CHILD DEVELOPMENT

Child Development, xxxx 2020, Volume 00, Number 0, Pages 1–15

Cognitive Prerequisites for Generative Learning: Why Some Learning Strategies Are More Effective Than Others

Jasmin Breitwieser
DIPF Leibniz Institute for Research and Information in Education

Garvin Brod
DIPF Leibniz Institute for Research and Information in Education and Goethe University Frankfurt

This study examined age-related differences in the effectiveness of two generative learning strategies (GLSs). Twenty-five children aged 9–11 and 25 university students aged 17–29 performed a facts learning task in which they had to generate either a prediction or an example before seeing the correct result. We found a significant Age × Learning Strategy interaction, with children remembering more facts after generating predictions rather than examples, whereas both strategies were similarly effective in adults. Pupillary data indicated that predictions stimulated surprise, whereas the effectiveness of example-based learning correlated with children's analogical reasoning abilities. These findings suggest that there are different cognitive prerequisites for different GLSs, which results in varying degrees of strategy effectiveness by age.

Constructivism conceives learning as an active process whereby learners themselves construct relations between their existing knowledge and new experiences (Piaget, 1926; Vygotsky, 1978; Wittrock, 2010). An increasingly prominent example of a group of learning strategies inspired by constructivist theories of learning are called generative learning strategies (GLSs; Fiorella & Mayer, 2016). GLSs prompt learners to actively make sense of the to-be-learned information and to integrate it with their prior knowledge (Fiorella & Mayer, 2016; Wittrock, 2010). GLSs can be expected to improve memory performance because they introduce a desirable difficulty (Bjork, 1994; Bjork & Bjork, 2011); that is, learners are asked to retrieve prior knowledge instead of simply re-reading the information. The retrieval process itself presents a more powerful learning opportunity than re-studying of the material for long-term retention (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006). In line with these considerations, GLSs have been found to foster various learning outcomes compared to more passive learning activities (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Fiorella & Mayer, 2016; Lee, Lim, & Grabowski, 2008).

Although constructivist theories of learning were initially put forward with a view to explain developmental differences in learning (Piaget, 1926; Vygotsky, 1978), research comparing the effectiveness of GLSs between different age groups has been surprisingly sparse. Studies that did include a wider age range typically focused only on one strategy and either did not compare age groups directly or, if they did, treated age as a surrogate for varying levels of prior knowledge (e.g., Chularut & DeBacker, 2004; Gurlitt & Renkl, 2008). A key finding from these studies has been that GLSs should be adapted to provide greater support for (mostly younger) learners with lower prior knowledge (Gurlitt & Renkl, 2008). However, in addition to age-related differences in prior knowledge, the effectiveness of GLSs is likely also affected by distinct cognitive mechanisms underlying different GLSs. Since different GLSs achieve active integration of new information in different ways (e.g., by asking learners to generate explanations or to

Garvin Brod was supported by a Jacobs Foundation Early Career Research Fellowship. This study was preregistered on the Open Science Framework. All data and materials, along with analysis scripts, have been made publicly available and can be accessed at https://osf.io/e4h9n/?view_only=37f9109f122b42faaec678796b225037. We thank Freya Moßig and Manuel Lieb for their help in collecting the data and Silvia Bunge for valuable discussions.

Open access funding enabled and organized by Projekt DEAL.

[Corrections added on July 31, 2020, after first online publication. the open access funding statement has been added.]

Correspondence concerning this article should be addressed to Jasmin Breitwieser or Garvin Brod, DIPF|Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt am Main, Germany. Electronic mail may be sent to breitwieser@dipf.de or garvin.brod@dipf.de.

0009-3920/2020/xxxx-xxxx
DOI: 10.1111/cdev.13393

generate examples), they rely on different cognitive abilities (e.g., verbal abilities, reasoning abilities), which follow different developmental trajectories (see Li et al., 2004). It is therefore plausible to assume that the relative effectiveness of different GLSs varies depending on the age of the learner. Knowledge of these differences would be crucial for selecting optimal GLSs for learners of all ages.

The present study tested these notions by comparing two commonly used GLSs that could be expected to differ in their underlying cognitive mechanisms: generating predictions and generating examples.

