The Role of Surprise in Learning:

Different Surprising Outcomes Affect Memorability Differentially

Meadhbh I. Foster[1] and Mark T. Keane[1]

[1]School of Computer Science and Informatics,

University College Dublin, Ireland

Running head: SURPRISE IN LEARNING

Correspondence concerning this article should be addressed to:

Meadhbh I. Foster

School of Computer Science and Informatics

University College Dublin

Belfield, Dublin 4

Ireland

E-mail: meadhbh.foster@ucdconnect.ie

Telephone: 00353 (0)1 7162907

Abstract

Surprise has been explored as a cognitive-emotional phenomenon that impacts many aspects of mental life from creativity to learning to decision making. In this paper, we specifically address the role of surprise in learning and memory. Although surprise has been cast as a basic emotion since Darwin's (1872) *The Expression of the Emotions in Man and Animals*, recently more emphasis has been placed on its cognitive aspects. One such view casts surprise as a process of "sense making" or "explanation finding": Metacognitive Explanation Based (MEB) theory proposes that people's perception of surprise is a metacognitive assessment of the cognitive work done to explain a surprising outcome. Or, to put it more simply, *surprise increases with the explanatory work required to resolve it*. This theory predicts that some surprises should be more surprising than others because they are harder to explain. In the current paper, this theory is extended to consider the role of surprise in learning as evidenced by memorability. This theory is tested to determine how scenarios with differentially-surprising outcomes impact the memorability of those outcomes. The results show that surprising outcomes (less-known outcomes), that are more difficult to explain, are recalled more accurately than less-surprising outcomes that require little (known outcomes) or no explanation (normal).

Keywords: surprise, learning, explanation, recall

Why are some events more surprising than others? Why are we less surprised to hear that Kurt Cobain or Amy Winehouse have died, but more surprised when we hear Michael Jackson or Lady Diana have died? The Metacognitive Explanation Based (MEB) theory of surprise proposes that surprise is fundamentally about explanation; explanations that make sense of the world and help to resolve the surprise we feel (see Foster & Keane, 2013, 2015a, 2015b, 2015c; Foster et al., 2014). So, the death of Michael Jackson was much more surprising than that of Kurt Cobain because it was so much harder to explain (i.e., Jackson was older, clean-cut and had no advertised addiction history); Jackson did not fit the early-career, rock-and-roll-suicide explanation exemplified by Cobain, Joplin, Hendrix and others.

This theory does not deny that surprise has a significant emotional component, but it does focus more on the cognitive aspects of surprise in an attempt to understand its adaptive role; namely, that surprise helps people make sense of a sometimes bewildering and uncertain world. Typically, when one resolves a surprising event by explaining it, one learns something new about the world – something that, hopefully, helps one to better deal with such events in the future. Hence, one would expect the memorability of surprising events to vary with the amount of explanation they elicit. For instance, intuitively, one would expect very surprising events to be more memorable than less surprising events, as they involve elaborations that are known to foster memorability (e.g., inferring causal structure). In this paper, we test this prediction using a novel experimental paradigm, in the surprise literature, to explore the subtle interactions between surprise and memorability. But, first we briefly review some of the relevant prior work.

How Surprise Fits with Explanation & Learning

The above proposals are just a variant of the long-standing view that surprise plays a key role in learning. In the literature on child development (Piaget, 1952) and education (Adler, 2008) it is generally accepted that people learn about their environment by explaining it (see Lombrozo, 2012). Such explanations are seen as being triggered by surprise (Ramscar, Dye, Gustafson, & Klein, 2013; Tsang, 2013) or inconsistencies (Johnson-Laird, Girotto, & Legrenzi, 2004; Hastie, 1984). For instance, the use of explanation to deal with inconsistencies has been shown to influence categorisation, concept development and communication (e.g., Khemlani & Johnson-Laird, 2011; Khemlani, Sussman, & Oppenheimer, 2011); also, inconsistent outcomes have also been shown to trigger explanation in early childhood (Legare, Schult, Impola, & Souza, 2016).

Adler (2008) proposes that surprises are of great value in learning; that as students encounter a surprising piece of information their attention is aroused (i.e., they notice the surprise), provoking more intensive processing of the to-be-learned material (i.e., resolving; there is a call to correct and better understand the material). Previous work has also shown that surprise can increase the retention of information (e.g., Munnich, Ranney, & Song, 2007), perhaps because surprise can make an event more interesting and likeable (Loewenstein & Heath, 2009). Related views are echoed in Artificial Intelligence, where surprise has been proposed as *the* cognitive mechanism to identify events that are learning opportunities in agent architectures for robots (Bae & Young, 2008, 2009; Macedo & Cardoso, 2001; Macedo & Cardoso, 2012; Macedo, Cardoso, Reisenzein, Lorini, & Castelfranchi, 2009; Macedo, Reisenzein, & Cardoso, 2004).

