Tackling Scientific Misconceptions: The Element of Surprise

Maria Theobald [iD]
DIPF Leibniz Institute for Research and Information in Education and Center for Individual Development and Adaptive Education of Children at Risk (IDeA)

Garvin Brod [iD]
DIPF Leibniz Institute for Research and Information in Education and Center for Individual Development and Adaptive Education of Children at Risk (IDeA) and Goethe University Frankfurt

Misconceptions about scientific concepts often prevail even if learners are confronted with conflicting evidence. This study tested the facilitative role of surprise in children's revision of misconceptions regarding water displacement in a sample of German children ($N = 94$, aged 6–9 years, 46% female). Surprise was measured via the pupil dilation response. It was induced by letting children generate predictions before presenting them with outcomes that conflicted with their misconception. Compared to a control condition, generating predictions boosted children's surprise and led to a greater revision of misconceptions ($d = 0.56$). Surprise further predicted successful belief revision during the learning phase. These results suggest that surprise increases the salience of a cognitive conflict, thereby facilitating the revision of misconceptions.

Teaching science is challenging because it entails changing students' naïve theories about the world. Prominent examples include children's misconceptions about buoyancy (Potvin, Masson, Lafortune, & Cyr, 2015) or about solids and liquids (Babai & Amsterdamer, 2008). Misconceptions are persistent because they provide plausible explanations for everyday phenomena (Vosniadou & Ioannides, 1998), and because they can coexist alongside the accepted scientific concept (Shtulman & Valcarcel, 2012). Following the framework by Chi (2008, 2013), misconceptions subsume false beliefs and flawed mental models. False beliefs refer to misconceptions at the level of a single idea, while flawed mental models contain a network of false beliefs. In either case, overcoming a misconception requires learners to revise (one or many) false beliefs, which is a cumbersome process (Chi, 2013; Vosniadou, 2019).

A prominent instructional tool to facilitate belief revision is to induce a cognitive conflict between the misconception and the scientific concept (Posner, Strike, Hewson, & Gertzog, 1982). Typically, this is done by confronting learners with evidence that is in conflict with their misconception. While intuitively plausible, a wealth of research has found that conflicting evidence does not suffice to change the learner's misconception (see Limón, 2001). Instead, learners interpret conflicting evidence as an exception and ignore or only slightly modify their misconception (Chinn & Brewer, 1993). For instance, children frequently have the misconception that the mass of an object determines how much water it displaces when fully immersed in water (e.g., Dawson & Rowell, 1984; Linn & Eylon, 2000; Piaget & Inhelder, 1941). It has been shown that children often stick with their misconception even if they are repeatedly confronted with conflicting evidence and even if they are explicitly told that the volume of an object and not its weight

determines water displacement (Burbules & Linn, 1988). Hence, simply presenting conflicting evidence seems not to suffice to overcome a misconception.

Cognitive developmental research suggests that learners leverage conflicting evidence if the evidence clearly violates their beliefs. For instance, it has been shown that children who already have an initial belief only revise it if they cannot "explain away" the conflicting evidence (Bonawitz, van Schijndel, Friel, & Schulz, 2012). This result suggests that the extent to which children revise their beliefs depends on the perceived strength of the conflicting evidence. Another study found that this is true for higher-level belief revision as well, and that it depends also on the strength of the prior belief (Kimura & Gopnik, 2019): Children with weak prior beliefs revise their beliefs faster than children with strong prior beliefs. In sum, previous results from developmental research suggest that children's belief revision depends both on the strength of the conflicting evidence and on the strength of the prior belief, in accordance with Bayesian learning.

Thus far, research on the effects of violations of expectations on belief revision has largely overlooked the role that emotions may play in this process. Epistemic emotions are considered to be of particularly high relevance for science learning because they occur while processing conflicting information (Sinatra, Broughton, & Lombardi, 2014; Vogl, Pekrun, Murayama, & Loderer, 2020). Surprise can be conceptualized as the initial emotional response to a violation of expectations (Noordewier, Topolinski, & Van Dijk, 2016). Surprise, in turn, triggers curiosity or leads to confusion if the conflict remains unresolved (Vogl et al., 2020). Hence, surprise may serve as a starting point to engage in deeper processing of expectancy-violating information.

Physiologically, surprise has been shown to go along with a short burst in physiological arousal, as indexed by a pupil dilation response (PDR, e.g., Kloosterman et al., 2015; Preuschoff, t Hart, & Einhauser, 2011). Already 6-month-old infants show a PDR to events that violate their expectations (Zhang, Jaffe-Dax, Wilson, & Emberson, 2018). The PDR has been shown to be sensitive to violations of more complex beliefs as well, such as in Theory of Mind tasks (Dörrenberg, Rakoczy, & Liszkowski, 2018). Age-comparative research suggests that the PDR to expectancy-violating events constitutes a reliable, age-invariant measure (Krüger, Bartels, & Krist, 2020). In sum, these results suggest that the PDR to expectancy-violating events constitutes a good marker of surprise.

The PDR to expectancy-violating events has been linked to successful belief revision already. It has recently been shown that the average magnitude of the PDR to expectancy-violating events predicts belief revision (Brod, Hasselhorn, & Bunge, 2018). In this study, university students had to predict the results of soccer games or predict which of two countries has the larger population size. Students with a larger PDR to expectancy-violating results showed higher learning gains from pretest to posttest. These findings suggest that surprise drives cognitive processes related to learning. However, this study focused on the interindividual relation between the average PDR aggregated over all expectancy-violating events and learning gains from pre- to posttest. This type of analysis cannot capture belief revision as an intraindividual, accruing process that occurs in response to repeated expectancy-violating events. In other words, it remains unclear whether surprise about an unexpected event predicts beliefs revision on a trial-by-trial level throughout the learning phase. There is, thus, a clear lack of research on how intraindividual differences in learners' response to conflicting evidence are related to successful belief revision.

How could surprise facilitate belief revision? It is assumed that surprise leads to an interruption of ongoing cognitive processes and directs attention toward the unexpected outcome (Reisenzein, Horstmann, & Schützwohl, 2019). In line with this claim, it has been shown that surprising feedback leads to increased attention to and learning of unexpected information (Fazio & Marsh, 2009). Moreover, it has been shown that surprise triggers curiosity which is associated with exploratory behavior (Vogl et al., 2020). Exploration and the search for alternative explanations for the conflicting evidence could in turn benefit belief revision (Bonawitz et al., 2012; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2017). Hence, surprise could facilitate belief revision by helping learners to focus on and remember alternative explanations for the conflicting evidence.

In this study, we examined whether surprise facilitated the correction of misconceptions in the domain of water displacement. As mentioned earlier, a common misconception among children is that the weight of an object determines how much water it will displace (Burbules & Linn, 1988; Dawson & Rowell, 1984; Piaget & Inhelder, 1941). The scientifically correct answer, however, is that only the size of the object determines how much water it will replace when it is forcibly held under water.

