100$. There was no interaction between the two variables as both the subjects in the Early Training condition (Rating Change M = 13.09) and Late Training (Rating Change M = 12.05) increased their ratings to similar degrees, $F(1,149) = 0.03$, $p = .86$, $\eta^2_p < .001$.

Experiment 3

There were several important findings from Experiment 2. First is that we replicated the causal imprinting effect from Taylor and Ahn (2012) in the first set of causal strength ratings. B was seen as equally strong as A. Second is that the second set of ratings showed belief revision wherein the ratings moved towards the pattern from Experiment 1 when A was seen as the common cause of B and C, and B was not seen as a strong cause of C. Interestingly, the timing of the category training did not seem to have an effect on this pattern. It still required two considerations of the ABC data for the subjects who had the category training from the beginning.[1]

The results of experiment 2 suggested that the category training provided a tool to foster belief revision if the subjects are given multiple chances to consider the statistical pattern. However, we cannot yet reject another plausible hypotheses- that having a second chance to consider the statistical pattern in the ABC block is sufficient to change the causal strength ratings, and that the category training had no effect. In Experiment 3, we replicate the Late Training condition of Experiment 2, but add a control task in between the first and second ABC block.

Methods

Subjects. Sixty-five subjects were recruited from Amazon's Mechanical Turk and were compensated $8 US for their participation.

Materials and Procedure. The materials and procedure for thirty-five of the subjects (randomly assigned) were that of Experiment 2's Late Training condition. The observational learning task was identical here for the Control condition, but instead of the category training between the two ABC blocks, the subjects read a passage about the logic of randomized control trials for medical research and then answered comprehension questions about that passage.

Results

Figure 6 shows the full pattern of results. Compared to Experiment 2, ratings across the board were higher, and B was actually seen as a stronger cause than A in the first rating (unlike in Experiment 2 when they were rated almost equally); this is most likely because E1 and E2 recruited undergraduates while E3 recruited from MechTurk. What is more important however are condition differences, not how the scale is used on average. Crucially, the second rating shows a reduction of B's strength relative to A's for the Late Training condition, while the Control condition shows no such reduction. Again, summary ratings scores are analyzed. Higher scores reflect a greater recognition of A as the common cause of B and C, while lower scores shows a greater effect of causal imprinting wherein B is seen as having causal strength comparable to or greater than A's.

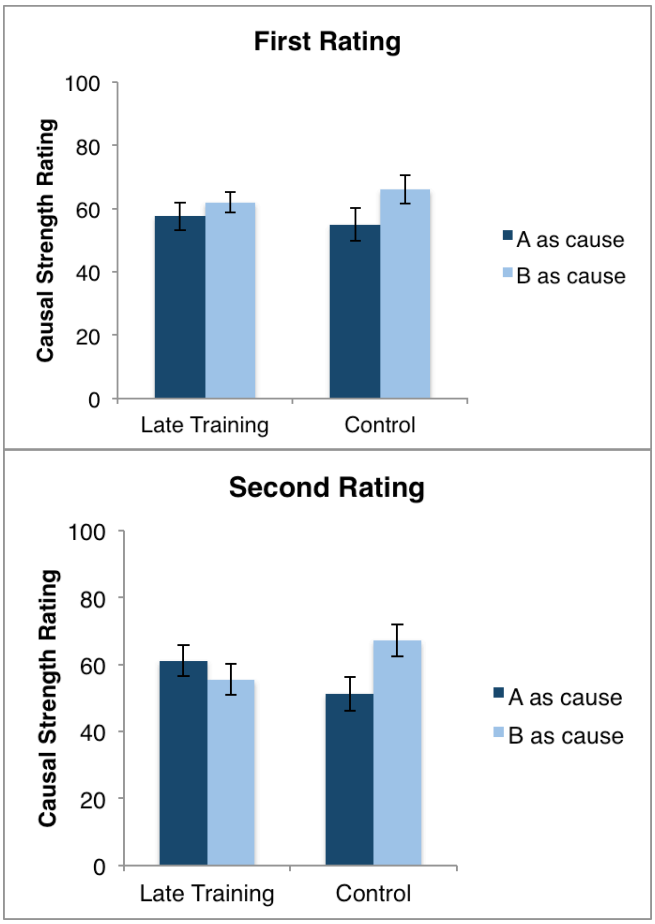

Figure 6: Mean (and standard error) causal strength ratings for Experiment 3.

