Brod, Garvin; Hasselhorn, Marcus; Bunge, Silvia

When generating a prediction boosts learning. The element of surprise

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Learning and instruction 55 (2018), S. 22-31

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction, 55*, 22-31.

When generating a prediction boosts learning: The element of surprise

Garvin Brod[a,b], Marcus Hasselhorn[a,b], & Silvia A. Bunge[c]

[a]German Institute for International Educational Research (DIPF) & IDeA Center for Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany; [b]Department of Psychology, Goethe University Frankfurt, Frankfurt am Main, Germany; [c]Department of Psychology & Helen Wills Neuroscience Institute, University of California, Berkeley, USA

Correspondence concerning this article should be addressed to Garvin Brod, German Institute for International Educational Research, Schloßstr. 29, 60486 Frankfurt am Main, Germany. E-mail: garvin.brod@dipf.de

Abstract

Using both behavioral and eye-tracking methodology, we tested whether and how asking students to generate predictions is an efficient technique to improve learning. In particular, we designed two tasks to test whether the surprise induced by outcomes that violate expectations enhances learning. Data from the first task revealed that asking participants to generate predictions, as compared to making post hoc evaluations, facilitated acquisition of geography knowledge. Pupillometry measurements revealed that expectancy-violating outcomes led to a surprise response only when a prediction was made beforehand, and that the strength of this response was positively related to the amount of learning. Data from the second task demonstrated that making predictions about the outcomes of soccer matches specifically improved memory for expectancy-violating events. These results suggest that a specific benefit of making predictions in learning contexts is that it creates the opportunity for the learner to be surprised. Implications for theory and educational practice are discussed.

Keywords: knowledge activation, hypothesis generation, prediction error, memory, eye-tracking

1. Introduction

Activating students' prior knowledge has been identified as the cornerstone of high-quality instruction (Alexander, 1996; Ausubel, 1968; Bransford, Brown, & Cocking, 2000). Activating prior knowledge in the learner strongly improves their comprehension and memory of new material (Bransford & Johnson, 1972). Thus, a key question for educators is how to best activate relevant prior knowledge in their students. Various techniques to activate prior knowledge in students have been proposed (for an overview, see Krause & Stark, 2006). One technique is to ask students to make a prediction (also called 'generate a hypothesis') before receiving the new information. This technique has been successfully employed in studies that investigated ways to improve students' learning of various materials, including learning from text (Fielding, Anderson, & Pearson, 1990), physics (Champagne, Klopfer, & Gunstone, 1982; Crouch, Fagen, Callan, & Mazur, 2004; Inagaki & Hatano, 1977), and biology (Schmidt, De Voider, De Grave, Moust, & Patel, 1989).

It has been suggested that making a prediction requires accessing prior knowledge and connecting it to the new information being learned (Schmidt et al., 1989). Furthermore, it may stimulate curiosity for the correct answer (Inagaki & Hatano, 1977) and, if the answer was not correctly predicted, trigger conceptual change because the learners realize that there is a flaw in their concept (cf. Anderson, 1977, p. 427). Not surprisingly, then, asking students to make a prediction forms part of many prototypical instructional curricula (e.g., Champagne et al., 1982; Hardy, Jonen, Möller, & Stern, 2006).

However, despite its widespread use, very little is known about the mechanism(s) by which making a prediction may improve learning. In addition, a potential caveat to the prediction method is that students spend a lot of time and effort generating a prediction and might thus remember their wrong prediction instead of the correct result, as theorized by proponents of errorless learning (e.g., Baddeley & Wilson, 1994). Another caveat is that learners might not experience meaningful conflict despite having made a wrong prediction,

thereby leading to no conceptual change (Limón, 2001). Thus, knowledge of the specific mechanisms by which making a prediction affects learning seems crucial to resolve these opposing views.

A relevant line of work that has recently gained momentum in cognitive psychology research concerns the effects of guessing on learning. Kornell, Hays, and Bjork (2009) showed that testing can be beneficial for memory even during novel learning, when participants can only guess the answer and nearly all guesses are incorrect. They argued that this so-called *errorful* generation instantiates a special case of the well-known generation effect (Slamecka & Graf, 1978) and may promote learning because it requires great retrieval effort. Study methods that make use of this effect (e.g., flashcards) have been shown to substantially enhance memory retention (the so-called *testing effect*, see Karpicke & Roediger, 2008; Pyc & Rawson, 2009). Kornell et al.'s (2009) finding has sparked considerable interest and has been replicated and extended by various labs (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Potts & Shanks, 2014). A boundary condition that seems to be emerging from these studies is that, for guessing to be beneficial, timely corrective feedback is crucial, giving participants an opportunity to encode the correct answer (Vaughn & Rawson, 2012). Other than that, however, this line of work has focused mainly on the retrieval effort explanation as to why making a guess is beneficial for memory.

Another related line of work concerns the role of surprise – i.e., the emotional response to outcomes that do not match the prediction (see Ekman, 1992) – in enhancing learning. This work is grounded in now-classic research on reinforcement learning showing that discrepancies between what is expected and what occurs trigger learning (Rescorla & Wagner, 1972), as well as in a rich neuroscience literature suggesting that prediction errors play a universal role in driving learning throughout the human brain (for an overview, see Bar, 2007; Henson & Gagnepain, 2010). From a cognitive psychology perspective, Fazio and Marsh (2009) showed that increased attention is allocated to surprising

feedback, which then leads to better memory (see also Butterfield & Metcalfe, 2006). In line with this account, a recent study demonstrated that the degree to which expectancies are violated predicts later memory (Greve, Cooper, Kaula, Anderson, & Henson, 2017).

In a new line of work, Stahl and Feigenson (2015) demonstrated that 11-months-old infants show enhanced information-seeking and hypothesis-testing behaviors and learning for objects that appeared in episodes that violated expectations as compared to ones that were consistent with expectations. Recently, they demonstrated this benefit of surprise in children (aged 3–6) as well (Stahl & Feigenson, 2017). These findings led the authors to suggest that expectancy-violating events present special opportunities for learning. The facilitatory role of surprise for learning is in line with recent research showing that inducing confusion in a learner, for example by presenting contradictory information, leads to enhanced learning and transfer performance (D'Mello, Lehman, Pekrun, & Graesser, 2014). Confusion is suggested to occur after a surprise reaction when the expectancy-violating new information cannot be resolved right away, inducing a cognitive disequilibrium (D'Mello et al., 2014). In sum, this line of work has shown that expectancy-violating events can trigger learning, which might be due to the surprise response that is evoked by these events.

Based on these prior studies on surprise, we hypothesize that one specific mechanism by which making a prediction is beneficial for learning is that it enables a learner to be surprised by events that refute the prediction. Many processes that are known to improve learning, including effortful retrieval, self-generation of a solution, curiosity, and learning from feedback, are invoked when generating a prediction. Here, we sought to test whether predicting outcomes boosts subsequent learning when controlling for various potentially confounding factors.

Further, we sought to assess the extent to which surprise accounts for the benefit of prediction on learning. However, a common problem in research on surprise is how to measure and compare it across individuals, because asking participants to report their level of

surprise in response to an outcome is prone to systematic distortions (Schützwohl, 1998). One way to measure surprise objectively is via the pupillary response. Dilation of the pupil has been repeatedly shown to signal surprise (e.g., Kloosterman et al., 2015; Preuschoff, 't Hart, & Einhäuser, 2011) and reflects the release of the neurotransmitter norepinephrine in the brainstem's locus coeruleus, which regulates arousal (for an overview, see Aston-Jones & Cohen, 2005). Thus, surprise can be measured indirectly using pupillometry.