Explanation, Learning & Memory

Traditionally, explanation is seen as playing a role in building causal models or predictive schemas to deal with future events (Heider, 1958; Lombrozo & Carey, 2006). In AI, Leake (1992; also, Schank, 1986; Schank, Kass, & Riesbeck, 1994) have argued that when anomalies are detected (e.g., surprising events) an explanation is built (or retrieved) to account for the anomaly with that explanation being indexed for future use. Explanations may also serve to help people decide how information should be weighted or how attention should be allocated as events unfold (Dehghani, Iliev, & Kaufmann, 2012; Keil, 2006; Keil, Carter Smith, Simons, & Levin, 1998). As such, explanation has been linked to a variety of cognitive processes, including diagnosis (Graesser & Olde, 2003), categorization (Murphy & Medin, 1985), and reasoning (e.g., Lombrozo, 2006). In the education literature, self-explaining and self-explanation training are known to improve text comprehension and learning (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, De Leeuw, Chiu, & LaVancher, 1994; Durkin, 2011; Roy & Chi, 2005), particularly for low-knowledge readers (McNamara, 2001; McNamara & Scott, 1999). Indeed, these effects are enhanced if there are reliable patterns and consistencies in the material that the explanation uncovers (Williams, Lombrozo, & Rehder, 2010, 2013).

Finally, in the field of neuropsychology, many findings from ERP studies show that, in comprehension, prediction is an effective strategy except when such predictions could be misleading (as in within-category violations), in which case a more complex comprehension process is required (namely, explanation). Similarly, and more generally, the left-hemisphere of the brain appears to be biased towards processing in an anticipatory, predictive fashion, while the right hemisphere appears to process information in a more post-hoc, integrative way (for a review, see Federmeier, Wlotko & Meyer, 2008).

Adaptation & Explanation

Piaget's theory of child development (e.g., Piaget, 1952), and appraisal theories of emotion have both previously recognised that people are skilled explainers of their environment. Research in hindsight bias has determined that people attempt to understand events as soon as they occur (e.g., Pezzo, 2003), and studies of affective adaptation have shown that when an explanation has been found, an event seems more predictable (Wilson & Gilbert, 2008). To capture this phenomenon, Wilson and Gilbert have developed a model of affective adaptation (i.e., when affective responses are weaker with repeated exposure to an emotional event) in which people Attend to self-relevant, unexplained events, React emotionally to these events, Explain or reach an understanding of these events, and thereby Adapt to these events (their AREA model). This model proposes that people try to make sense of events as best as they can, a process that may include comparing them to their past experiences, but could also include comparing them with hypothetical alternative scenarios, or assimilating the event to existing knowledge structures. Similarly, Pyszczynski and Greenberg (1981) found that people were more likely to search for explanations of behaviour that did not conform to expectancies than behaviour that did; to search for a new explanation would require additional, and unnecessary, cognitive work (see also Hastie, 1984; Lau and Russell, 1980; Wong and Weiner, 1981). MEB theory proposes that for situations where the outcome is one that is expected to occur, the 'explanation' has been computed in advance, so no further cognitive work is necessary to resolve the outcome."

Surprise Judgements, Explanatory Knowledge & Memorability

These diverse literatures on surprise, explanation and learning all converge on a common theory for how surprise and explanation impact learning and, by extension, memorability. When people encounter surprising, anomalous or unexpected events they are prompted to explain these anomalies, a process of causal elaboration that produces richer memory encodings, thus improving the memorability of the focal event. However, recent work on surprise suggests a novel extension to this common theory; namely, that relative differences in the surprisingness of events elicit different explanatory processing that have consequential effects on memorability.

MEB theory (Foster & Keane, 2015c) states that the perception of surprisingness is based on a metacognitive assessment of the effort-to-explain. The theory maintains that the cognitive system roughly tracks the amount of work done in explaining a surprising outcome and this assessment underlies the perceived surprisingness of the target event. Foster & Keane argued that there are, at least, two distinct classes of surprise scenarios; some surprising events are resolved by pre-packaged, explanatory knowledge (i.e., *known surprise scenarios*), whereas others are truly surprising (i.e., *less-known surprise scenarios*) that require the construction of explanations *ab initio*, from scratch. Foster & Keane (2015c) illustrated these classes with the following example:

(1) Walking down the street, John finds that his wallet is missing from his trouser pocket, where he put it this morning. (*known surprise*)

(2) Walking down the street, John finds that his belt is missing from his trousers, where he put it this morning. (*less-known surprise*)