Furthermore, the concept of water displacement is closely related to the concept of buoyancy. To evaluate buoyancy, children need to understand the concepts of mass and volume and their respective influence on water displacement. The understanding of the concept of water displacement, thus, constitutes an important prerequisite to develop a coherent understanding of the more complex concept of buoyancy (see Hardy, Jonen, Möller, & Stern, 2006).

Throughout our experiment, children saw pairs of spheres and indicated which sphere displaces more water. Children generated predictions or gave post hoc expectancy ratings (i.e., postdictions) before seeing the correct outcome. Surprise was measured via the PDR to expectancy-violating outcomes. Based on previous findings (Breitwieser & Brod, 2021; Brod, Breitwieser, Hasselhorn, & Bunge, 2020; Brod et al., 2018), we assumed that the PDR to expectancy-violating outcomes would be stronger if a prediction was made beforehand. The design, thus, yields an indirect experimental manipulation of surprise intensity.

We hypothesized that, compared to children in the postdiction condition, children in the prediction condition would exhibit a greater PDR to expectancy-violating outcomes and greater belief revision, as indicated by better performance in a transfer test. We further hypothesized an intraindividual, positive link between surprise and subsequent belief revision in the prediction condition.

Method

Participants

We tested 94 six- to nine-year-old children (M_{Age} = 8.00, SD_{Age} = 0.96; 46% female) who were randomly assigned to a prediction (n = 48) or postdiction condition (n = 46). The target sample size (n = 84) was determined using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) with the following settings: Difference between two independent means, effect size d = 0.65 (based on pilot results), α = .05, β = .90. We tested more children than the target sample size to compensate for the exclusion of children who did not exhibit a misconception in the pretest. Eight children were excluded because they correctly solved at least 7 of 8 trials in the pretest (prediction: n = 2; postdiction: n = 6). The randomly assigned groups (prediction vs. postdiction) did not differ regarding their average age (t (92) = −.86, p = .394). Children's age was further unrelated to pretest performance (r = −.06,

p = .548), posttest performance (r = −.04, p = .747), and transfer test performance (r = .12, p = .284).

Participants were recruited by student assistants in a natural history museum in Germany. The museum is one of the largest natural history museums in Europe and attracts almost 400 thousand visitors a year from all over the country, many of them children. Parents gave written informed consent prior to testing. Children received a small gift for their participation. Ethics approval was obtained from the ethics committee of DIPF, Leibniz Institute for Research and Information in Education.

Design

Children participated in a pretest, learning phase, posttest, and transfer test, in this order (see Figure 1). For the learning phase, children were randomly assigned to a prediction or postdiction condition (between-subject design). Children in the prediction condition predicted the outcome of the learning task before they were presented with the correct solution. Children in the postdiction condition first saw the correct solution of the learning task and only then stated which outcome they would have predicted. This experimental manipulation was adapted from Brod et al., (2018) who (using different stimuli) demonstrated that the prediction condition elicited a pupillary surprise response whereas the postdiction condition did not. This experimental manipulation, thus, promised to be an indirect experimental manipulation of surprise intensity during the learning phase.

To examine children's correction of misconceptions, several dependent variables were assessed: (a) performance gains from pretest to posttest, (b) performance in a transfer test, and (c) performance during the learning phase (see behavioral analyses). Furthermore, pupil size was recorded throughout the learning phase to assess children's pupillary surprise response (see pupillary analyses).

Procedure

Children were tested individually in a quiet room and saw a short introductory video clip that introduced the water displacement experiment. The clip showed an experimenter demonstrating how water gets displaced by pressing a sphere under water. The experimenter stressed that the spheres were held under water to avoid that children evaluate buoyancy instead of water displacement.
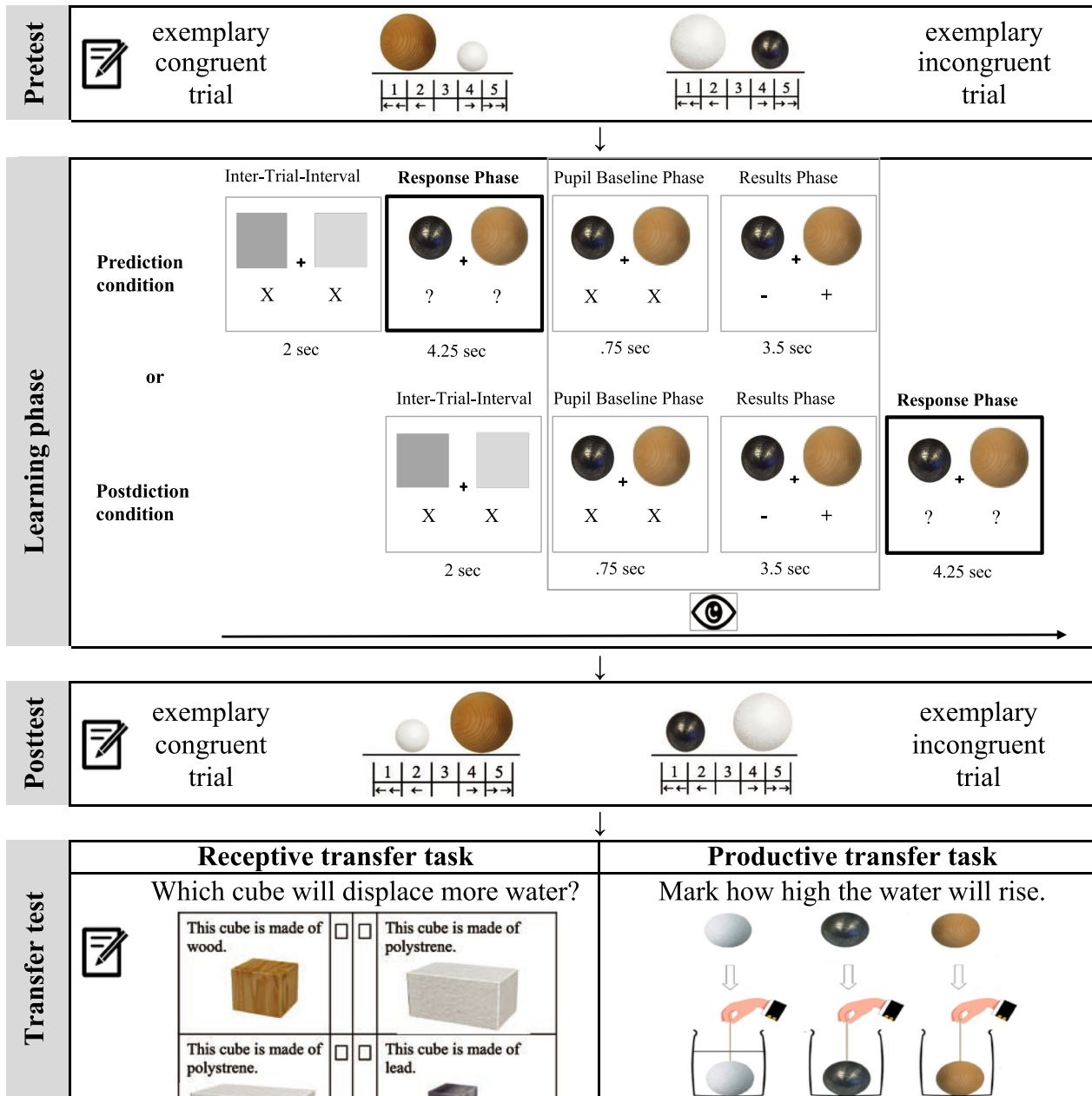
Figure 1. Schematic overview of the experimental procedures. The study followed a pretest–learning phase–posttest–transfer test design. The learning phase consisted of 34 trials that were presented in one of two conditions (between-subjects): prediction or postdiction. The figure shows an exemplary incongruent trial in both conditions. Children in the prediction condition stated their expectation about which sphere displaces more water before results were presented. Children in the postdiction condition first saw the results and only then stated their expectation.

