Similarities and differences: Comment on Chan et al.

Peter M. Jones, Chris J. Mitchell, Andy J. Wills, and Stuart G. Spicer

University of Plymouth

Correspondence:
Peter M. Jones
School of Psychology
University of Plymouth
Plymouth
PL4 8AA
United Kingdom
Email: peter.m.jones@plymouth.ac.uk

Abstract:

Spicer et al. (2020) reported a series of causal learning experiments in which participants appeared to learn most readily about cues when they were not certain of their causal status, and proposed that their results were a consequence of participants' use of "theory protection". In the present issue Chan et al. present an alternative view, using a modification of Rescorla and Wagner's (1972) influential model of learning. Although the explanation offered by Chan et al. appears very different to that suggested by Spicer et al., there are conceptual commonalities. Here we briefly discuss the similarities and differences of the two approaches, and agree with Chan et al.'s proposal that the best way to advance the debate will be to test situations in which the two theories make differing predictions.

Spicer et al. (2020) used Rescorla's (2000) compound test procedure to assess the idea that humans engage in "theory protection". In their experiment, as a result of early training, participants were certain that one cue (Y) was non-causal. Another cue (X) was assigned intermediate causal ratings, with participants indicating that they were uncertain about whether or not X was a cause. Both cues were then presented in compound with the outcome (XY+). The test data suggested that X had gained more associative strength than Y on XY+ trials. Hence, learning on XY+ trials seemed to be at odds with Rescorla's (2001) proposal that learning should be determined by prediction error. Rather, Spicer et al. proposed that participants engaged in theory protection for Y because they were certain that it was non-causal, and attributed the outcome to X because they were less certain of its status. Chan et al. argue (current issue) that theory protection is not required to explain Spicer et al.'s findings. They argue that X and Y may have accrued equal associative strength on XY+ trials, but that the differing associative strengths of the two cues at the start of XY+ training may, due to a learning-performance function proposed by Holmes et al. (2019), have led to a greater increment in responding to X than to Y.

At first glance, Chan et al.'s explanation appears very different to the theory protection proposal. However, they are different levels of explanation. The theory protection proposal is a description of a psychological process, whereas Holmes et al. (2019) described a mathematical model. According to their model, changes in associative strength bring about smaller changes in performance when a cue is located at a point on the mapping function with a shallow gradient, and larger changes when the mapping function is steeper. But what does the gradient of the mapping function represent in psychological terms? Holmes et al. did not describe the psychological processes that their mapping function represents, leaving open the possibility that their model is not in conflict with the theory protection idea at all. Here we first comment on the similarities between the two approaches in their explanations of Spicer

et al.'s (2020) data, and then compare the two approaches more broadly, considering whether the quantitative predictions of Holmes et al.'s model are a good fit for theory protection.

The two approaches offer similar accounts of Spicer et al.'s (2020) compound test data. Suppose one were to represent the theory protection proposal in a mathematical model. This model might include a function in which more experience of X-outcome pairings, but not Y-outcome pairings, would result in an increase in responding. In other words, the function should be steeper where X is located, and flatter where Y is located. In the specific design used by Spicer et al., Holmes et al.'s (2019) model therefore captures the properties of theory protection. It is not so surprising, therefore, that Chan et al.'s retrofit of the Holmes et al. model was successful. The two approaches also make further predictions in common. For instance, Chan et al. predicted that the difference in learning about X and Y observed by Spicer et al. should be largest when the associative strength of Y, prior to XY+ training, is close to zero. This is because, under these circumstances, the gradient of the mapping function is also close to zero. The theory protection account makes the same prediction, because participants who are certain that Y is not a cause of the outcome prior to compound conditioning should be especially unwilling to change this belief. Although the expression differs, these two proposals in essence offer the same explanation for the data from the compound test. However, there are two reasons to prefer the theory protection account as an explanation. Firstly, Spicer et al.'s design was motivated, and the results predicted, by the theory protection idea. Secondly, as Chan et al. have conceded, Holmes et al.'s model can only accommodate the results of the compound test if it allows the associative strength of X to be higher than that of Y immediately before XY+ training. However, they have not provided an account of how that happens. By contrast, Spicer et al. offered an account of their entire experiment.