Both of these scenarios describe events that are surprising to the actor. But, the missing-wallet scenario (*known scenario*) intuitively seems to elicit many ready-made explanations that we have used before to explain such missing-wallet events (e.g., it was robbed, I left it in a shop, I left it in my other trousers). In contrast, the missing-belt scenario (*less-known scenario*) intuitively seems to have no ready-made explanations and, thus, requires building an explanation from scratch (e.g., using knowledge about assumed events, such as, that John travelled through airport security and left his belt there). In AI, Schank (1986) makes exactly the same distinction between *canned explanations* and *additive explanations*; the former are already-encoded, explanation schemas in memory that are "ready to go", whereas the latter are "critical to the learning process…(as) after the explanation is finished, the explainer now knows something that he didn't know before" (p. 28). Indeed, Leake (1992) elaborates a whole taxonomy of pre-stored "explanation patterns" at different levels of abstraction to handle such anomalous events. In a series of experiments, Foster & Keane (2015c) have found that scenarios involving known-surprise scenarios were judged to be less surprising, were easier to explain (as measured by total number of explanations generated by participants), but produced fewer distinct explanations (as measured by proportion-of-agreement between explanations produced by participants) than scenarios involving less-known surprises. This converging evidence supports the view that the explanatory processes elicited to these differentially surprising events are quite different, perhaps in ways that affect learning and memorability.

It can be argued that these known-surprise scenarios involve a type of schema filling using pre-stored explanation patterns, akin to the sort of schema-filling one does in normal comprehension of expected events. However, less-known-surprise scenarios have no ready-made explanatory schemas and require active construction of some causal model of what might have occurred. If this theoretical account is indeed the case, then differential impacts on memorability should follow. Namely, that less-known-scenarios should engender the sort of causal elaboration that increases memorability, whereas known-scenarios and normal-scenarios should not increase memorability in the same way, as they do not elicit such elaboration. Indeed, the latter two scenarios both involve schema-filling, albeit perhaps different types of schema, and therefore could be predicted to have equivalent levels of memorability. Here, we explore these predictions in a novel paradigm examining surprise and memorability.

The Present Study: A Tale of Three Scenarios

This study used materials describing a variety of everyday scenarios – short stories with a setting and an outcome – where the outcome was varied to have a known-surprise, a less-known surprise or a normal, expected outcome (see Table 1 for one example). This manipulation of the known and less-known outcomes, and the materials used, corresponded to those used in previous studies (see Foster & Keane, 2015c), though the normal scenario with its expected outcome was a new addition.

Table 1. Example of Materials Used in the Study

(only the label "outcome" was shown to participants in presented materials).

Setting	Rebecca is on the beach. She goes for a swim in the water.		
Outcome	**Known**	**Less-Known**	**Normal**
	After she dries herself off she notices that her skin has turned red.	After she dries herself off she notices that her skin has turned turquoise.	After she dries herself off she notices that her skin is no longer wet.

The design for this experiment manipulated Outcome-Type (known, less-known, or normal), making the prediction for the judgement-of-surprise measure that less-known outcomes would be rated as more surprising than known outcomes, which would in turn be rated as more surprising than the normal outcomes (which are unsurprising). For the memorability or recall-measure, the prediction was that less-known outcomes would be recalled more accurately than either the known- or normal-outcomes because less-known outcomes involve causal elaboration to explain the scenario. However, the known- and normal-outcomes should be equally memorable as they both involve a schema-filling type of comprehension.

A standard cued-recall paradigm was adapted for the experiment involving three tasks: participants first carried out a task in which they judged the surprisingness of outcomes of the scenarios, were then distracted by a filler task, and then lastly completed an unexpected cued-recall task (see, e.g., Ham & Vonk, 2003; Hassin, Aarts, & Ferguson, 2005; Hassin, Bargh, & Uleman, 2002; McCulloch, Ferguson, Kawada, & Bargh, 2008).

Method

Participants and design. Thirty UCD students (13 male, 17 female) with a mean age of 20.97 years (*SD* = 3.65, range = 18-38) took part voluntarily in this study. Informed consent was obtained prior to the experiment. A within-subjects (Outcome-Type: known, less-known, normal) repeated-measures design was used.

Procedure, materials and scoring. The experiment had three distinct parts: a surprise-judgement task where participants were first exposed to the materials, a filler task, and an unannounced cued-recall task. In the first task, participants were asked to read each of nine randomly-presented stories, and to judge the surprisingness of their outcomes after reading each story. Each story was presented on a separate page with the scenario setting on the top of the page, followed by the outcome (known/less-known/normal), and a 7-point scale on which to provide surprise judgements.