Pretest

Children's concepts of water displacement were assessed in a paper-pencil pretest (eight trials, McDonald's ω = .81). Children were presented with pairs of spheres that varied in material and, thereby, weight (polystyrene, wood, lead) as well as in size (small, medium, large; stimuli were adapted from Potvin, Sauriol, & Riopel, 2015). A common misconception about water displacement is that the weight instead of the size of the object determines how much water will be displaced. In congruent trials, the common misconception (heavier object displaces more water) led to the correct

solution because the heavier sphere was also the larger one, while in incongruent trials the misconception led to the wrong solution. For each trial, children stated whether they expected the left sphere or the right sphere to displace more water, or whether they thought that both spheres displace an equal amount of water. Expectations were assessed on a 5-point-scale (1 = *clearly the left sphere*, 2 = *rather the left sphere*, 3 = *spheres displace an equal amount of water*, 4 = *rather the right sphere*, 5 = *clearly the right sphere*).

Learning Phase

Children were randomly assigned to a prediction or a postdiction condition in which they performed a computerized trial-and-error learning task. The learning task encompassed the same kind of stimuli as in the pretest. The learning phase started with eight practice trials (only congruent trials) followed by a pseudorandomized sequence of 10 congruent and 16 incongruent trials. Depending on the experimental condition, children stated their expectation about which sphere displaces more water either before seeing the correct answer ("Which of the two spheres displaces more water?"; prediction condition) or as a post hoc judgment ("Which outcome would you have predicted, independent of the actual result?"; postdiction condition). Importantly, the way in which the stimuli were presented as well as the duration of the presentation was identical for both conditions; the only difference was the order in which these stimuli were presented (see Figure 1).

Posttest

After the learning phase, children completed a paper-pencil posttest (eight trials, McDonald's ω = .92). The posttest was identical to the pretest but with the position of the spheres (left/right) being switched.

Transfer Test

Finally, a paper-pencil transfer test (8 trials, McDonald's ω = .92) was administered (see Figure 1). The transfer test differed from the pretest, learning phase, and posttest because it included other kinds of stimuli. Hence, the transfer test served to test whether children could apply their newly acquired model of water displacement to different tasks. The transfer test included a receptive and a productive transfer task. The receptive transfer task (six trials) was presented first and was similar to the pre- and posttest whereby spheres were replaced by cubes. For each pair of cubes, the child had to decide which cube would displace more water. A correct answer was dummy-coded with "1". Hence, overall, children could earn 6 points in the first transfer task. The productive transfer task was presented second. In the productive transfer task, children indicated how high the water level in a glass would rise for spheres of different size and weight that were fully immersed in water. Children were presented with sets of spheres. The first set included three spheres and is displayed in Figure 1. The second set included four different spheres. In each set, the water level for one sphere was already shown, which served as a baseline. Children then had to mark how high the water level would rise for the other spheres relative to this baseline water level. All possible pairwise comparisons within each item set were dummy-coded in terms of accuracy. For instance, in the set displayed in Figure 1, the water level of the leaden sphere and the wooden sphere were each compared to the baseline water level of the polystyrene sphere as well as among each other. This procedure, thus, yielded three comparisons in the first set of spheres and six comparisons in the second set. Hence, overall, children could earn 9 points in the productive transfer task.

After the transfer test, children performed a brief executive functions task (Hearts and Flowers Task). The executive functions task was included because it could explain interindividual differences in belief revision. We do not report results of the executive function task in this manuscript because of our focus on intraindividual processes of belief revision. The whole experiment took about 30 min in total.

Stimulus Presentation and Eye-Tracking Procedure

Stimuli were presented using PsychoPy v1.83.03 (Peirce et al., 2019). Children were seated about 68 cm from the computer screen in a dimly lit room without windows. The eye-tracking camera (Eye-Link 1000; SR Research, Osgoode, Ontario, Canada) was placed below the computer screen and recorded continuously throughout the learning phase at a frequency of 500 Hz. Eye-tracking was performed to record children's pupil size when seeing the correct result. The pupil size is highly reactive to changes in luminance and eye movements. Therefore, a luminance-matched "Inter-Trial-Interval" was presented at the beginning of each trial to avoid carry-over effects from the previous

trial. In addition, a short "Pupil Baseline Phase" was included 750 ms prior to the "Results Phase." This was done to enhance comparability of pupil size changes in response to seeing the correct result between the prediction and postdiction condition. The duration (750 ms) was deliberately kept short to avoid that children in the postdiction condition could generate a prediction as well (see Brod et al., 2018, 2020). Throughout the whole trial sequence, a fixation cross was presented to guide the children's view to the center of the screen.

Data Analyses

All analyses were carried out using R (R Core Team, 2019). Significance levels were set at .05 throughout the analyses. We conducted confirmatory, one-sided tests for our directional hypotheses that (a) children in the prediction condition would show better performance in the transfer test compared to children in the postdiction condition, (b) generating a wrong prediction would evoke a pupillary surprise response, and (c) the pupillary surprise response would predict belief revision in the prediction condition.

For the behavioral analyses, we excluded 6 children with implausible performance in the posttest or transfer test (prediction: $n = 3$; postdiction: $n = 3$). These children scored more than 3 SD below average and showed no performance increase or even reduced performance from pretest to posttest. We ensured that including those outliers did not alter our main results (see Supporting Information for robustness checks). Hence, the sample of the behavioral analyses comprised 80 children ($n = 43$ prediction, $n = 37$ postdiction). For the pupillary analysis, data from an additional three children were excluded due to erroneous data recording, mostly because the eye tracker focused on children's glasses instead of on their pupil. Thus, the reduced sample for pupillary analyses comprised 77 children ($n = 41$ prediction, $n = 36$ postdiction).