We conducted a 2 (Late Training vs. Control; between-subjects) X 2 (1st Rating vs. 2nd Rating; within-subjects) mixed-measures ANOVA. There was no statistically significant main effect comparing the Late Training condition (1st Rating M = -4.29; 2nd Rating M = 5.70) to the Control condition (1st Rating M = -11.15; 2nd Rating = -16.05), $F(1, 63) = 3.62$, $p = .06$, $\eta^2_p = .054$. Nor was there a significant difference between the first and second rating, $F(1, 63) = 0.62$, $p = .43$, $\eta^2_p = .010$. Critically there was an interaction between the two variables, $F(1, 63) = 5.30$, $p < .05$, $\eta^2_p = .078$ because the Late Training condition's ratings increased from the first to the second (Rating Change M = 9.99), while the Control condition's rating actually

[1] The Early Training condition had a slightly higher summary score, but this was indistinguishable from random error.

decreased (Rating Change M = -4.90). The Category Training condition's increase was statistically significant, $t(34) = 2.15$, $p < .05$; the Control condition's decrease was not $t(29) = 1.11$, $p = .28$.

These results replicate the key finding of Experiment 2, that when following the category training, a second chance to consider the statistical pattern amongst the A, B, and C variables reduces the causal imprinting effect. While Experiment 3 seemed to show a larger causal imprinting effect overall than Experiment 2, the ratings change across the two experiments were similar (Experiment 2 Late Training M = 12.85; Experiment 3 Late Training M = 9.99).

General Discussion

Across three experiments a clear pattern emerged. When two correlated variables were observed together, subjects inferred that one caused the other. This inference became imprinted on the subjects' minds. New statistical evidence showing that a third variable was in fact the common cause of the two original variables (which have no direct causal link) was not alone sufficient to change their minds. However, learning that the common cause relation was a more general causal structure that many phenomena shared, supported changing the interpretation of the statistical evidence. When subjects considered the statistical evidence for a second time, they (on average) applied their causal category knowledge and shifted their understanding in the direction of the true causal structure. That Experiment 2 showed no effect of category training at the first rating, and that in Experiment 3 only the category training condition elicited a significant difference between the first and second ratings suggests that both category training and having a second chance to interpret the evidence are necessary for belief revision.

We suggest that causal system category training supports belief revision by offering a conceptual tool to make sense of statistical evidence. Many models of causal learning suggest that with each new piece of evidence, multiple causal structure hypotheses are evaluated (e.g., Tenenbaum, Griffiths, & Kemp, 2006). The current work is consistent with prior findings that explicit training in specific causal structures is sometimes necessary for those causal structures to be directly evaluated (Fernbach & Sloman, 2009). Here, learning about the common cause relation at a general, categorical level was necessary to revise beliefs to be more consistent with a common cause structure. In contrast, Experiment 1 showed category training was not necessary to infer a common cause structure when all relevant variables were considered from the beginning.

In the category training task, subjects learned about common cause structures (relevant to the observational learning task) and causal chain structures (irrelevant to the observational learning task). Learning two categories suggests that subjects' knowledge transfer from the category training task to the observational learning task was selective. In Experiments 2 and 3, category training elicited changes to causal strength ratings to be more consistent with a common cause structure, not a causal chain structure. The latter would have seen no decrease in B's strength to cause C, but may have seen a decrease in the direct link from A to C. It is possible that by teaching them two causal system categories we made the revision task harder than the prior work on false beliefs (reviewed by Lewandowsky and colleagues, 2012). In that work, the original (false) causal explanation was explicitly refuted, and sometimes (depending on the study) an alternative explanation was given and stated as the true explanation. Here we were less direct, simply providing conceptual tools to the learner, and many were able to use them appropriately.

Of course, we recognize that the success was not total. Comparing Experiments 2 and 3 to Experiment 1 shows that the category training did not make the ratings as consistent with a common cause structure as if causal imprinting was never induced. Still, given how stubborn false beliefs can be, any success in revising them is quite promising. The next steps in our research will examine whether this form of training can help change subjects' pre-existing beliefs.

Acknowledgments

This research was funded by Australian Research Council Grant DP150104267 awarded to MBG.

References

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678- 693.

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition*, *137*, 137-153.

Hornsey, M. J., & Fielding, K. S. (2017). Attitude roots and Jiu Jitsu persuasion: Understanding and overcoming the motivated rejection of science. *American Psychologist*, *72*(5), 459-473.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*(5), 406-412.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive science*, *36*(5), 919-932.

Taylor, E. G., & Ahn, W. K. (2012). Causal imprinting in causal structure learning. *Cognitive psychology*, *65*(3), 381-413.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309-318.