Then, after handing this booklet back to the experimenter, participants engaged in a filler task that served to clear short-term memory. For this, they were presented with a new booklet which provided them with the filler task instructions; they were asked to write down different uses for two common objects (a brick and a car tyre) on separate pages of this booklet. After four minutes on this task, the experimenter interrupted them and introduced the third part of the experiment, where they were unexpectedly presented with a cued-recall task for the scenarios they had rated for surprise in the first part of the experiment.

For this recall task, participants were provided with a booklet that began with a reminder that they had read nine short stories in the first part of the study. They were informed that in the booklet they would be provided with the first sentence of each story, and asked to "*write down everything that you can remember about what happened next in that story*". Each story was presented in a randomized order on a separate page with the word "recall" followed by the first sentence of that story on the top of the page, and a sentence that reminded them to write down everything that they could remember from the story. Two measures were recorded: the 7-point scale judgment of surprise from the first task, and the answers that they provided in the third part, the recall task. The answers provided for the recall task were submitted to a propositional analysis and assigned a recalled-items score.

Results and Discussion

Overall, the results confirmed the prediction that Outcome-Type impacts people's perception of surprise, with less-known outcomes being rated as more surprising than known outcomes, which were rated as more surprising than the normal outcomes. Analysis of memorability of outcomes in recall task showed that highly surprising events are more memorable; more items were recalled for the surprising less-known outcomes, than for the less surprising known outcomes and normal outcomes. Indeed, the known and normal outcomes both produced equivalent recall levels, perhaps as a result of how they were understood.

Surprise judgments. A one-way repeated measures ANOVA[1] showed a main effect of Outcome-Type, where participants judged stories with less-known outcomes to be more surprising (M = 5.93, SD = 1.13) than known outcomes (M = 4.43, SD = 1.26), and these were both judged to be more surprising than the normal outcomes (M = 1.72, SD = .72), $F(2, 58) = 130.67$, $p < .001$,

[1] See Norman (2010) for a discussion on the use of parametric statistics in Likert scale data.

$\eta_p^2 = .82$ (see Figure 1). Pairwise comparisons (with Bonferroni corrections) showed that these differences were significant between all possible pairings (p's $< .001$).

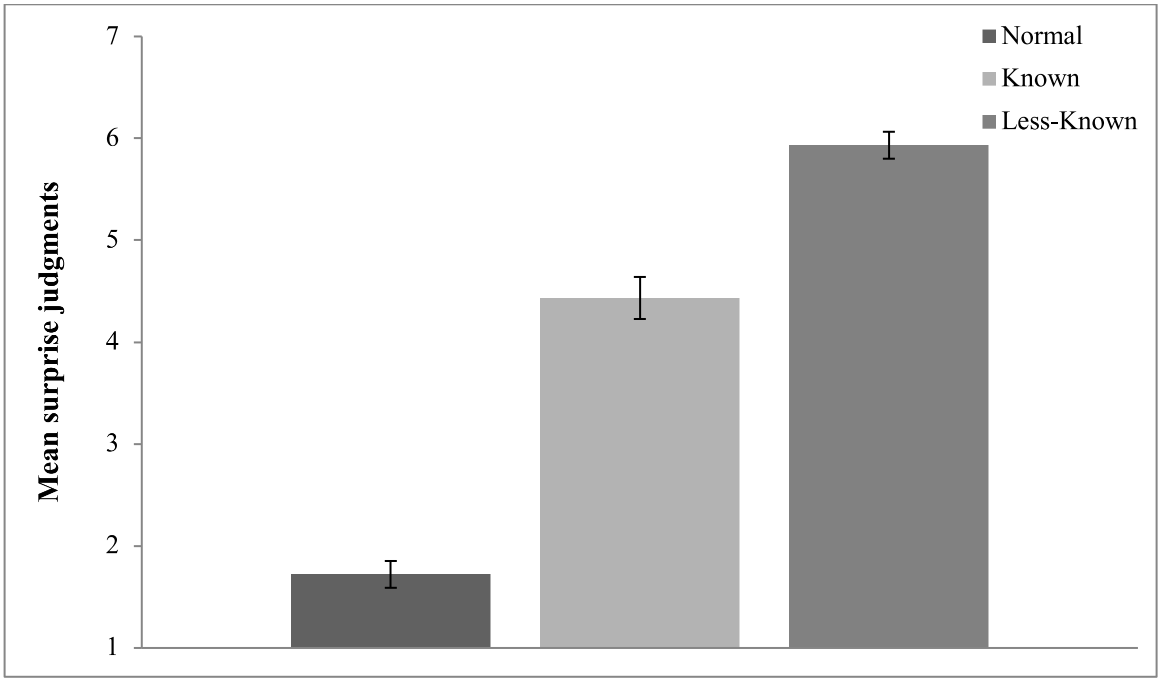

Figure 1. Mean surprise judgments for all levels of Outcome-Type (normal, known, less-known) with standard errors ($N = 30$).