Each child in the prediction condition had on average three unexpected, incongruent trials ($M = 2.63$, $SD = 2.42$). However, seven children had to be excluded from this analysis as they did not have any unexpected, incongruent trials, leaving 34 children for the analysis. The lack of unexpected (i.e., incorrectly answered) incongruent trials does not fit with the observation that those children had misconceptions in the pretest. The time restrictions for making a prediction might offer an explanation. We had to discard trials in which children did not make a prediction (within 4.25 s), which affected 49 trials in total. That is, a response could only be registered if the child made a prediction in the respective time frame. Predictions from children who failed to respond within this time frame were missing. The average number of missed trials per child was 1.33 (Median = 1.00, $SD = 1.91$, range: [0–10]). The likelihood that children missed a trial decreased over the course of the experiment ($b = -0.04$, $p = .036$), and children were more likely to miss incongruent trials compared to congruent trials ($b = 1.49$, $p < .001$). These results suggest that many children missed at least one—mostly incongruent—trial at the beginning of the experiment. However, even if children failed to make a prediction in time, they were presented the correct answer during the results phase, which gave them the opportunity to revise their beliefs. One could expect that children who had difficulties to make a prediction in time would have likely produced an incorrect prediction. This might, in part, explain the comparably high average percentage of correctly solved incongruent trials in the prediction group throughout the learning phase ($M = 0.84$, $SD = 0.15$), despite the fact that children had misconceptions in the pretest. If missed trials were coded as errors, the average percentage of correctly solved incongruent trials drops ($M = 0.79$, $SD = 0.15$).

We also tested whether children in the postdiction condition were able to differentiate between their prior false belief and the actual outcome. We found that the post hoc judgments matched the actual outcome in 66% of cases ($SD = 0.30$), which is similar to the accuracy in the prediction condition. This finding indicates that children in the postdiction condition could differentiate between their prior false belief and the correct outcome. We further tested whether repeating a false belief impaired belief revision in the postdiction condition. We found that the correctness of children's postdictions during the learning phase did not moderate the increase in performance from pretest to posttest ($F(1, 35) = 0.309$, $p = .582$). Hence, belief revision from pretest to posttest was not affected by the correctness of children's postdictions.

Pupillary Data Preparation and Data Analysis

We used a self-developed analysis script to analyze the PDR during the learning phase (for details, see Breitwieser & Brod, 2021). First, we preprocessed the pupillary data by removing blinks and

saccades. Blinks and saccades resulted in missing values, which were interpolated using fitted values. We calculated these fitted values by estimating local regressions using 300 data points on each side of the regressed data point. Second, pupillary data were epoched relative to the onset of the "Results Phase." The average pupil diameter during the final 300 ms of each trial's "Pupil Baseline Phase" served as a baseline for the pupil analyses. This late and relatively short interval was chosen because pupil size is sensitive to changes in higher-level image content (even if luminance is matched, Naber & Nakayama, 2013), and because we wanted to keep this influence to a minimum. It is, thus, plausible that the pupil was adapting during the Pupil Baseline Phase in the postdiction condition, which could explain why the overall time course of the PDR looks different between the prediction and postdiction condition (see Figures 3A and 3B). It is important to note, though, that these overall differences in time course are unlikely to affect expected and unexpected trials differently, which were the main target of our analyses.

As a further sanity check for the pupillary data, we tested whether children's gaze patterns during the pupil baseline were comparable for expected and unexpected trials (see Supporting Information). As heavier spheres were darker than lighter spheres, we tested whether children fixated them more, which could distort the pupillary baseline. We found no evidence that children looked more at the heavier spheres nor that this differed between expected and unexpected trials. Hence, it seems unlikely that the differences in luminance between the spheres caused systematic differences in the pupil baseline between expected and unexpected trials.

To establish a marker of surprise in the pupillary data, we calculated the average change in pupil diameter 250–2,000 ms after the onset of the "Results Phase" relative to each trial's "Pupil Baseline Phase." This analysis window was chosen based on the expected time course of the PDR as observed in previous studies with similar paradigms (see Brod et al., 2018, 2020). In these studies, the PDR peaked between 500 and 1,000 ms after the result presentation and unfolded for about 1,000 ms afterward. We discarded the first 250 ms to account for the typical delay in PDR. Furthermore, to identify potential outliers, we inspected the average PDR in the prediction and postdiction conditions for each trial. We removed trials in which the PDR deviated more than 3 SD from the average pupillary response of a particular person throughout the experiment.

We calculated a linear mixed-effects model to test whether generating predictions elicited a greater PDR for unexpected compared to expected results (i.e., a pupillary surprise response). To capture children's surprise most clearly, trials were defined as unexpected if the child (a) expected that the left sphere would displace more water (i.e., pressed "1") when it was, in fact, the right sphere or when both spheres displaced an equal amount ("3"), (b) expected that the right sphere would displace more water ("5") when it was, in fact, the left sphere or when both spheres displaced an equal amount ("3"), or (c) expected that both spheres would displace an equal amount of water ("3") when it was, in fact, either the right sphere or the left sphere that displaced more water. Hence, correctly answered trials in which the child was uncertain about the correct answer (i.e., pressed "2" or "4") were not classified as unexpected. Likewise, incorrectly answered trials were only coded as unexpected if the child was certain about the answer (i.e., pressed "1" or "5"). To classify the correctness of an answer, we did not differentiate the degree of certainty, meaning that it did not matter whether children were confident in their prediction/postdiction.

To examine the relation between surprise and belief revision, we then examined whether children's PDR to unexpected outcomes predicted subsequent belief revision. For this purpose, we conducted a logistic mixed-effects regression analysis to analyze whether a higher PDR after an incorrectly predicted incongruent trial was associated with an increased likelihood to correctly solve the subsequent incongruent trial. Using a multilevel approach (trials clustered in children), we could account for between-subject differences in average PDR to test the intraindividual process account that a larger PDR to unexpected events (compared to each child's usual PDR) increases the likelihood of belief revision in children. Doing so, we were able to examine the intraindividual process of belief revision in response to conflicting evidence. We conducted this analysis in the prediction group only because children in the postdiction group did not make a prediction beforehand that could then be violated. Accordingly, the postdiction group did not show a PDR to outcomes that were afterward classified as unexpected during the postdiction ($t(59) = -.41$, $p = .683$, see also results section below).

Results

Behavioral Performance

We first tested whether generating predictions facilitated the correction of misconceptions compared to giving post hoc expectancy ratings (i.e., postdictions). We used the percentage of correctly solved incongruent trials as a measure of children's pretest and posttest performance, respectively (see Figure 2). There were no pretest differences between groups ($t(78) = .54$, $p = .592$), and both groups answered most trials incorrectly, indicating the existence of misconceptions (prediction: $M_{Pretest} = 0.33$, $SD_{Pretest} = 0.28$; postdiction: $M_{Pretest} = 0.36$, $SD_{Pretest} = 0.24$). While children in both conditions strongly improved from pre- to posttest ($F(1, 78) = 397.07$, $p < .001$, $d = 4.53$), children in the prediction condition showed higher learning gains than children in the postdiction condition (see Figure 2A, $F(1, 78) = 4.75$, $p = .032$, $d = 0.49$, prediction: $M_{Posttest} = 0.97$, $SD_{Posttest} = 0.10$; $\Delta = .64$; postdiction: $M = 0.87$, $SD = 0.20$; $\Delta = .51$). Concurrent evidence was obtained in the analysis of process data during the learning phase. A logistic mixed-effects regression analysis revealed a performance increase in incongruent trials over time in the prediction condition ($\beta = .81$, $p < .001$, 95% CI [0.55, 1.06]). In contrast, there was no clear time trend in the postdiction condition ($\beta = -.11$, $p = .324$, 95% CI [−0.32, 0.11]). Note, however, that children in the postdiction condition made post hoc judgments of whether they would have predicted the outcome, which cannot be interpreted as a direct performance measure.