Here, we report the results of an experiment with two tasks involving university students. These experimental tasks probed different domains of knowledge, but both involved a within-subject experimental design that contrasted a condition in which participants had to make a prediction (henceforth called 'prediction condition') with a condition in which participants had to make a post-hoc evaluation (henceforth called 'postdiction condition').

The prediction and postdiction conditions differ only in the presentation order of the stimuli; participants have to state their expectations either before or after seeing the actual outcome (see Figure 1 for a graphical depiction of the study phase). Critically, both conditions require answering questions about the stimuli, and thus active engagement with the material and the activation of relevant prior knowledge. Better learning performance in the prediction condition would, thus, suggest that there are specific beneficial effects of generating a prediction that go beyond prior knowledge activation or active encoding. The current design represents a conservative test of the benefits of prediction for memory, because participants had considerably more time to encode the correct result in the postdiction condition (7.75 s instead of 3.5 s in the prediction condition), which they could use to engage in mnemonic strategies.

The first experimental task, referred to below as the geography task, tested whether asking participants to make predictions as to which of two countries has a larger population helps them to learn about European geography. The second experimental task, referred to below as the soccer task, tested how generating predictions about the result of a soccer match

affects memory for results that conform to or violate expectancies, based on prior knowledge about various German soccer teams' performance. The episodic nature of the task allowed us to directly test whether expectancy-violating events are better remembered in the prediction condition as compared to the postdiction condition.

We collected eyetracking data while participants performed these tasks, with a view to measuring pupil diameter with high temporal resolution over the course of a trial. Comparing pupillary response patterns for the prediction and postdiction conditions allowed us to test whether generating a prediction increases surprise about expectancy-violating outcomes and thereby enhances learning.

This study (including hypotheses, sampling, design, and analysis plan) was preregistered on the Open Science Framework (Brod, G., Bunge, S. A., & Hasselhorn, M. (2017, January 26, Does making a prediction improve memory? Retrieved from osf.io/v9fpu). In short, our main hypotheses were that generating a prediction would lead to a) better learning performance than post-hoc evaluation for both tasks, b) higher surprise about expectancy-violating outcomes, as indexed by a larger pupillary dilation, and c) the degree of this surprise reaction would be positively related to amount of learning via the updating of prior beliefs.

2. Methods

2.1 Participants

Thirty-six students of Goethe University Frankfurt (20 women; mean age 23.1 years; range 19–29) who gave written informed consent participated in the study. Sample size was determined *a priori* using G*Power with the following settings: t-test for dependent means, .05 alpha error, .90 power to detect a medium effect size of half a standard deviation (as found in pilot studies). Participants were recruited through bulletins within the university's psychology and education building and student email lists of these two departments. The

advertisements stressed that participants had to have at least some interest in soccer. The two experiments took 60–90 minutes in total and participants were paid €15 or received course credit for their participation. Ethics approval was obtained from the ethics committee of the German Institute for International Educational Research.

2.2 Study Design & Testing Procedures

2.2.1 Overview

Two computerized experimental tasks were performed in a single session: the geography task was performed first, followed by the soccer task. The two tasks were separated by a short break, during which participants could use the restroom. Each task took approximately 30 minutes, including an initial knowledge assessment, a computerized eyetracking task (the study phase), and a final assessment of knowledge or memory. After each experimental task, participants were given a brief questionnaire in which they had to indicate on a scale from 1–6 which of the two conditions (1= clearly prediction, 6 = clearly postdiction) they thought was more fun. Including the time required to provide instructions and calibrate the eye-tracker, this resulted in a total time of about 75 minutes to complete the whole experiment.

Both tasks included two conditions: one prompted participants to make a prediction about the outcome before the answer was revealed (prediction condition); the other presented the outcome first and then asked participants to indicate which outcome they would have predicted (referred to below as the postdiction condition). Thus, in the prediction condition, prior knowledge had to be activated before seeing the actual outcome, whereas in the postdiction condition, prior knowledge had to be activated after seeing the outcome. The two conditions were performed within-subjects in separate blocks, and differed only in the presentation order of the stimuli; participants had to state their expectations either before or after seeing the actual outcome (see Figure 1). Critically, total presentation time of the stimuli was identical in the two conditions, and both conditions required the activation of relevant

prior knowledge. In the next two sections, we describe the specifics of the task designs and study procedures.

2.2.2 Geography Task

2.2.2.1 Design

In the geography task, participants were asked, on each of a series of trials, to consider which of two countries had a larger population. The dependent measure was the change in hierarchy knowledge of the population size of European countries. Changes in knowledge were assessed via two knowledge tests, which consisted of rank ordering European countries by their number of inhabitants. Between these two assessments, participants completed the study phase, during which they made predictions for one block of trials, and post hoc evaluations for another block (Figure 1; see Procedures section for additional details). The length of the study phase (40 trials) was piloted to enable participants to gain knowledge of the countries' population sizes while not enabling them to memorize the exact number of inhabitants per country so that they could not merely remember the exact number of inhabitants per country but had to perform inference. Participants, thus, acquired relational knowledge of European country populations – that is, a hierarchical knowledge structure that contains a consistent mapping of elements and relations, which enables transitive inference (see Halford, Wilson, & Phillips, 2012). Both the study phase and the final knowledge test, thus, involved a relational reasoning component, as participants had to try to infer and remember relations between countries that followed a consistent, hierarchical structure (Alexander, 2016). Assignment of hierarchy to condition as well as the ordering of the conditions was counterbalanced across participants. This design enabled us to compare the improvement in hierarchy knowledge between the prediction and postdiction condition.

Two hierarchies were used, with 12 different countries each; one for the prediction and one for the postdiction condition. The 24 most populous European countries (not including Germany) were used for this experiment, and were distributed to the two hierarchies using an

odd/even procedure (see Appendix 1). In between pre- and post-test, the study phase was performed (see Figure 1), in which participants saw 40 unique pairs of countries that were taken from the current 12-country hierarchy. Given the limited number of potential pairings (66 in total), we used all of the adjacent countries (14 pairs) and odd/even countries (1–3, 2–4, etc, 13 pairs). The remaining 13 pairs were selected pseudo-randomly from the remaining potential pairings.

2.2.2.2 Testing Procedures

The testing session started and ended with the knowledge test, in which the participants were asked to rank order 2 decks of 12 European countries (represented by the flag and the name below) by number of inhabitants, starting with the country that they thought had the most inhabitants. After participants were finished sorting the first deck (no time limit was imposed), the cards were removed and participants repeated the procedure with the second deck of cards. The second deck always contained the twelve countries they saw in the first computer task. Before the computerized task blocks were administered, participants were given time to familiarize themselves with the flags, and they were made aware of the fact that no country names would be shown in the task. Participants were told that during the following study phase they would see pairs out of these 12 countries along with the correct population sizes. They were not told to memorize those numbers, but were instead informed that they would be asked to sort the cards again after the computerized study phase was finished and thus that they should try to figure out the correct rank order.

Each of the two blocks started with four practice trials to familiarize participants with the task (prediction or postdiction). Next, participants saw 40 unique pairs of countries. To facilitate learning of the hierarchy, participants saw only six of the twelve countries during the first half of each block; only during the second half of each block did they see all of the countries. In the prediction block, participants were instructed to predict which country of each pair had the greater population size, and to do so while the question marks appeared on

the screen (i.e., 'Response Phase', see Figure 1). Participants had to state their expectancy on a five-point scale (Far left: clearly the left country, Left: probably the left country, Middle: don't know, Right: probably the right country, Far right: clearly the right country). The same scale was used in the postdiction condition, in which participants were instructed to make a post-hoc evaluation ("What would you have expected?"). After each block was completed, participants sorted the respective 12 countries again, following the same instructions as during the first knowledge test. Between the end of each block and the knowledge test, participants performed a 30 sec distractor task, which was counting backwards from 200 by sevens/threes as rapidly as possible.