Recall. The recollections that participants provided in the recall task were submitted to a propositional analysis noting the key items mentioned from the scenario's outcome (i.e., actors/objects/attributes; see, e.g., Brown & Yule, 1983). Each recollection was scored proportionally for its completeness in mentioning all the items in a given outcome. So, in the Rebecca story where she is on the beach (see Table 1), and then notices that her *skin* has turned *red*, the recalled answer that "she was *red*" would be given a score of .5, as 1 out of 2 items are recalled correctly (*red*), while the answer that "her *skin* was *red*" would be given a score of 1, as 2 out of 2 (*skin, red*) items are recalled correctly. Note, one reviewer queried whether the less-surprising materials (known and normal outcomes) could invite more concise responses as they involve commonly understood implications; for instance, people might be more likely to truncate their responses, saying "she was red" instead of "she had red skin". However, an inspection of the responses did not show any systematic pattern of such responding in these conditions compared to the most surprising condition (i.e., less-known; e.g., people seemed just as likely to say "she was turquoise" in the less-known condition).

A one-way repeated-measures ANOVA on this Outcome-Type variable showed a main effect, where participants remembered more items from the less-known outcomes (M = .75, SD = .21), than from known outcomes (M = .66, SD = .21) or normal outcomes (M = .65, SD = .17), $F(2, 58) = 3.30$, $p = .04$, $\eta_p^2 = .10$ (see Figure 2). Planned pairwise comparisons (with Bonferroni corrections for two comparisons, following our predictions detailed above) showed significantly more items recalled for less-known outcomes compared to both known and normal outcomes (p's $< .05$, 1-tailed), with post-hoc pairwise comparisons showing no significant difference between recall scores for known and normal outcomes ($p = .76$).

Finally, by-materials Pearson product-moment correlation coefficients and regression analyses were computed to assess the relationship between surprise judgments and the recall scores. The surprise judgments and recall scores for less-known outcomes correlated ($r(28) = .36$, $p = .05$); greater surprise predicted greater recall (β= .359, $t(28) = 2.034$, $p = .05$), and explained a small but marginally significant proportion of this variance ($R^2 = .13$, $F(1, 28) = 4.139$, $p = .05$), but no such relationship was found for known outcomes, $r(28) = .077$, $p = .68$), or normal outcomes, $r(28) = -.32$, $p = .08$)[2]. Overall, this result suggests that the less-known outcomes are

[2] If we accept this correlation of marginal significance as a 'real' finding, it might perhaps invite the conclusion that low-surprise scores predict recalled, suggesting that when events follow a course that they always have followed in the past, we recall them more accurately. However, this improved accuracy may not be due to remembering the scenario in the experiment; it could just be due to people giving the most available information for the scenario in memory (namely, what always occurred before). Clearly, this issue could be explored further, though it is less about the phenomenon of surprise.

quite different than the known and normal outcomes, both of which produce near identical responding with respect to recall (though they differ with respect to judged surprise). This result supports the view that even though people are sensitive to the relative differences in surprisingness of these three outcomes, the outcomes differ in memorability because of differences in how they are explained/comprehended; namely, that the known and normal outcomes are processed by schema filling, which decrements memorability, relative to more actively and causally-elaborated less-known outcomes.

This conclusion is supported by the convergence of these results with those of many other experiments using these same materials (see Foster & Keane, 2015c). Foster & Keane (2015c) have found that these known-surprise scenarios are easier to explain than the less-known ones, as measured by total number of explanations generated by participants; that is, people readily generate explanations to the known-surprise scenarios because they have readily-available knowledge to use, whereas people struggle to generate explanations to the less-known-surprise scenarios. Foster & Keane (2015c) also found that people generated fewer distinct explanations to known surprises than they did to less-known surprises, as measured by the proportion-of-agreement between the explanations produced by participants; that is, the explanations generated to the known-surprise scenarios tend to be very similar presumably because they arise from shared schematic knowledge, whereas the explanations to the less-known-surprise scenarios are much more varied presumably because they are built from more idiosyncratic, personal knowledge.