Turning to the transfer tasks, in line with our hypothesis, a between-subject t-test revealed a higher percentage of correctly solved trials in the prediction compared to the postdiction condition (see Figure 2B, prediction: $M = 0.86$; $SD = 0.17$; postdiction: $M = 0.74$, $SD = 0.25$; $t(78) = -2.48$, one-tailed $p = .008$, $d = 0.56$). The pattern of results was similar for the receptive transfer task (prediction: $M = 0.89$; $SD = 0.21$; postdiction: $M = 0.72$, $SD = 0.34$; $t(78) = -2.68$, one-tailed $p = .005$, $d = 0.61$) and the productive transfer task (prediction: $M = 0.83$; $SD = 0.22$; postdiction: $M = 0.76$, $SD = 0.21$; $t(78) = -1.47$, one-tailed $p = .072$, $d = 0.33$). Taken together, these results indicate that generating predictions facilitated the correction of misconceptions in children.

Pupillometry

The following set of analyses examined surprise as a potential mechanism by which generating predictions promotes belief revision during the learning phase. In a first step, we tested whether unexpected results that contradicted children's misconception elicited a PDR. In line with our hypothesis, children in the prediction condition (Figure 3A) showed a larger PDR for unexpected than expected results, that is, a pupillary surprise response ($\beta = .11$, one-tailed $p = .005$, 95% CI [0.03, 0.19]). A model with random slopes for expectancy-consistent versus expectancy-violating trials did not provide a better model fit compared to a model with random intercepts only ($\chi^2(2) = 1.16$, $p = .559$). That is, the relation between expectancy and PDR
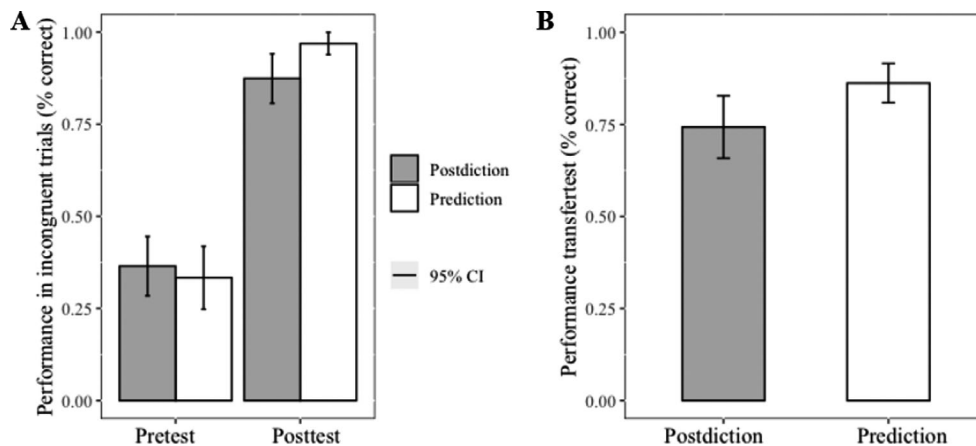

Figure 2. Generating predictions facilitates belief revision. Compared to children in the postdiction condition, children in the prediction condition showed a steeper performance increase from pretest to posttest (A) and a better performance in the transfer test (B). Error bars represent 95% confidence intervals.
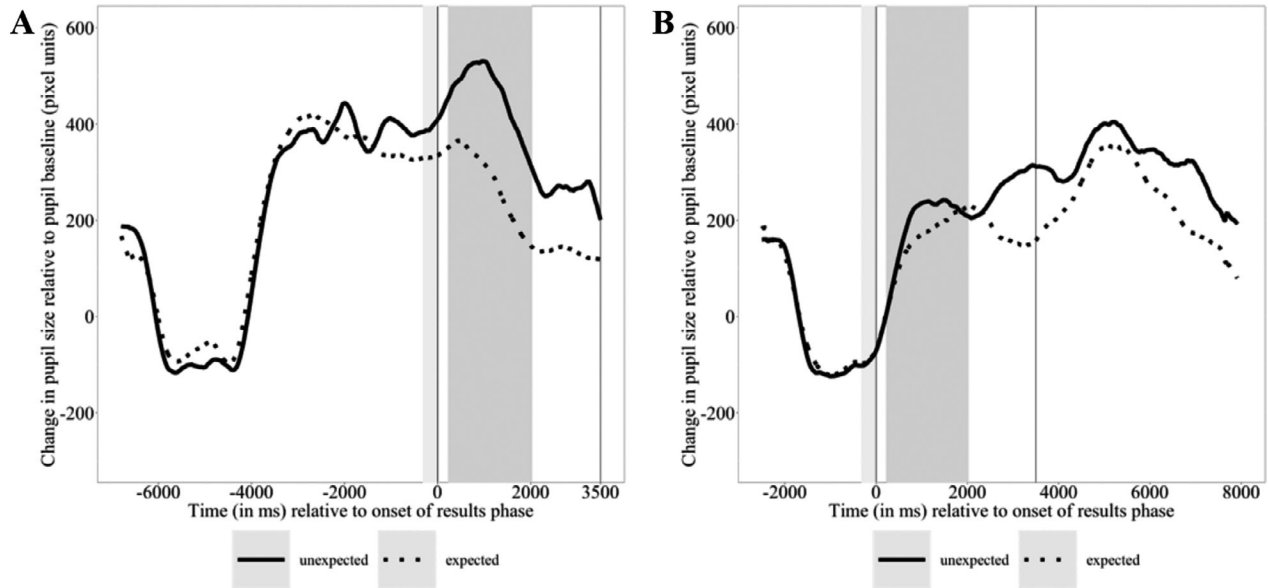
Figure 3. Generating predictions elicits a pupillary surprise response to unexpected results. Children in the prediction condition (A) showed a stronger pupil dilation response (PDR) to seeing expectancy-violating than expectancy-consistent results. In the postdiction condition (B) the PDR for expectancy-violating and expectancy-consistent results did not differ significantly. Change in pupil size relative to the pupil baseline (light gray area) was calculated for the time interval from 250 to 2,000 ms after the onset of the results phase (dark gray area).

was comparable between children. In contrast to the prediction condition, children in the postdiction condition did not show a PDR to unexpected compared to expected results (Figure 3B; $\beta = .02$, $p = .764$, 95% CI [-0.08, 0.11]). Hence, consistent with previous reports (Brod et al., 2018, 2020), unexpected results elicited a PDR only when a prediction was made beforehand. The prediction versus postdiction design can, thus, be seen as an indirect experimental manipulation of surprise intensity.