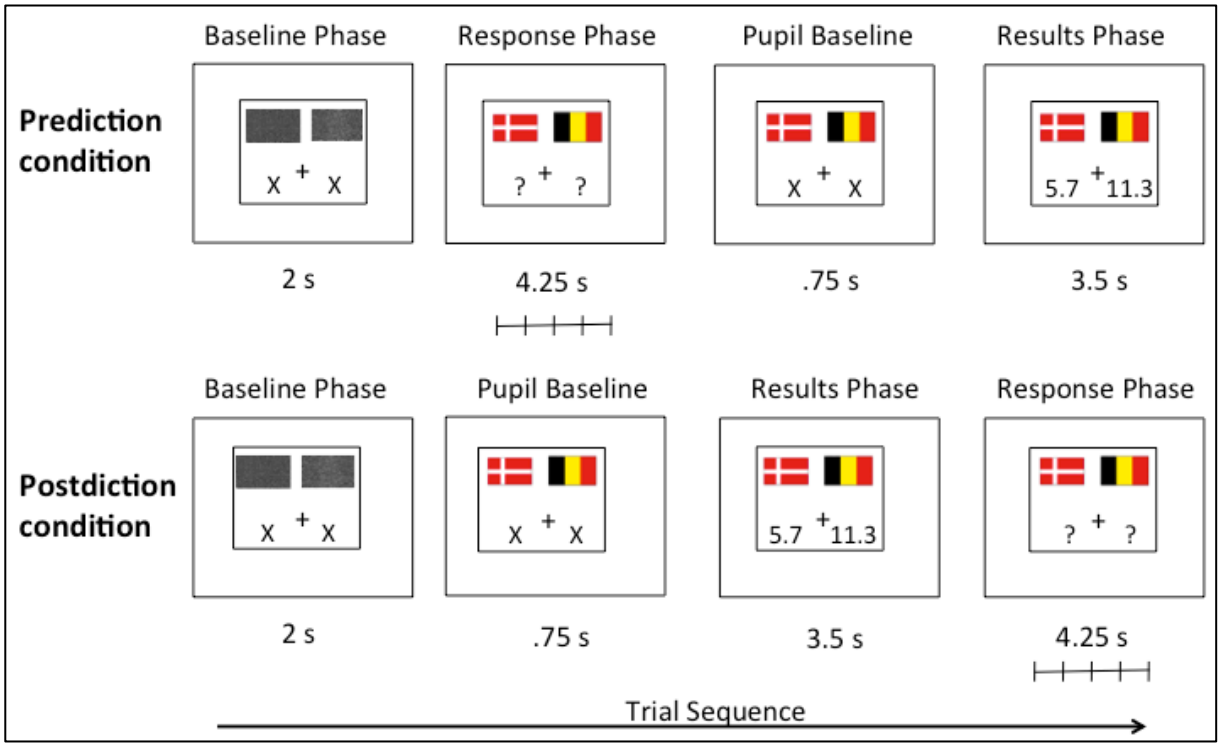

Figure 1. Schematic overview of the common study phase of the two paradigms, exemplified by the geography task. One exemplary trial is depicted per condition, which consisted of four different slides presented in the depicted order (duration times per slide are presented below the screens). In the prediction condition (upper half), participants had to make a prediction first and then saw the correct population sizes (in millions), whereas in the postdiction condition, they first saw the population sizes and then had to make a post-hoc statement regarding which results they would have predicted. Participants were only able to respond when the question marks appeared on the screen, using the same five-point scale for both conditions: Far left: clearly the left country, Left: probably the left country, Middle: don't know, Right: probably the right country, Far right: clearly the right country). Details

regarding the purposes of the 'Baseline Phase' and 'Pupil Baseline' can be found in section 2.4. For illustrative purposes, the background is shown in white and the print in black. For the real experiment, the background was gray and the print was white, so as to reduce luminance contrasts. The following details were changed for the soccer task (not shown due to copyright regulations): country flags were replaced by club logos; country populations were replaced by scores; and the labels of the five-point scale were adapted to the scores: Far left: >1 goal difference victory for the left team, Left: 1 goal victory for the left team, Middle: draw, Right: 1 goal victory for the right team, Far right: >1 goal victory for the right team.

2.2.3 Soccer Task

2.2.3.1 Design

In the soccer task, participants were asked to consider which of two soccer teams had won a particular match, and by how many points. Akin to the geography task, the prediction and postdiction conditions were performed in different blocks. Each block consisted of a study phase followed by a test phase. The order of the blocks as well as the assignment of matches to blocks was counterbalanced across participants. In the study phase, participants had to predict/postdict the result of a soccer match between two teams of Germany's first division and were then provided with the actual result (see Figure 1, country flags were replaced by club logos and country populations by scores). Participants saw 30 unique pairs of soccer teams in each study phase. In the test phase, participants saw all 30 pairings again and had to state the actual results of the match. Match results were taken from real matches that took place during the 2014/15 season. Matches were drawn from match days 24 to 33 during the soccer season and randomly assigned to the lists, thus assuring unique matches and similar frequency of teams.

We hypothesized that taking real results and telling participants that the results were real should enhance the relevance of participants' prior knowledge (as noted previously, recruitment materials indicated that participants should be at least somewhat interested in soccer). However, since two teams always play twice against each other in the course of a season and the matches dated back two seasons, even participants with high soccer knowledge

could not know the actual results beforehand. This assumption was confirmed with a questionnaire administered after the experiment.

The dependent variable for the soccer task was the percentage of correctly retrieved results (i.e., correct differences), independent variables were condition (prediction/postdiction) and expectancy (match/violation). To assess participants' prior knowledge of the relative strengths of the 18 teams and to ensure familiarity with the stimulus material (they were shown the club logos and the names), a knowledge test was performed prior to the beginning of the soccer task, as described below.

2.2.3.2 Testing Procedures

First, participants were instructed to rank order the 18 teams of the 2014-15 season of Germany's premier soccer division by their final standing. They were then given time to familiarize themselves with the club logos, and were made aware of the fact that no club names would be shown during the computerized study phase. Before starting the study phase, they were told that they would now see real results of match from this season, which they should memorize for a subsequent memory test. No details were given regarding the specifics of the later memory test.

Procedure and instructions for the study phase were very similar to those for the geography task, i.e., the blocks also started with four practice trials and participants were instructed to predict / make a post-hoc evaluation regarding the likely outcome of the match. Participants again had to respond on a five-point scale: Far left: >1 goal difference victory for the left team, Left: 1 goal victory for the left team, Middle: draw, Right: 1 goal victory for the right team, Far right: >1 goal victory for the right team.

For the test phase, which followed shortly after the study phase, participants were told that they would now see all match pairs again and that they should try to recall the actual result of the match. They were instructed to answer using the same five-point scale that they had used during the study phase.

2.3 Stimulus Presentation & Eye-Tracking Data Acquisition

Subjects were seated about 68 cm from the screen in a dimly lit room. The eye-tracking camera (EyeLink 1000, SR Research, Osgoode, Ontario, Canada) was located below the computer screen and recorded continuously throughout both experiments at a frequency of 500Hz. Eye-tracking was performed to record changes in participants' pupil size in response to the presentation of the correct outcome (i.e., during the 'Results Phase', see Figure 1). The key measure was the difference in the pupillary response between outcomes that match expectancies and those that violate expectancies. This difference can be interpreted as a measure of the amount of surprise experienced by a participant, and can be compared between the prediction and postdiction conditions.