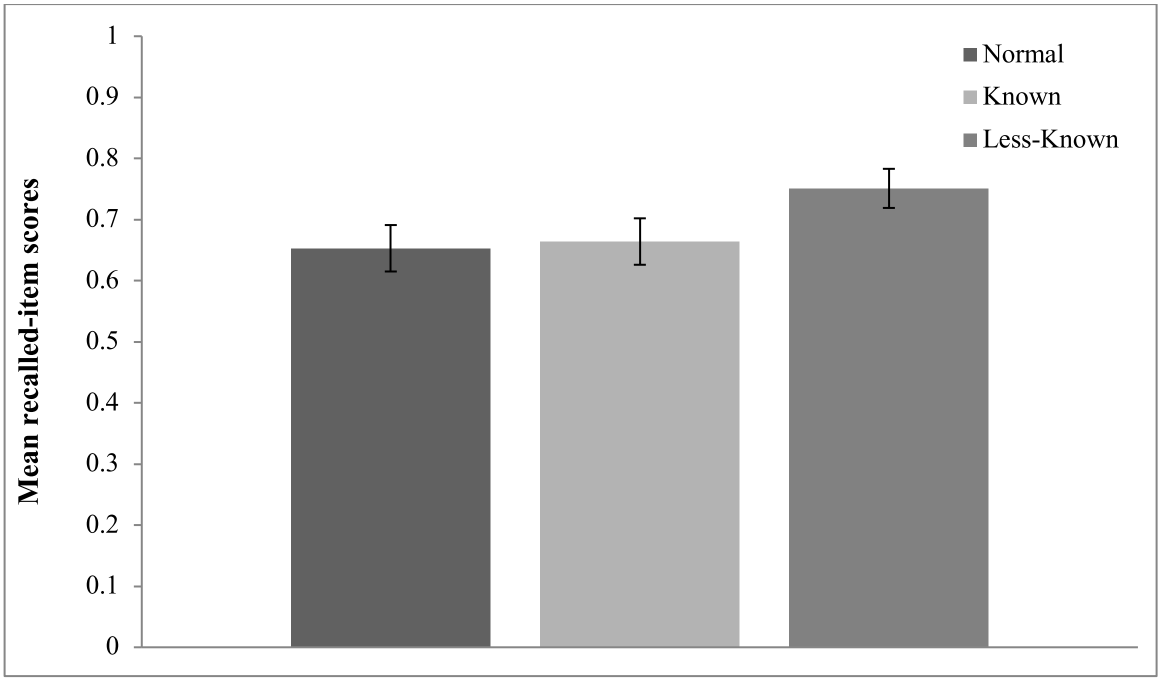

Figure 2. Mean recalled-item scores for all levels of Outcome-Type (normal, known, less-known, normal) with standard errors (N = 30).

Conclusions

This paper advances a new theoretical proposal about the role of explanation in surprise; specifically, it argues that different surprise scenarios elicit different explanatory strategies (e.g., explanation-schema filling versus causal model building) that have differential impacts on memorablility. This work thus contributes to the surprise literature but also has wider resonances in other topic areas (e.g., research on explanation, attribution and religious concepts).

The present experiment replicates and extends previous findings about surprise judgements, in showing that less-known outcomes are judged to be more surprising than known surprising outcomes which are, in turn, more surprising than normal outcomes. It also shows that surprise scenarios bear a complex relationship to memorability. Less-known surprises were remembered more accurately than known surprises and unsurprising outcomes, but there was no difference in recall for outcomes that had lower levels of surprise or no surprise at all (i.e., the known or normal outcomes). As there is no linear relationship between judged surprise and memorability, it is not the case that all surprising scenarios automatically result in higher memorability. Rather, it appears that memorability depends on the surprising outcome that results in causal elaboration of the scenario, as a novel explanation is built. When the scenario is less surprising (the known outcome), memorability drops on instantiation of an explanation-schema; recall in this condition is effectively identical to the normal scenario, where people where people apply normative, schematic knowledge to comprehend the scenario without need for elaboration. We believe that that theoretical account opens up several new vistas on how we think about surprise and, indeed, explanation, though it should be said that similar patterns of findings do occur in other topic areas.

For instance, in anthropology, researchers have argued that religious concepts are more memorable because they are *minimally counterintuitive* (e.g., a talking bird or a walking stone). varying somewhat from *intuitive* concepts (e.g., a flying bird, a heavy stone) while not being *bizarre* or *maximally counterintuitive* (e.g., a walking stone that quacks)[3]. In our terms, intuitive, minimally counterintuitive and maximally counterintuitive concepts are likely to be surprising to varying degrees. One set of findings from this literature suggests that minimally counterintuitive concepts are better remembered than intuitive and bizarre concepts (Boyer & Ramble, 2001; Barrett & Nyhof, 2001; Upala, Gonce, Tweney, and Slone, 2007). While there are controversies about how these effects interact with other variables, the core findings on the memorability of *minimally counterintuitive* concepts seems quite consistent with the present findings. When people understand a *counterintuitive concept* they have to "justify" or explain to themselves how such an

[3] We thank one of our reviewers for pointing out these correspondences to us.

ontological violation arises (e.g., how a bird could talk) and this elaboration makes the concept more memorable than intuitive concepts that are consistent with prior knowledge (e.g., a bird that flies). However, it should also be noted that *bizzare* or *maximally counterintuitive* concepts can also become less memorable if they are unexplainable; that is, if one cannot square them causally with one's knowledge, their causal elaboration does not succeed, and improved memorability does not follow. To quote, Upala et al (2007) the "memory of an intelligent agent should evolve to preferentially remember and recall those events/objects that violate the agent's expectations about the future but can be justified once they have been observed" (p. 432); to translate this to the current context one might say "unexpected, surprising events/objects that can be explained will be more memorable".