To examine the link between surprise and belief revision, we then tested whether the PDR in the prediction condition was related to children's performance in the subsequent incongruent trial. That is, we tested whether children's PDR after unexpected, incongruent trials predicted a change toward the scientifically correct model in the next incongruent trial (see Figure 4). In line with our hypothesis, a larger PDR to an unexpected outcome was associated with a higher likelihood of predicting the next incongruent trial correctly ($\beta = .55$, one-tailed $p = .014$, 95% CI [0.06, 1.04]). Put differently, incongruent trials that were predicted correctly were preceded by a larger PDR in the previous incongruent trial than incongruent trials that were predicted incorrectly. A model with random slopes for the PDR after unexpected,

incongruent trials did not provide a better model fit compared to a model with random intercepts only ($\chi^2(2) = 0.002$, $p = .999$). That is, the relation between the PDR after unexpected, incongruent trials and performance in the next incongruent trial was comparable between children. We further explored whether the PDR to expected, incongruent trials was positively related to performance in the next incongruent trial as well. We found a nonsignificant trend in the opposite direction: A larger PDR to expected, incongruent trials were associated with a lower likelihood of answering the next incongruent trial correctly ($\beta = -.35$, $p = .055$, 95% CI [-0.70, 0.01]). Taken together, the pupillometry data suggest that incorrect predictions elicited surprise and that the degree of surprise was positively related to successful belief revision during the learning phase.

Discussion

This study examined whether surprise about conflicting outcomes predicts how much children revise their misconceptions. Children who generated predictions before seeing outcomes that conflicted with their misconception showed a pupillary surprise response. The magnitude of this response
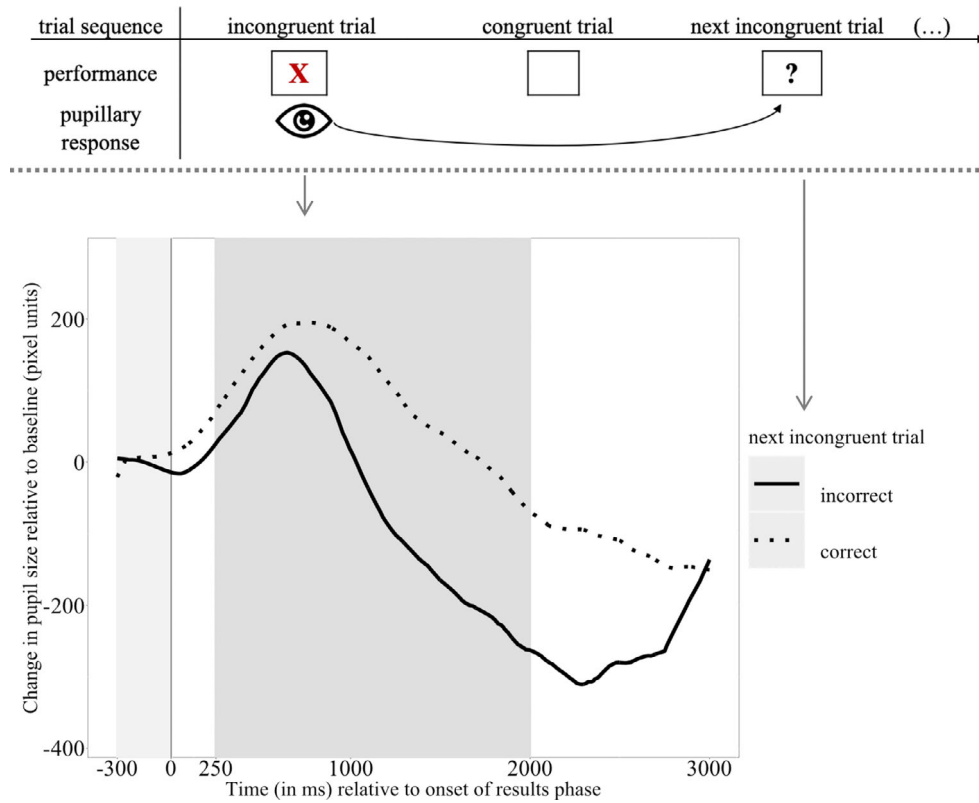
Figure 4. Surprise predicts successful belief revision. Incongruent trials that were predicted correctly were preceded by a larger pupil dilation in the previous incongruent trial (dotted line) compared to incongruent trials that were predicted incorrectly (solid line). Change in pupil size relative to the pupil baseline (light gray area) was again calculated for the time interval from 250 to 2,000 ms after the onset of the results phase (dark gray area).

positively predicted subsequent belief revision. In contrast, children who stated their expectation *after* seeing the conflicting outcome did not show a pupillary surprise response and exhibited less belief revision. In conclusion, this study indicates that surprise plays an important role in predicting how much children revise their misconceptions.

Results suggest that incorrect predictions induce surprise. Surprise signals a cognitive conflict between the initially held belief and the scientifically correct concept. Physiologically, conflicting information that violates prior expectations evokes a PDR (Kloosterman et al., 2015; Preuschoff et al., 2011). The PDR is age-invariant (e.g., Krüger et al., 2020; Zhang et al., 2018) and known to be a good proxy for enhanced physiological arousal that is induced by increased activity of the autonomic nervous system (e.g., Bradley, Miccoli, Escrig, & Lang, 2008). Recent theories provide a physiologically plausible mechanism of how the pupillary surprise response could facilitate belief revision. Changes in pupil size are affected by the release of norepinephrine in the brainstem's locus coeruleus (Joshi,

Li, Kalwani, & Gold, 2016). Norepinephrine release has been suggested as a global model failure signal that leads to the interruption of ongoing processes, and to promote attention and memory for goal-relevant information (Clewett, Huang, Velasco, Lee, & Mather, 2018). It is also thought to initiate updating in a Bayesian fashion (Dayan & Yu, 2006; Yu & Dayan, 2005). Taken together, although clearly speculative for a high-level belief revision task such as ours, these theories suggest an intricate neurobiological mechanism for the link between the pupillary surprise response and subsequent belief revision observed in our study.

On a cognitive level, surprise signals a discrepancy between the learners' expectations and new information (Reisenzein et al., 2019). This discrepancy causes an interruption in ongoing cognitive processes and draws learners' attention to the conflicting information, which benefits memory for it (see Fazio & Marsh, 2009; Stahl & Feigenson, 2017). Surprise has further been suggested to function as a metacognitive cue to engage in deep processing of the conflicting information and to search for

alternative explanations (Munnich & Ranney, 2018). Active evaluation of alternative explanations has been shown to support students' science learning (Lombardi, Bailey, Bickel, & Burrell, 2018). In sum, surprise might lead learners to elaborate on conflicting information, which paves the way for belief revision.