Since the pupil is highly reactive to changes in luminance as well as to eye movements, the design of the study phase had to be tailored to the measurement of changes in pupil size. First, the 'Baseline Phase' was luminance-matched to the subsequent slides of the trial by presenting reshuffled images of the club logos in which their original luminance was preserved. The 'Baseline Phase' was included to avoid carry-over effects in pupil size from the previous trial. Second, a short 'Pupil Baseline' phase was introduced right before the 'Results Phase'. In the 'Pupil Baseline' phase, participants saw the flags/logos alone for 750 msec before the results were displayed on the screen. This was done to increase comparability of pupil size changes in the 'Results Phase' between the prediction and postdiction conditions. We piloted the duration of this phase to make sure that it was short enough to prevent participants from forming a prediction in the postdiction condition, but long enough to allow the pupil to adapt to the image. Third, to eliminate the need for larger saccades, which would interfere with accurate measurement of pupil diameter, all stimuli were presented close to the center of the screen, within a marked square. Fourth, stimuli were presented against a gray background, and white print was used to reduce luminance contrasts.

Stimuli were presented using PsychoPy v1.8 (Peirce, 2007), an open-source application for conducting psychology experiments written in Python, and devices (including the eye-tracker) were controlled by the ioHub Event Monitoring Framework, a Python package.

2.4 Data Analyses

2.4.1 Performance data analysis

Data were analyzed using R (R Core Team, 2014). The α level was set at 0.05 throughout the analyses. For both tasks, an expectancy-violation was defined as a scale difference between expected and actual result of 2 or greater, which means that the actual result is also qualitatively different than the expected one.

For the geography task (Experiment 1), hierarchy knowledge was assessed by calculating the mean absolute difference between the estimated rank position and the true rank position. Thus, smaller differences represent greater knowledge. Improvement in hierarchy knowledge was defined as pretest - posttest accuracy. A within-subject t-test was calculated to test for condition differences in change in hierarchy knowledge. Two participants were excluded because they had exceptionally high prior knowledge of the country populations (defined as a mean absolute difference at pretest < 1), which left little room for improvement. Due to the fact that we did not anticipate such high prior knowledge and therefore did not specify this data exclusion criterion in the preregistration, we confirmed that including these participants in the analyses would not have altered the results reported below.

For the soccer task (Experiment 2), a repeated-measures ANOVA was performed with the percentage of correctly retrieved results (i.e., correct differences) as the dependent variable and condition (prediction, postdiction) as well as expectancy (consistent, violating) as within-subject factors. To be able to directly compare memory performance between

conditions for expectancy-consistent and expectancy-violating events, respectively, within-subject t-tests were performed for each event type. Six participants were excluded due to chance level performance (20%), as specified in the preregistration. This left 30 participants for the analyses (see Figure 3 for a graphical depiction of the results). Including the six participants with chance level performance would not have altered the significance of any of the results reported for the smaller sample.

To assess performance differences between conditions, it was necessary to eliminate floor and ceiling effects. The between-task differences in data exclusion criteria stem from the different natures of the two tasks. In the geography task, we needed to ensure that participants did not have such high prior knowledge as to make learning impossible. In the soccer task, we needed to ensure that participants did not perform the episodic memory task at chance levels. Questionnaire data were evaluated using a one sample t-test comparing participants' responses to the mean of the scale (3.5).

2.4.2 Eye-Tracking data analysis

For this study, we focused on the pupillometry data recorded during the study phase of the geography task. We originally sought to use the pupillometry data recorded during the study phase of the soccer task as well. However, in preparing to conduct these analyses, we found that the number of expectancy-violating trials was very low (mean: 8.5 trials per condition), which meant that eight participants did not even meet a liberal trial number criterion (> 5 trials per condition), and there were many participants with less than expectancy-violating 10 trials. This made us decide to not pursue pupillometry analyses in Experiment 2.

Pupil data were analyzed in R using itrackR (https://github.com/jashubbard/itrackR) along with self-developed analysis scripts. First, eye-tracking data and behavioral data were merged. Second, periods of blinks were removed and interpolated using cubic spline interpolation. Third, pupil data were epoched relative to the onset of the 'Results Phase'. To facilitate comparison of the pupillary response to seeing the actual outcome within and across

subjects, pupil data were normalized by subtracting the diameter at each time point from the average diameter during the final 400ms of each trial's 'Baseline Phase' and dividing by it. This results in a percentage signal change measure relative to the 'Baseline Phase'. With this normalization, any nonspecific effect that lasts longer than an individual trial (e.g., arousal, fatigue) cannot confound the results. The average percentage change in pupil diameter was calculated per participant across the full 'Results Phase' (3.5 s), separately for outcomes that were consistent with vs. violated expectancies, and separately for the prediction and postdiction condition.

To determine the pupil surprise response, the average percentage change in pupil diameter for expectancy-consistent outcomes was subtracted from the change in pupil diameter for expectancy-violating outcomes. T-tests were performed to determine statistical significance of the pupil surprise response in each condition and to test for condition differences. Finally, participants' pupil surprise responses were correlated with their improvements in hierarchy knowledge to determine whether surprise enhances learning of the hierarchy.

3. Results

3.1 Generating a prediction improves learning

In the geography task, participants strongly improved their hierarchy knowledge from pretest to posttest in both conditions (Prediction pre-test: 2.33 ± .62 (M ± SD); post-test: 0.90 ± .72; Postdiction pre-test: 2.12 ± .60; post-test: 1.11 ± .76). As is apparent in Figure 2A, a within-subject t-test revealed a stronger improvement in hierarchy knowledge for the prediction as compared to the postdiction condition ($t(33)=2.5$, $p = .01$, Cohen's $d = .497$). These results support our hypothesis that making a prediction benefits the updating of relational knowledge.