Surprise has long been thought to play a key role in learning, increasing the retention of information (see, e.g., Adler, 2008; Loewenstein & Heath, 2009; Munnich et al., 2007). It has also been suggested that people learn about their environment by explaining it (e.g., Lombrozo, 2012; Piaget, 1952), and that surprise may trigger this process (e.g., Adler, 2008; Ramscar et al., 2013; Tsang, 2013). With respect to learning, the present results suggest that it is not sufficient, in itself, to present learners with unexpected outcomes to guarantee better retention; rather, the degree of surprise must be high relative to the prior knowledge of the learner. From a practical perspective, this suggests that any educational materials using surprise to aid learning, will have to be carefully tailored to the knowledge of those who are expected to learn. The present paper suggests that if we can readily explain something, then it is more likely to be forgettable and no learning will occur, but if it challenges our knowledge to explain how it could have happened, then something will be remembered and, indeed, learned.

References

Adler, J. E. (2008). Surprise. *Educational Theory, 58*(2), 149-173.

Bae, B.C., & Young, R. M. (2008). A use of flashback and foreshadowing for surprise arousal in narrative using a plan-based approach. In U. Spierling & N. Szilas (Eds.), *Proceedings of ICIDS, LNCS, 5334* (pp. 156-167). Heidelberg: Springer.

Bae, B. C., & Young, R. M. (2009). Suspense? Surprise! or How to generate stories with surprise endings by exploiting the disparity of knowledge between a story's reader and its characters. In I. A. Iurgel, N. Zagalo & P. Petta (Eds.), *Proceedings of ICIDS, LNCS, 5915* (pp. 304-307). Berlin: Springer.

Barrett, J. L., & Nyhof, M. A. (2001). Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *Journal of cognition and culture*, *1*(1), 69-100.

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: Cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science*, *25*(4), 535-564.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge University Press.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145-182.

Chi, M. T. H., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Darwin, C. R. (1872). *The expression of the emotions in man and animals.* London: John Murray.

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language, 27*(1), 55-85.

Durkin, K. (2011). The self-explanation effect when learning mathematics: a meta-analysis. *Science*, *26*(2), 147-179.

Federmeier, K. D., Wlotko, E. W., & Meyer, A. M. (2008). What's "right" in language comprehension: ERPs reveal right hemisphere language capabilities. *Language and Linguistics Compass, 2*(1), 1–17.

Foster, M. I., & Keane, M. T. (2013). Surprise! You've got some explaining to do. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2321-2326). Austin, TX: Cognitive Science Society.

Foster, M. I., & Keane, M. T. (2015a). Predicting surprise judgments from explanation graphs. In N. A. Taatgen, M. K. van Vugt, J. P. Borst, & K. Mehlhorn (Eds.), *Proceedings of the 13th International Conference on Cognitive Modeling* (pp. 124-129). The Netherlands: University of Groningen.

Foster, M. I., & Keane, M.T. (2015b). Surprise as an ideal case for the interplay of cognition and emotion. *Behavioral and Brain Sciences, 38,* e74.

Foster, M. I., & Keane, M. T. (2015c). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology, 81*, 74-116.

Foster, M. I., Keane, M. T., Loewenstein, J., Maguire, P., Maguire, R., May, R. W., Smith-Rodden, M., Ash, I. K., Munnich, E., & Ranney, M. A. (2014). Triangulating surprise: Expectations, uncertainty, and making sense. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 218). Austin, TX: Cognitive Science Society.

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology, 95*(3), 524-536.

Ham, J., & Vonk, R. (2003). Smart and easy: Co-occurring activation of spontaneous trait inferences and spontaneous situational inferences. *Journal of Experimental Social Psychology*, *39*(5), 434-447.

Hassin, R. R., Aarts, H., & Ferguson, M. J. (2005). Automatic goal inferences. *Journal of Experimental Social Psychology*, *41*(2), 129-140.

Hassin, R. R., Bargh, J. A., & Uleman, J. S. (2002). Spontaneous causal inferences. *Journal of Experimental Social Psychology*, *38*(5), 515-522.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology*, *46*(1), 44-56.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: John Wiley & Sons.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111*(3), 640-661.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57,* 227-254.