We turn now to the more general question of whether Holmes et al. (2019) have provided a mathematical formulation of theory protection – are these two theories the same, but at different levels of explanation? For this to be true, participants must always be certain about cues that appear in the flat areas of the double-sigmoid function, where training translates into only small changes in responding. The curve is flattest when associative strength is close to 1, 0, or -1, so certainty would have to be highest at these same points for the theories to be the same. Conversely, responding is predicted to change most for cues with 0.5 and -0.5 strength because those are the steepest parts of the curve, so participants should be least certain about these cues. Informally, we have observed many situations in which these predictions are accurate. Participants often categorize cues as causal (value = 1), non-causal (value = 0), or preventative (value = -1) with high certainty, and assign intermediate ratings to cues about which they are uncertain. However, it seems unlikely that certainty and associative strength are very tightly linked in this way. We believe that there will be exceptions to the pattern described above, and describe one example here in detail.

The two accounts differ in their predictions for cues with associative strength of zero. According to the Holmes et al. (2019) model, changes in associative strength will have little effect on performance for cues with a starting associative strength of zero because the gradient of the mapping function is low. Furthermore, this should be the case both when the cue has been shown to be neutral (e.g. by being presented in the absence of the outcome) and when it is novel, because the translation of learning to performance depends only on associative strength. The theory protection account makes different predictions. For cues that have been presented without consequence, participants should try to protect their theory that the cue does not cause the outcome, changing their beliefs slowly when the cue and outcome are paired subsequently. For novel cues, on the other hand, participants should have no theory to protect and should therefore readily change their beliefs about the cue when it is paired

with the outcome. The theory protection proposal therefore predicts that the results of compound training that includes cues with zero associative strength will be influenced by the prior training history of those cues, whereas Holmes et al.'s model does not.

This analysis applies to compound training in which the outcome is present (e.g. XY+), but the theory protection principle makes different predictions when the outcome is absent during compound conditioning trials. Consider an experiment reported by Rescorla (2001), in which rats received pairings of one stimulus with food (A+) and nonreinforced presentations of another stimulus (B-). They then received AB- training. A subsequent compound test indicated that the rats learned more about A than B on the compound conditioning trials. Rescorla attributed this to unequal learning about the two stimuli, but Holmes et al. (2019) showed that their model predicts this result because B, having zero associative strength after B- training, would be located on the flattest part of their mapping function at the start of the AB- trials. The theory protection account makes a different prediction for an analogous experiment with humans. According to theory protection, B- training should result in participants having a theory that B does not cause the outcome. However, this theory does not conflict with the possibility that B prevents the outcome, provided the causal scenario is presented to participants in such a way that they should not expect preventative cues to have any effect when presented alone (see Melchers, Wolff, and Lachnit, 2006). On the other hand, participants should attempt to protect their theory that A causes the outcome following A+ training. We therefore predict that, provided the causal scenario allows cues to prevent the outcome and for preventative cues to appear neutral when presented alone, participants should learn more about B than A on AB- trials. This is the opposite prediction to that made by Holmes et al.

In summary, we agree that Holmes et al.'s (2019) model can accommodate some aspects of Spicer et al.'s (2019) results. For this experiment, their model resembles a

mathematical expression of the notion of theory protection. However, we propose that the two theories will not always align so neatly, and the best way to make progress will be to test cases in which certainty is dissociated from the gradient of Holmes et al.'s mapping function.

References

Chan, Y. Y., Westbrook, R. F., & Holmes, N. M. (current issue). Protecting the Rescorla-Wagner (1972) theory: a reply to Spicer et al. (2019). *Journal of Experimental Psychology: Animal Learning and Cognition*.

Holmes, N. M., Chan, Y. Y., & Westbrook, F. (2019). A combination of common and individual error terms is not needed to explain associative changes when cues with different training histories are conditioned in compound: A review of Rescorla's compound test procedure. *Journal of Experimental Psychology: Animal Learning and Cognition, 45,* 242-256.

Melchers, K. G., Wolff, S., & Lachnit, H. (2006). Extinction of conditioned inhibition through nonreinforced presentation of the inhibitor. Psychonomic Bulletin & Review, 13(4), 662–7.

Rescorla, R. A. (2000). Associative changes in exciters and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes, 26,* 428-438.

Rescorla, R. A. (2001). Unequal associative changes when excitors and neutral stimuli are conditioned in compound. *Quarterly Journal of Experimental Psychology, 54B,* 53–68.

Rescorla, R. A., and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Spicer, S. G., Mitchell, C. J., Wills, A. J., and Jones, P. M. (2020). Theory protection in associative learning: humans maintain certain beliefs in a manner that violates prediction error. *Journal of Experimental Psychology: Animal Learning and Cognition, 46,* 151-161.