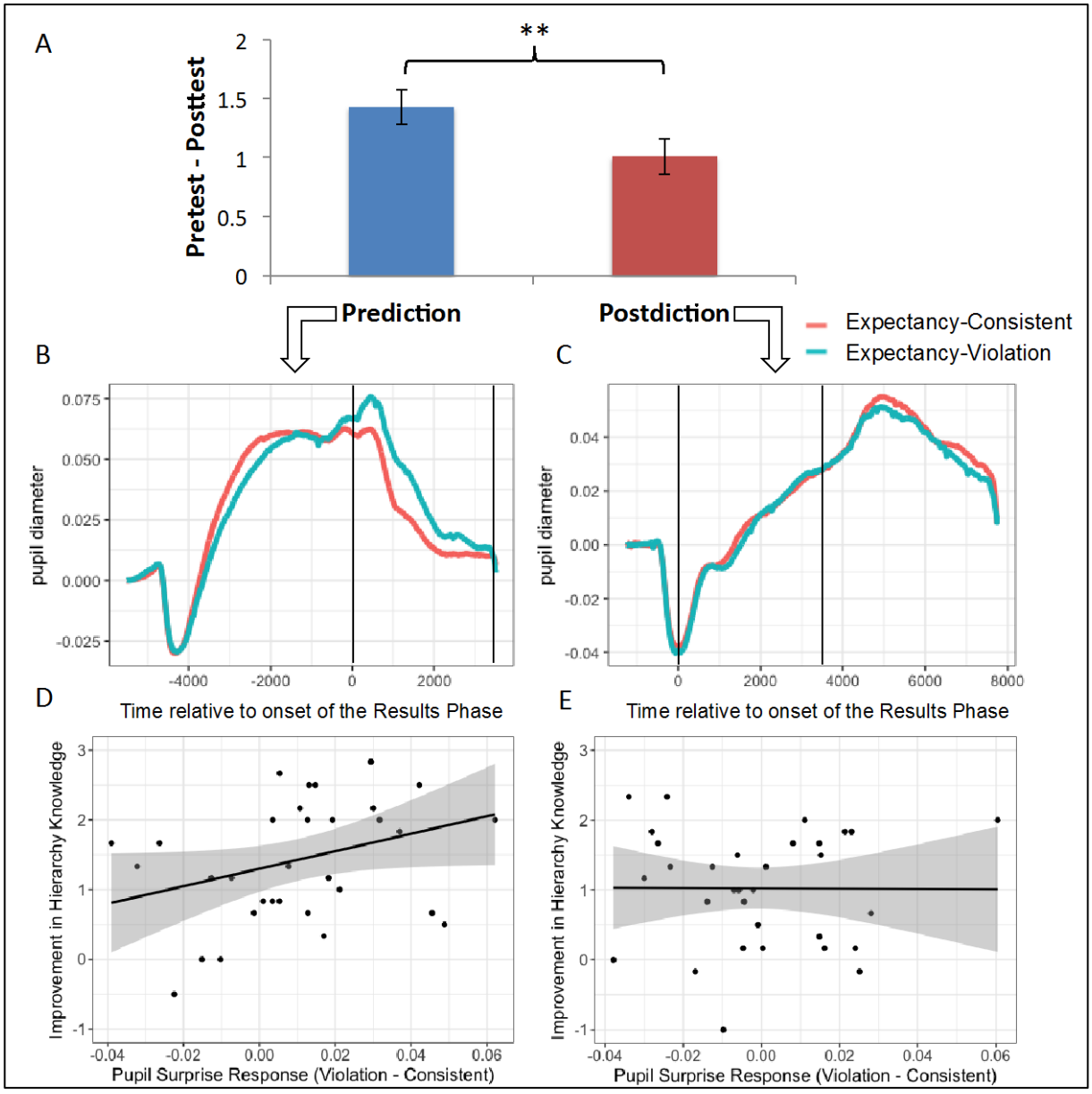

Figure 2. Geography Task Results. Panel A shows a greater increase in hierarchy knowledge in the prediction condition than in the postdiction condition. Error bars represent within-subject standard error. Panels B and C show the full time series of the pupillary response in the prediction (B) and postdiction condition (C), separately for expectancy-consistent and expectancy-violating outcomes. Black lines indicate the duration of the 'Results Phase'. Panels D and E show scatterplots relating the increase in hierarchy knowledge and the pupillary surprise response (expectancy-violating – expectancy-consistent during 'Results Phase'), separately for prediction (D) and postdiction (E) conditions.

For the soccer task, a repeated-measures ANOVA revealed no main effect of condition ($F(1, 29) = 1.03$, $p = .32$, $eta^2_G = .004$), but a main effect of expectancy ($F(1, 29) = 9.29$, $p = .005$, $eta^2_G = .09$), indicating better memory for events that were consistent with

expectancies. This effect was qualified by a condition x expectancy interaction ($F(1, 29) = 6.28$, $p = .018$, $eta^2_G = .04$). Follow-up t-tests indicated no memory differences between conditions for results that were consistent with expectancies ($t(29) = -1.56$, $p = .13$), but better memory in the prediction condition for expectancy-violating results ($t(29) = 2.34$, $p = .013$). These results support our hypothesis that generating a prediction boosts memory for expectancy-violating results, and suggest that it does not affect memory for expectancy-consistent results.

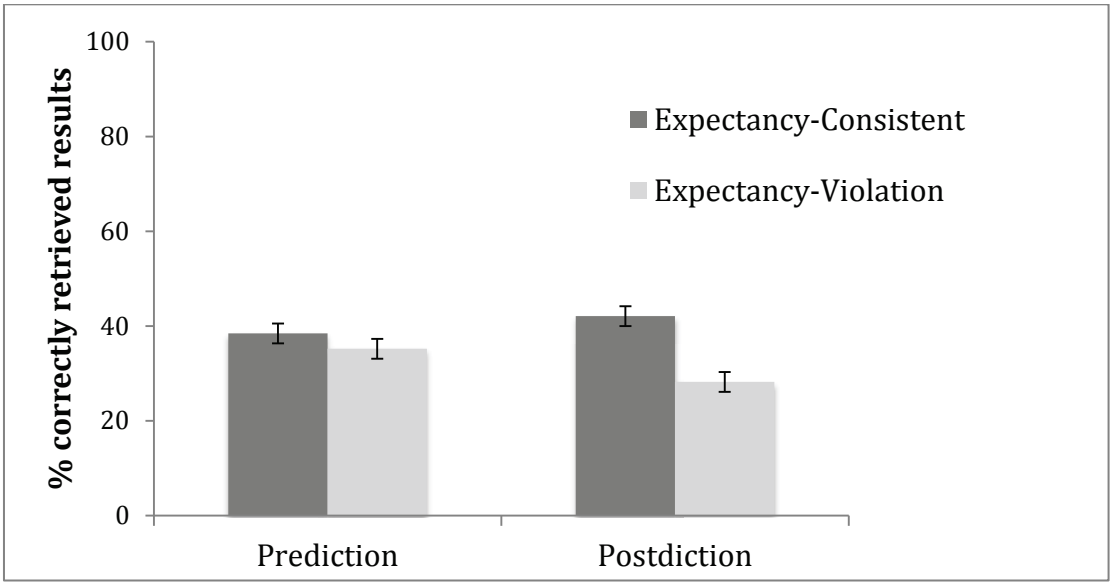

Figure 3. Soccer Task Results. Memory performance, separately for the prediction and postdiction condition, and for expectancy-consistent and expectancy-violating events. Error bars represent within-subject standard error.

Questionnaire data revealed that participants found the prediction condition more fun than the postdiction condition in both the geography task (mean = 1.88, t(33) = -7.1, $p < .001$) and the soccer task (mean = 1.77, t(29) = -7.1, $p < .001$).

3.2 Generating a prediction elicits surprise about expectancy-violating outcomes

Data from one participant were discarded because of a lack of expectancy-violating trials (criterion: > 5 usable trials per condition), leaving 33 participants for analyses. For illustrative purposes, we plotted the full time series of changes in pupil size for both conditions in Figure 2. The planned analyses focused on the pupillary response to seeing the outcome – i.e., the

'Results Phase'. In line with our predictions, the pupillary response was enhanced for expectancy-violating as compared to expectancy-consistent events in the prediction condition ($t(32) = 2.61$, $p = .007$). By contrast, there was no pupil surprise response in the postdiction condition ($t(32) = -.25$, $p = .81$). Accordingly, the pupillary violation of expectation response was greater in the prediction than in the postdiction condition ($t(32) = 1.85$, $p = .037$). These findings confirm that expectancy-violating events evoke a surprise response – but only when a prediction was made beforehand.

3.3 Surprise is associated with learning

To test whether surprise was associated with learning, we correlated participants' pupil surprise responses with their subsequent improvements in hierarchy knowledge (see Figure 2). For the prediction condition, this analysis revealed a positive correlation between the strength of the pupil surprise response and the degree of improvement in hierarchy knowledge ($r = .34$; $t(31)=2.0$, $p = .027$). For the postdiction condition, no such relationship was observed ($r = -.007$; $t(31)=-0.04$, $p = .971$). Thus, the pupillometry data suggest that generating a prediction increases surprise about expectancy-violating outcomes, and that the degree of surprise experienced by participants is positively related to their updating of relational knowledge.

4. Discussion

Asking students to generate a prediction is a popular technique for activating prior knowledge and improving student learning, probably because it entails many of the cognitive processes that are known to improve learning in general, including engaging in effortful retrieval, generating a solution to solve a problem, eliciting curiosity, and learning from feedback. Whether there are specific beneficial effects of generating predictions has been unclear, however. Here, we sought to test whether guessing the answer to a question is more conducive to learning than reflecting on the answer after it is revealed.