Of note, simply presenting conflicting information did not suffice to induce surprise. The pupillary surprise response only occurred if a prediction was made before the correct result was revealed (i.e., in the prediction condition). The postdiction condition was designed in a way that children did not have enough time to generate and commit to a prediction before seeing the correct result. Children in the postdiction condition did not show a pupillary surprise response, which replicated prior findings (Brod et al., 2018, 2020). This experimental design, thus, yielded an indirect manipulation of surprise intensity. Children in the postdiction condition (i.e., low surprise condition) showed smaller learning gains from pretest to posttest and displayed more misconceptions in the transfer test compared to children in the prediction condition (i.e., high surprise condition). In sum, results suggest that generating predictions enabled children to be surprised about conflicting information, which benefitted their belief revision.

Our findings also bear on related areas of research on correcting misconceptions, such as refutation texts. In refutation texts, the misconception is typically stated before the scientifically accepted view is presented. It has been argued that the co-activation of the misconception and the scientifically accepted view makes the conflict more salient, which benefits its resolution (for an overview see Broughton, Sinatra, & Reynolds, 2010). For instance, eighth-graders who read a refutation text were more likely to revise the misconception than those who read an explanatory text that only stated the scientifically accepted view without mentioning the misconception (van Loon, Dunlosky, van Gog, van Merriënboer, & de Bruin, 2015). This finding suggests that activating the prior theory is necessary for the beneficial effect to occur. Building on the findings of this study, it seems worthwhile to test whether a reason for this effect is the greater surprise response in the refutation text condition.

The current findings give further support for the increasingly recognized importance of epistemic emotions in belief revision (Sinatra et al., 2014; Vogl et al., 2020). In line with recent proposals (Inzlicht, Bartholow, & Hirsh, 2015), our findings suggest a link between emotions and conflict resolution. Emotions make a cognitive conflict more salient, which might prevent learners from ignoring conflicting information. Following conflict detection, learners may engage in effortful processes to resolve the conflict. Our findings, thus, complement studies showing that children change their beliefs when confronted with strongly conflicting information (e.g., Kimura & Gopnik, 2019) by indicating that belief revision is not a "cold" process. In summary, our results provide evidence for the claim that belief revision constitutes both a cognitive and an emotional process.

Limitations and Future Directions

There are a number of limitations of this study, which provide avenues for future research. First, children acquired the scientific concept rather fast. Children had to learn one simple rule: volume alone determines water displacement. We intentionally chose this simple task because we wanted to tackle a simple, yet common, misconception. Only then could we expect to observe surprise-related belief revision during such a short amount of time. While children in the prediction (/high surprise) condition demonstrated faster belief revision than children in the postdiction (/low surprise) condition, children in the postdiction condition showed substantial belief revision as well. This pattern indicates that while surprise is clearly not the only factor driving belief revision, it does serve an accelerating role. Future studies should examine whether this pattern holds in more complex belief revision tasks where the correct answer depends on more than one rule (e.g., buoyancy). We assume that the proposed link from surprise to belief revision would also hold for more complex tasks that include more than one plausible alternative hypothesis. However, in such a more complex scenario, it has to be ensured that the correct alternative is deducible for the learner. Furthermore, the task instructions have to be very clear. For instance, as buoyancy and water displacement are related concepts, some children in our study may have initially confused the two concepts. Future studies that use more complex tasks should further test whether greater surprise goes along with more exploratory behavior by the learner.

The second limitation refers to the measurement and conceptualization of surprise. In this study, we assessed the physiological component of surprise, that is, a brief increase in physiological arousal as indexed by a succinct PDR (Kloosterman et al., 2015; Preuschoff et al., 2011). We observed that

higher surprise (as assessed via the PDR) predicted the likelihood of belief revision later on. Since it seems virtually impossible to manipulate surprise directly, we indirectly manipulated surprise intensity by letting children generate predictions (vs. making post hoc judgments). Therefore, we cannot completely rule out that a third variable caused by generating predictions could have led to both higher surprise and a higher likelihood of belief revision. It is further questionable whether a brief increase in arousal following a violation of expectation is sufficient to infer that children also *felt surprised* (for a recent summary of the debate on the state of surprise as an emotion, see Munnich, Foster, & Keane, 2019; Reisenzein et al., 2019). Future research could include self-reports of surprise intensity and correlate those with pupil dilation data.

Third, while generating prediction was more effective in promoting the revision of misconceptions than generating postdictions, the durability of the observed effects is debatable. According to the coexistence claim, misconception never fully disappear (Shtulman & Valcarcel, 2012). That is, misconceptions continue to interfere with the correct scientific concept, but this interference may be reduced with repeated practice. Hence, it seems likely that, without repeated practice, the beneficial effects obtained in this study would attenuate over time. To test this hypothesis, future studies should conduct follow-up tests after a substantial delay, test whether findings hold for live demonstration (ideally in classrooms), and include far transfer tasks to demonstrate true and lasting belief revision.

Finally, we cannot rule out that repeating false beliefs in the postdiction condition impaired belief revision. Children in the postdiction condition were asked to state retrospectively which outcome they would have predicted. Inevitably, children therefore sometimes repeated their false belief. We chose this experimental manipulation to avoid confounding effects that could arise if only the prediction group had to make an active choice that requires semantic elaboration. To test whether repeating a false belief impaired belief revision, we tested whether the correctness of children's postdictions during the learning phase moderated the increase in performance from pretest to posttest, which was not the case. To examine whether repeating a false belief impairs belief revision, future studies could include another condition in which children are not asked to make a post hoc judgment after each trial. A previous study with children that included such an additional baseline condition found that children's memory for unexpected outcomes was significantly better in the postdiction condition than in the baseline condition (Brod et al., 2020). This result suggests that the postdiction condition does not generally impair learning of unexpected outcomes that refute prior beliefs.

Conclusion

Picture a typical science class in school: A teacher asks students to first *predict* the outcome of an experiment, then *observe* the outcome, and finally *explain* the outcome. "Predict—Observe—Explain" (POE; White & Gunstone, 1992) constitutes a popular instructional practice to foster believe revision in learners. This study offered an explanation why POE could be effective. Letting students generate predictions constitutes a simple, instructional tool to elicit surprise and grasp students' attention. Students become cognitively and emotionally engaged, which builds an ideal starting point to encourage deeper scientific reasoning. Teaching science entails discussing conflicting outcomes and explanations; surprise could help students to leverage these conflicts for learning.