Keil, F. C., Carter Smith, W., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition*, *65*(2), 103-135.

Khemlani, S. S., & Johnson-Laird, P. N. (2011). The need to explain. *The Quarterly Journal of Experimental Psychology*, *64*(11), 2276-2288.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*(3), 527-535.

Leake, D. B. (1991). Goal-based explanation evaluation. *Cognitive Science, 15*(4), 509-545.

Legare, C. H., Schult, C. A., Impola, M., & Souza, A. L. (2016). Young children revise explanations in response to new evidence. *Cognitive Development, 39*, 45-56.

Loewenstein, J., & Heath, C. (2009). The Repetition ‐ Break plot structure: A cognitive influence on selection in the marketplace of ideas. *Cognitive Science*, *33*(1), 1-19.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464-470.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning* (pp. 260-276), United Kingdom: Oxford University Press.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99*(2), 167-204.

Macedo, L., & Cardoso, A. (2001). Creativity and surprise. In G. Wiggins (Ed.), *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science* (pp. 84-92). United Kingdom: University of York.

Macedo, L., & Cardoso, A. (2012). The exploration of unknown environments populated with entities by a surprise-curiosity-based agent. *Cognitive Systems Research, 19-20,* 62-87.

Macedo, L., Cardoso, A., Reisenzein, R., Lorini, E., & Castelfranchi, C. (2009). Artificial surprise. In Vallverdú, J. & Casacuberta, D. (Eds.) *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence* (pp. 267-291). United Kingdom: Information Science Reference.

Macedo, L., Reisenzein, R., & Cardoso, A. (2004). Modeling forms of surprise in artificial agents: Empirical and theoretical study of surprise functions. In K. Forbus, D. Gentner

& T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 588-593). Mahwah, NJ: Erlbaum.

McCulloch, K. C., Ferguson, M. J., Kawada, C. C., & Bargh, J. A. (2008). Taking a closer look: On the operation of nonconscious impression formation. *Journal of Experimental Social Psychology*, *44*(3), 614-623.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*(1), 51-62.

McNamara, D. S., & Scott, J. L. (1999). Training reading strategies. In *Proceedings of the 21st Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Lau, R. R., & Russell, D. (1980). Attributions in the sports pages. *Journal of Personality and Social Psychology*, *39*(1), 29-38.

Leake, D. (1992). *Evaluating Explanations.* Hillsdale, NJ: Erlbaum.

Munnich, E., Ranney, M., A. & Song, M. (2007). Surprise, surprise: The role of surprising numerical feedback in belief change. In D.S. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 503-508). Mahwah, NJ: Erlbaum.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*(3), 289-316.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education, 15*(5), 625-632.

Pezzo, M. V. (2003). Surprise, defence, or making sense: What removes hindsight bias? *Memory, 11*(4/5), 421-441.

Piaget, J. (1952). *The origins of intelligence in children.* New York: W. W. Norton & Company.

Pyszczynski, T. A., & Greenberg, J. (1981). Role of disconfirmed expectancies in the instigation of attributional processing. *Journal of Personality and Social Psychology, 40*(1), 31-38.

Ramscar, M., Dye, M., Gustafson, J. W., & Klein, J. (2013). Dual routes to cognitive flexibility: Learning and response ‐ conflict resolution in the dimensional change card sort task. *Child Development, 84*(4), 1308-1323.

Roese, N. J., & Olson, J. M. (1996). Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology, 32*(3), 197-227.

Roy, M., & Chi, M. T. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning.* (pp. 271-286). New York: Cambridge University Press.

Shank, R. C. (1986). *Explanation Patterns.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. C., Kass, A., & Riesbeck, C. K. (Eds.) (1994). *Inside case-based explanation.* Hove, UK: Lawrence Erlbaum Associates.

Scheffler, I. (1977). *In praise of the cognitive emotions.* In I. Scheffler (2010) In Praise of the Cognitive Emotions: And Other Essays in the Philosophy of Education (pp.2-13). London: Routledge.

Tsang, N. M. (2013). Surprise in social work education. *Social Work Education: The International Journal, 32*(1), 55-67.

Upala, M. A., Gonce, L. O., Tweney, R. D. and Slone, D. J. (2007). Contextualizing counterintuitiveness: How context affects comprehension and memorability of counterintuitive concepts. *Cognitive Science, 31*, 415–439.

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. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*(4), 1006-1014.

Wilson, T. D., & Gilbert, D. T. (2008). Explaining away: A model of affective adaptation. *Perspectives on Psychological Science, 3*(5), 370-386.

Wong, P. T., & Weiner, B. (1981). When people ask" why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology*, *40*(4), 650-663.