Our study included two tasks that tapped into different types of memory: the first a relational knowledge task that tested knowledge of the relative sizes of different European countries, and the second an episodic memory task that tested memory for the results of soccer matches. The first task allowed us to test whether making a prediction benefits updating knowledge about the relative differences in countries' populations, whereas the latter, testing memory for unique events, afforded a direct comparison between memory for expectancy-consistent and expectancy-violating events. We had hypothesized that one specific mechanism by which generating a prediction is beneficial for learning is that it enables a learner to be surprised by events that refute their prediction. Thus, we tested whether larger pupillary responses to unexpected outcomes would be associated with better learning. The results of the two tasks are not intended to be compared directly, but rather to be considered as complementary sources of evidence. Although the tasks were comparable in structure, they were not identical. For one thing, the nature of the pre- and post-tests were necessarily different, given the type of memory being probed. For another, the two tasks included different numbers of pairs (30 for the soccer task; 40 for the geography task). This difference emerged during piloting of the tasks and resulted from the different goals we had for the two tasks. For the soccer task, the goal was to keep episodic memory performance above chance. For the geography task, the goal was to enable the participants to gain knowledge of the countries' population sizes while not enabling them to memorize the exact number of inhabitants per country. Participants were asked to focus on the relative differences in population, and to use relational reasoning to infer the correct rank ordering of the countries.

In keeping with our hypotheses, we observed a greater extent of learning of the relative population sizes of European countries in the prediction than postdiction condition. Moreover, the pupillary surprise response to expectancy-violating events was present only in the prediction condition, and correlated positively with the improvement in relational

knowledge, as measured on the schema test. To ensure that the expectancy-violating events are the ones that benefit most from having made a prediction, we also collected learning success data on the individual trial level. These data, collected in the episodic memory paradigm, indeed revealed a specific memory benefit for expectancy-violating events in the prediction condition. No memory benefit was found for expectancy-consistent events. To conclude, findings of this study support our hypothesis that there is a specific benefit of prediction for learning, and that this effect is related in part to the surprise generated by expectancy-violating events. Furthermore, participants, who were students of education or psychology, found generating predictions to be more enjoyable than making post hoc judgments.

In our episodic memory task, overall memory was better for events that matched expectancies as compared to those that violated expectancies (even though the difference was not significant in the prediction condition). This result is in line with a rich literature on the memory congruency effect (see Brod, Werkle-Bergner, & Shing, 2013; Stangor & McMillan, 1992)), which is often observed in naturalistic memory tasks in which participants can successfully guess based on their prior knowledge (Bayen, Nakamura, Dupuis, & Yang, 2000). Guessing, then, benefits episodic memory accuracy for expectancy-matching events in the absence of true recollection. Due to this feature of expectancy-consistent outcomes, they are often excluded from further analysis. We chose to keep these events in the analyses to explore whether our condition manipulation also affected memory for expectancy-consistent outcomes, which was not the case. As a result of this null effect, overall memory performance did not differ between the prediction and postdiction condition in the episodic memory task.

This study contributes to an understanding of the specific mechanisms by which generating a prediction can improve learning. It suggests that generating a prediction enables the learner to be surprised about outcomes that refute the prediction, and that this surprise leads to an updating of knowledge structures. The elicited surprise, thus, makes generating

wrong predictions a productive exercise in failure (see also Kapur, 2016). It is worth noting that expectancy-violating outcomes do not in and of themselves seem to trigger surprise, as indicated by a lack of the pupil surprise response in the postdiction condition. Thus, explicit generation of a prediction seems necessary for surprise – and its beneficial effects – to occur.

The pupillary surprise response can be considered a proxy for physiological arousal that is induced by norepinephrine release in the cortex by neurons in the locus coeruleus of the brainstem (Aston-Jones & Cohen, 2005). Release of norepinephrine has been shown to promote long-term memory formation as well as behavioral and neural adaptation by interacting with other neuromodulators in the hippocampus (for a review, see (McGaugh & Roozendaal, 2009). On a cognitive level, this increased arousal likely increases attention to surprising outcomes (see Fazio & Marsh, 2009; Stahl & Feigenson, 2015). This enhanced attention may in turn lead to more effortful retrieval in attempting to resolve the incongruity, which is known to improve memory. Additionally or alternatively, this enhanced attention may induce a longer-lasting state of confusion (D'Mello & Graesser, 2014), which prompts more elaborative encoding (as suggested by D'Mello et al., 2014).

A useful framework for how surprise may trigger learning has been provided by Mandler's discrepancy theory (Mandler, 1990). The discrepancy theory posits that unpredicted outcomes result in a conflict between new information and existing schemata. This conflict/discrepancy results in an increase in arousal (i.e., a response by the autonomic nervous system) and a shift of attention to the discrepant information. According to Mandler (1990), the surprise response (i.e., the emotional reaction by an individual) can be interpreted as the initial, value-neutral consequence of this discrepancy. This suggestion is in line with the classification of surprise as an epistemic emotion (Pekrun & Stephens, 2012). Our findings provide support for these notions in that they demonstrate a pupillary surprise response to unpredicted outcomes, which was furthermore related to learning.

The present research also contributes to the long-standing debate among memory researchers as well as among social psychologists about the circumstances under which expectancy-consistent or expectancy-violating events are better remembered. Meta-analytic studies on the memory congruency effect (e.g., Stangor & McMillan, 1992) revealed several factors that influence whether expectancy-consistent or expectancy-violating events are better remembered. These factors include strength of expectancy, overall cognitive demands, participants' goals, and the ratio between expectancy-consistent and expectancy-violating events. We can now add another circumstance to this list, which is whether a specific prediction has been made prior to seeing the event. Making a prediction likely increases the strength of expectancy and, thereby, the perceived expectancy-violation and the surprise experienced by the learner, which then leads to better memory.

Results of these meta-analytic studies also suggest that there may be situations in which being asked to make a prediction is not beneficial, for example when overall cognitive demands are already high. This possibility could be tested in future pupillometry studies, given that pupils dilate as cognitive demands increase (Kahneman & Beatty, 1966; Van Gerven, Paas, Van Merriënboer, & Schmidt, 2004). It seems plausible to assume that the pupillary surprise response will be dampened if general arousal is high, but this hypothesis needs to be tested empirically.

On the whole, we do not mean to imply that asking students to generate a prediction is always the best way to activate prior knowledge and boost learning. First, there is simply a lack of studies directly comparing the effectiveness of different knowledge activation strategies. Second, generating a prediction is probably not feasible under all circumstances (e.g., for learning non-categorical information). Third, outcomes that were predicted incorrectly may not necessarily yield surprise in all learners, for all materials. As stressed by Limón (2001), instructional strategies that build upon inducing conflict in learners often fail in the classroom because the learners do not experience meaningful conflict. Our findings are

in line with this account as the observed lack of a surprise response in the postdiction condition indicates that the existence of a conflict is not enough to trigger a physiological trace of surprise, which in turn may be a prerequisite for experiencing conflict or confusion. Reasons for this lack of a surprise response could include a lack of interest, motivation, or prior knowledge. Thus, it seems reasonable to assume that generating a prediction will only yield a surprise response when at least basic levels of engagement and knowledge are present in the learner.

Aside from surprise, other factors could have contributed to the beneficial effect of prediction. This condition likely evoked curiosity (Inagaki & Hatano, 1977), which enhances motivation and is posited to boost learning (Kang et al., 2009). Students' self-reported higher enjoyment in the prediction condition is in accordance with this speculation. Also, one might argue that retrieval was less effortful in the postdiction condition because participants did not need to generate their own prediction but only had to assess the plausibility of the actual outcome given their prior knowledge. More difficult retrievals have been shown to lead to better memory than less difficult retrievals (Pyc & Rawson, 2009). While we cannot definitively rule out the possibility that differences in curiosity and effortful retrieval contributed to the observed benefits for the prediction condition, they are highly unlikely to have driven the condition differences entirely given the predictive value of the pupil surprise response and the results of the episodic memory task. In the latter task, we found a condition x expectancy interaction, suggesting that generating a prediction is only beneficial for remembering events that violated expectancies. Future studies might further explore the contributions of curiosity and effortful retrieval to the beneficial effects of prediction by assessing and/or manipulating these factors directly.