References

Babai, R., & Amsterdamer, A. (2008). The persistence of solid and liquid naive conceptions: A reaction time study. *Journal of Science Education and Technology*, *17*, 553–559. https://doi.org/10.1007/s10956-008-9122-6

Bonawitz, E., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*, 215–234. https://doi.org/10.1016/j.cogpsych.2011.12.002

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*, 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Breitwieser, J., & Brod, G. (2021). Cognitive prerequisites for generative learning: Why some learning strategies are more effective than others. *Child Development*, *92*, 258–272. https://doi.org/10.1111/cdev.13393

Brod, G., Breitwieser, J., Hasselhorn, M., & Bunge, S. A. (2020). Being proven wrong elicits learning in children —But only in those with higher executive function skills. *Developmental Science*, *23*, 1–14. https://doi.org/10.1111/desc.12916

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction*, *55*, 22–31. https://doi.org/10.1016/j.learninstruc.2018.01.013

Broughton, S. H., Sinatra, G. M., & Reynolds, R. E. (2010). The nature of the refutation text effect: An

investigation of attention allocation. *Journal of Educational Research*, *103*, 407–423. https://doi.org/10.1080/00220670903383101

Burbules, N. C., & Linn, M. C. (1988). Response to contradiction: Scientific reasoning during adolescence. *Journal of Educational Psychology*, *80*, 67–75. https://doi.org/10.1037/0022-0663.80.1.67

Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 61–82). New York, NY: Routledge.

Chi, M. T. H. (2013). Two kinds and four sub-types of misconcieved knowledge, ways to change it, and the learning outcomes. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 49–70). https://doi.org/10.4324/9780203154472

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*, 1. https://doi.org/10.2307/1170558

Clewett, D. V., Huang, R., Velasco, R., Lee, T.-H., & Mather, M. (2018). Locus coeruleus activity strengthens prioritized memories under arousal. *The Journal of Neuroscience*, *38*, 1558–1574. https://doi.org/10.1523/JNEUROSCI.2097-17.2017

Dawson, C., & Rowell, J. (1984). Displacement of water: Weight or volume? An examination of two conflict based teaching strategies. *Research in Science Education*, *14*, 69–77. https://doi.org/10.1007/BF02356792

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, *17*, 335–350. https://doi.org/10.1080/09548980601004024

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12–30. https://doi.org/10.1016/j.cogdev.2018.01.001

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

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

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*, 307–326. https://doi.org/10.1037/0022-0663.98.2.307

Inzlicht, M., Bartholow, B. D., & Hirsh, J. B. (2015). Emotional foundations of cognitive control. *Trends in Cognitive Sciences*, *19*, 126–132. https://doi.org/10.1016/j.tics.2015.01.004

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, *89*, 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kimura, K., & Gopnik, A. (2019). Rational higher-order belief revision in young children. *Child Development*, *90*, 91–97. https://doi.org/10.1111/cdev.13143

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*, 1068–1078. https://doi.org/10.1111/ejn.12859

Krüger, M., Bartels, W., & Krist, H. (2020). Illuminating the dark ages: Pupil dilation as a measure of expectancy violation across the life span. *Child Development*, *91*, 2221–2236. https://doi.org/10.1111/cdev.13354

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

Linn, M. C., & Eylon, B. S. (2000). Knowledge integration and displaced volume. *Journal of Science Education and Technology*, *9*, 287–310. https://doi.org/10.1023/A:1009451808539

Lombardi, D., Bailey, J. M., Bickel, E. S., & Burrell, S. (2018). Scaffolding scientific thinking: Students' evaluations and judgements during Earth science knowledge construction. *Contemporary Educational Psychology*, *54*, 184–198. https://doi.org/10.1016/j.cedpsych.2018.06.008

Munnich, E. L., Foster, M. I., & Keane, M. T. (2019). Editors' introduction and review: An appraisal of surprise: Tracing the threads that stitch it together. *Topics in Cognitive Science*, *11*, 37–49. https://doi.org/10.1111/tops.12402

Munnich, E. L., & Ranney, M. A. (2018). Learning from surprise: Harnessing a metacognitive surprise signal to build and adapt belief networks. *Topics in Cognitive Science*, *11*. https://doi.org/10.1111/tops.12397

Naber, M., & Nakayama, K. (2013). Pupil responses to high-level image content. *Journal of Vision*, *13*, 7. https://doi.org/10.1167/13.6.7

Noordewier, M. K., Topolinski, S., & Van Dijk, E. (2016). The temporal dynamics of surprise. *Social and Personality Psychology Compass*, *10*, 136–149. https://doi.org/10.1111/spc3.12242

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Piaget, J., & Inhelder, B. (1941). *Le développement des quantités chez l'enfant: Conservation et atomisme. [The development of the idea of quantity in the child]*. Neuchâtel, Paris: Delachaux et Niestlé.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception:

Toward a theory of conceptual change. *Science Education*, *66*, 211–227. https://doi.org/10.1002/sce.3730660207

Potvin, P., Masson, S., Lafortune, S., & Cyr, G. (2015). Persistence of the intuitive conception that heavier objects sink more: A reaction time study with different levels of interference. *International Journal of Science and Mathematics Education*, *13*, 21–43. https://doi.org/10.1007/s10763-014-9520-6

Potvin, P., Sauriol, É., & Riopel, M. (2015). Experimental evidence of the superiority of the prevalence model of conceptual change over the classical models and repetition. *Journal of Research in Science Teaching*, *52*, 1082–1108. https://doi.org/10.1002/tea.21235

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*, 1–12. https://doi.org/10.3389/fnins.2011.00115

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reisenzein, R., Horstmann, G., & Schützwohl, A. (2019). The cognitive-evolutionary model of surprise: A review of the evidence. *Topics in Cognitive Science*, *11*, 50–74. https://doi.org/10.1111/tops.12292

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209–215. https://doi.org/10.1016/j.cognition.2012.04.005

Sinatra, G. M., Broughton, S. H., & Lombardi, D. (2014). Emotions in science education. In R. Pekrun & L. Linnenbrink-Garcie (Eds.), *International handbook of emotions in education* (pp. 415–436). https://doi.org/10.4324/9780203148211

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

van Loon, M. H., Dunlosky, J., van Gog, T., van Merriënboer, J. J. G., & de Bruin, A. B. H. (2015). Refutations in science texts lead to hypercorrection of misconceptions held with high confidence. *Contemporary Educational Psychology*, *42*, 39–48. https://doi.org/10.1016/j.cedpsych.2015.04.003

Vogl, E., Pekrun, R., Murayama, K., & Loderer, K. (2020). Surprised–curious–confused: Epistemic emotions and knowledge exploration. *Emotion*, *20*, 625–641. https://doi.org/10.1037/emo0000578

Vosniadou, S. (2019). The development of students' understanding of science. *Frontiers in Education*, *4*, 1–6. https://doi.org/10.3389/feduc.2019.00032

Vosniadou, S., & Ioannides, C. (1998). From conceptual development to science education: A psychological point of view. *International Journal of Science Education*, *20*, 1213–1230. https://doi.org/10.1080/0950069980201004

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*, 229–246. https://doi.org/10.1111/cdev.12590

White, R. T., & Gunstone, R. F. (1992). *Probing understanding*. London, UK: Falmer Press.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

Zhang, F., Jaffe-Dax, S., Wilson, R. C., & Emberson, L. L. (2018). Prediction in infants and adults: A pupillometry study. *Developmental Science*, *22*, e12780. https://doi.org/10.1111/desc.12780

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. Supplementary materials.