This study was intended as a first attempt to identify the specific mechanisms by which generating a prediction and experiencing surprise are beneficial for learning. Further studies are needed to establish the external validity of these findings, by testing whether the

beneficial effects are long-lasting and can be observed in more complex domains, such as conceptual change or scientific reasoning. Future studies should also compare generating predictions to other generative learning activities, such as providing examples or explanations (e.g., Endres, Carpenter, Martin, & Renkl, 2017; Legare & Lombrozo, 2014). Nevertheless, having found beneficial effects in two different domains makes us optimistic that they can be found across a wide range of situations. Another question that this study has raised but not answered is how participants' familiarity with the to-be-remembered items interacts with their perceived surprise about expectancy-violating events; to answer this question, pretest familiarity ratings at the item level would be necessary.

A further next step will be to test whether the observed benefits of generating a prediction also hold for school children. Inviting children to generate predictions may be a promising tool for fostering their scientific literacy, for at least two reasons: First, children do not spontaneously use memory strategies that activate their prior knowledge until around the end of elementary school (e.g., Hasselhorn, 1990). Thus, techniques that lead children to activate their knowledge may prove beneficial. Second, the cognitive conflict induced by a wrong prediction may help children overcome scientific misconceptions (Vosniadou, Ioannides, Dimitrakopoulou, & Papademetriou, 2001). In sum, it is important to follow up on these findings in school children, and using more school-related tasks.

Successful classroom curricula exist that incorporate generating predictions and receiving feedback (e.g., "Predict-Observe-Explain", see Champagne, Klopfer, & Anderson, 1980; Gunstone & White, 1981; Liew & Treagust, 1995). Generating predictions has also been integrated in interactive computer programs in which students can individually go through prediction–feedback cycles (e.g., genotype–phenotype relations, see Tsui & Treagust, 2003). It also forms part of many study methods that include testing or self-testing (e.g., flashcards). However, although the utility of asking students to generate a prediction as a means to activate their prior knowledge has been proposed before (cf. Anderson, 1977), its

specifics have remained opaque. We have observed a specific beneficial effect of prediction that does not form part of other generative learning techniques (e.g., generating examples or explanations), which is that generating a prediction allows for surprise. Evoking surprise in students has not been a major target in educational curricula thus far, even though already a single, surprising numerical fact can lead to long-lasting conceptual change (Clark & Ranney, 2010; Munnich, Ranney, & Bachman, 2005). Our results suggest that asking students to *put their cards on the table* by making a prediction seems a fruitful approach to harnessing the power of surprise. In addition, recent advances in mobile technologies (e.g., smartphones, tablets) will further simplify the way in which students' predictions can be collected and feedback can be provided.

In conclusion, this research demonstrates that there is a specific benefit of generating a prediction for learning, and that this benefit is at least in part mediated by the surprise generated by expectancy-violating events. It presents convergent evidence from several approaches – in this case, behavioral and psychophysiological – and has theoretical implications for our understanding of human cognition as well as practical implications for the development of good instructional practices and study habits. This work establishes a solid foundation for further research on this pedagogical tool, both in laboratory and classroom settings.

Acknowledgements

We thank Jasmin Breitwieser for help in collecting the data, and Maria Eckstein, Belén Guerra-Carrillo, and Ariel Starr from the Bunge laboratory as well as Dan Schwartz for valuable discussions. S.A.B. was supported by a James S. McDonnell Foundation Scholar Award and a Jacobs Foundation Research Fellowship.

References

Alexander, P. A. (1996). The Past, Present, and Future of Knowledge Research: A Reexamination of the Role of Knowledge in Learning and Instruction. *Educational Psychologist*, *31*(2), 89–92. http://doi.org/10.1080/00461520.1996.10524941

Alexander, P. A. (2016). Relational thinking and relational reasoning: harnessing the power of patterning. *npj Science of Learning*, *1*, 16004. http://doi.org/ 10.1038/npjscilearn.2016.4

Anderson, R. C. (1977). The notion of schemata and the educational enterprise: General discussion of the conference. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the Acquisition of Knowledge* (pp. 415–31). Lawrence Erlbaum.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive Gain and Optimal Performance. *Annual Review of Neuroscience*, *28*(1), 403–450. http://doi.org/10.1146/annurev.neuro.28.061604.135709

Ausubel, D. P. (1968). *Educational Psychology: A Cognitive View*. New York, NY: Holt, Rinehart and Winston of Canada Ltd. http://doi.org/10.1107/S010827019000508X

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*(1), 53–68. http://doi.org/10.1016/0028-3932(94)90068-X

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289. http://doi.org/10.1016/j.tics.2007.05.005

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. http://doi.org/10.1016/0885-2014(91)90049-J

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 717–726. http://doi.org/10.1016/S0022-5371(72)80006-9

Brod, G., Werkle-Bergner, M., & Shing, Y. L. (2013). The Influence of Prior Knowledge on

Memory: A Developmental Cognitive Neuroscience Perspective. *Frontiers in Behavioral Neuroscience*, *7*, 2011–2013. http://doi.org/10.3389/fnbeh.2013.00139

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69–84. http://doi.org/10.1007/s11409-006-6894-z

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics*, *48*(1980), 1074. http://doi.org/10.1119/1.12290

Champagne, A. B., Klopfer, L. E., & Gunstone, R. F. (1982). Cognitive research and the design of science instruction. *Educational Psychologist*, *17*(1), 31–53. http://doi.org/10.1080/00461528209529242

Clark, D., & Ranney, M. A. (2010). Known Knowns and Unknown Knowns : Multiple Memory Routes to Improved Numerical Estimation. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Proceedings of the 2010 International Conference of the Learning Sciences* (Vol. 1, pp. 460–467). International Society of the Learning Sciences, Inc.

Crouch, C., Fagen, A. P., Callan, J. P., & Mazur, E. (2004). Classroom demonstrations: Learning tools or entertainment?. *American journal of physics*, *72*(6), 835-838. https://doi.org/10.1119/1.1707018

D'Mello, S., & Graesser, A. (2014). Confusion. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International Handbook of emotions in education* (pp. 289–310). New York, NY: Routledge.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170. http://doi.org/10.1016/j.learninstruc.2012.05.003

Ekman, P. (1992). Are there basic emotions? *Cognition & Emotion*, *99*(3), 550–553. http://doi.org/10.1080/02699939208411068

Endres, T., Carpenter, S., Martin, A., & Renkl, A. (2017). Enhancing learning by retrieval: Enriching free recall with elaborative prompting. *Learning and Instruction*, *49*, 13–20. http://doi.org/10.1016/j.learninstruc.2016.11.010

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88–92. http://doi.org/10.3758/PBR.16.1.88

Fielding, L. G., Anderson, R. C., & Pearson, P. D. (1990). *How discussion questions influence children's story understanding. Technical Report No. 490. University of Illinois at Urbana-Champaign*.

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165. http://doi.org/10.1016/j.jml.2016.11.001

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. http://doi.org/10.3758/s13421-011-0174-0

Gunstone, R. F., & White, R. T. (1981). Understanding of gravity. *Science Education*, *65*(3), 291–299. http://doi.org/10.1002/sce.3730650308

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: the foundation of higher cognition. *Trends in cognitive sciences*, *14*(11), 497-505. http://doi.org/10.1016/j.tics.2010.08.005

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking." *Journal of Educational Psychology*, *98*(2), 307–326. http://doi.org/10.1037/0022-0663.98.2.307

Hasselhorn, M. (1990). The emergence of strategic knowledge activation in categorical clustering during retrieval. *Journal of Experimental Child Psychology*, *50*(1), 59–80. http://doi.org/10.1016/0022-0965(90)90032-4

Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, *20*(11), 1315–1326. http://doi.org/10.1002/hipo.20857

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*, 514–27. http://doi.org/10.3758/s13421-011-0167-z

Inagaki, K., & Hatano, G. (1977). Amplification of Cognitive Motivation and Its Effects on Epistemic Observation. *American Educational Research Journal*, *14*(4), 485–491. http://doi.org/10.3102/00028312014004485

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science (New York, N.Y.)*, *154*(3756), 1583–5. http://doi.org/10.1126/science.154.3756.1583

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of leraning. *Psychological Science*, *20*(8), 963–974. http://doi.org/10.1111/j.1467-9280.2009.02402.x

Kapur, M. (2016). Examining Productive Failure, Productive Success, Unproductive Failure, and Unproductive Success in Learning. *Educational Psychologist*, *51*(2), 289–299. http://doi.org/10.1080/00461520.2016.1155457

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science (New York, N.Y.)*, *319*(5865), 966–8. http://doi.org/10.1126/science.1152408

Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A. F., Bonneh, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience*, *41*(8), 1068–1078. http://doi.org/10.1111/ejn.12859

Kornell, N., Jensen Hays, M., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. http://doi.org/10.1037/a0015729

Krause, U.-M., & Stark, R. (2006). Vorwissen aktivieren. In H. Mandl & H. F. Friedrich (Eds.), *Handbuch Lernstrategien* (Vol. 1, pp. 38–49). Göttingen: Hogrefe.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212. http://doi.org/10.1016/j.jecp.2014.03.001

Liew, C. W., & Treagust, D. F. (1995). A predict-observe-explain teaching sequence for learning about students' understanding of heat and expansion of liquids. *Australian Science Teachers' Journal*, *41*(1), 68–71.

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and Instruction*, *11*(4–5), 357–380. http://doi.org/10.1016/S0959-4752(00)00037-2

Mandler, G. (1990). A constructivist theory of emotion. In N. L. Stein, B. Leventhal, & T. R. Trabasso (Eds.), *Psychological and Biological Approaches to Emotion* (pp. 21–45). Hillsdale, NJ: Lawrence Erlbaum.

McGaugh, J. L., & Roozendaal, B. (2009). Drug enhancement of memory consolidation: Historical perspective and neurobiological implications. *Psychopharmacology*, *202*(1–3), 3–14. http://doi.org/10.1007/s00213-008-1285-6

Munnich, E. L., Ranney, M. A., & Bachman, M. L. N. (2005). The Longevities of Policy-Shifts and Memories Due to Single Feedback Numbers. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1553–1558). Mahwah, NJ: Erlbaum. Retrieved from http://csjarchive.cogsci.rpi.edu/proceedings/2005/docs/p1553.pdf

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. http://doi.org/10.1016/j.jneumeth.2006.11.017

Pekrun, R., & Stephens, E. J. (2012). Academic emotions. *APA Educational Psychology Handbook: Individual Differences and Cultural and Contexual Factors*. http://doi.org/10.1037/13274-000

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of*

Experimental Psychology: General, *143*(2), 644–667. http://doi.org/10.1017/CBO9781107415324.004

Preuschoff, K., ’t Hart, B. M., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline’s role in decision making. *Frontiers in Neuroscience*, *5*(SEP), 1–12. http://doi.org/10.3389/fnins.2011.00115

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. http://doi.org/10.1016/j.jml.2009.01.004

R Core Team. (2014). R: A Language and Environment for statitical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*(6), 64–99. http://doi.org/10.1101/gr.110528.110

Schmidt, H. G., De Voider, M. L., De Grave, W. S., Moust, J. H. C., & Patel, V. L. (1989). Explanatory Models in the Processing of Science Text: The Role of Prior Knowledge Activation Through Small-Group Discussion. *Journal of Educational Psychology*, *81*(4), 610–619. http://doi.org/10.1037/0022-0663.81.4.610

Schützwohl, A. (1998). Surprise and schema strength. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*(5), 1182–1199. http://doi.org/10.1037/0278-7393.24.5.1182

Slamecka, N. J., & Graf, P. (1978). The Generation Effect: Delineation of a Phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants’ learning and exploration. *Science*, *348*(6230), 91–94. http://doi.org/10.1126/science.aaa3799

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition*, *163*, 1–14. http://doi.org/10.1016/j.cognition.2017.02.008

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin*, *111*(1), 42–61. http://doi.org/10.1037//0033-2909.111.1.42

Tsui, C.-Y., & Treagust, D. F. (2003). Genetics Reasoning withMultiple External Representations. *Research in Science Education*, *33*(Figure 1), 111–35. http://doi.org/10.1023/A

Van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology*, *41*(2), 167–174. http://doi.org/10.1111/j.1469-8986.2003.00148.x

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*(5), 899–905. http://doi.org/10.3758/s13423-012-0276-0

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction*, *11*(4–5), 381–419. http://doi.org/10.1016/S0959-4752(00)00038-4

Figure Captions

Figure 1. Schematic overview of the common study phase of the two paradigms, exemplified by the geography task. One exemplary trial is depicted per condition, which consisted of four different slides presented in the depicted order (duration times per slide are presented below the screens). In the prediction condition (upper half), participants had to make a prediction first and then saw the correct population sizes (in millions), whereas in the postdiction condition, they first saw the population sizes and then had to make a post-hoc statement regarding which results they would have predicted. Participants were only able to respond when the question marks appeared on the screen, using the same five-point scale for both conditions: Far left: clearly the left country, Left: probably the left country, Middle: don't know, Right: probably the right country, Far right: clearly the right country). Details regarding the purposes of the 'Baseline Phase' and 'Pupil Baseline' can be found in section 2.4. For illustrative purposes, the background is shown in white and the print in black. For the real experiment, the background was gray and the print was white, so as to reduce luminance contrasts. The following details were changed for the soccer task (not shown due to copyright regulations): country flags were replaced by club logos; country populations were replaced by scores; and the labels of the five-point scale were adapted to the scores: Far left: >1 goal difference victory for the left team, Left: 1 goal victory for the left team, Middle: draw, Right: 1 goal victory for the right team, Far right: >1 goal victory for the right team.

Figure 2. Geography Task Results. Panel A shows a greater increase in hierarchy knowledge in the prediction condition than in the postdiction condition. Error bars represent within-

subject standard error. Panels B and C show the full time series of the pupillary response in the prediction (B) and postdiction condition (C), separately for expectancy-consistent and expectancy-violating outcomes. Black lines indicate the duration of the 'Results Phase'. Panels D and E show scatterplots relating the increase in hierarchy knowledge and the pupillary surprise response (expectancy-violating – expectancy-consistent during 'Results Phase'), separately for prediction (D) and postdiction (E) conditions.

Figure 3. Soccer Task Results. Memory performance, separately for the prediction and postdiction condition, and for expectancy-consistent and expectancy-violating events. Error bars represent within-subject standard error.

Appendix 1. Country lists used in the experiment.

List 1	Population (in million)	List 2	Population (in million)
France	66.35	Great Britain	64.77
Italy	60.8	Spain	46.44
Poland	38.01	Romania	19.86
Netherlands	16.9	Belgium	11.26
Greece	10.81	Czech Republic	10.54
Portugal	10.37	Hungary	9.85
Sweden	9.75	Austria	8.58
Switzerland	8.08	Serbia	7.1
Bulgaria	7.2	Denmark	5.66
Finland	5.5	Slovakia	5.42
Norway	5.08	Ireland	4.63
Croatia	4.23	Lithuania	2.92