Asking learners to generate a prediction (also known as generating hypotheses) before telling them the correct solution requires learners to engage in effortful retrieval of relevant prior knowledge, and thus qualifies as a comparably simple GLS. Even when participants do not have extensive prior knowledge to build their prediction on and can, thus, only guess, this has been shown to improve their learning (Huelser & Metcalfe, 2012; Metcalfe & Kornell, 2007; Potts & Shanks, 2014; Richland, Kornell, & Kao, 2009). However, retrieval failures that are committed with high confidence have been shown to lead to greater learning than those committed with low confidence (Butterfield & Metcalfe, 2001, 2006). A possible mechanism underlying this hypercorrection effect is increased attention to surprising outcomes (Butterfield & Metcalfe, 2006), which aligns well with research on generating predictions. A recent study (Brod, Hasselhorn, & Bunge, 2018) found that generating a prediction enabled learners to experience surprise about the correct solution. Surprise was linked to enhanced learning, presumably by increased attention to task-relevant information (see Fazio & Marsh, 2009; Stahl & Feigenson, 2019). In line with reconsolidation theory (Alberini & LeDoux, 2018) and predictive coding (Friston, Thornton, & Clark, 2012), surprise may drive the destabilization and updating of stored memories in light of new information (Sinclair & Barense, 2018). Furthermore, generating a prediction might also affect metamemory as a prediction followed by feedback forces learners to perform a memory search and gives them clues about the accuracy of this search (Koriat, 1993). Whether the surprise induced by violated predictions also facilitates learning in children has not yet been directly investigated. However, related research suggests that the surprise response to expectancy-violating events is age-invariant (Schützwohl & Reisenzein, 1999), that pupillary reactions reflect surprise already in infants (Jackson & Sirois, 2009), and that expectancy-violations promote declarative learning in children as young as 3 years old (Stahl & Feigenson, 2017).

Asking learners to generate examples is a technique often used in the context of concept learning (e.g., Gorrell, Tricou, & Graham, 1991; Rawson & Dunlosky, 2016). Tying an abstract concept to a concrete example and, thus, encoding it as part of an interrelated semantic network is meant to make the concept more meaningful, thus fostering its understanding (Bjork, 1994). Even though example quality is strongly correlated with successful concept learning, learning success rather depends on the processes involved in example generation than on the examples themselves (Rawson & Dunlosky, 2016). To successfully generate examples, learners need not only to activate relevant prior knowledge but also to compare potential examples with the abstract concept as well as among each other to determine their appropriateness. Example generation therefore requires analogical reasoning (Duit, 1991), which undergoes large developmental changes during childhood and adolescence (Gentner & Toupin, 1986; Richland, Morrison, & Holyoak, 2006). It is, thus, conceivable that the processes involved in successful example generation differ by age, which could lead to age-related differences in the effectiveness of this strategy.

We compared generating predictions and generating examples between 9- and 11-year-olds and university students. Participants had to learn numerical trivia facts for an immediate recall test under each GLS condition. We chose children in this age range because late childhood marks a period where basic reasoning abilities are in place but not yet fully mature, with large individual differences among children (see Richland et al., 2006). The key hypothesis guiding this study was that the learning benefits of prediction over example generation would be greater in children than in adults (i.e., Age × Learning Strategy interaction). Based on findings in adults suggesting that generating predictions has the specific benefit of boosting surprise (Brod et al., 2018), we assessed changes in pupil diameter to investigate whether surprise contributed to the assumed beneficial effects of generating predictions in children. In addition, we explored whether, among the children, reasoning abilities and executive functions (EFs) differentially affected the effectiveness of the two learning strategies.

Method

This study (including hypotheses, sampling plan, design, and analysis plan) was preregistered on the

Open Science Framework (https://osf.io/e4h9n/?view_only=37f9109f122b42faaec678796b225037).

Participants

We tested 26 university students of Goethe University Frankfurt and 26 children between November 2017 and June 2018. One university student did not meet the predefined inclusion criteria of speaking German at native-speaker level, and one child did not attend Grade 4 or 5; we therefore discarded their data. The final sample consisted of 25 young adults (17 female; $M_{\text{age}} = 21.24$, range = 17–29) and 25 children of Grades 4 and 5 (10 female; 19 in Grade 4; $M_{\text{age}} = 9.84$, range = 9–11). We sought to keep age variability among children low to ensure comparatively similar cognitive development. We therefore tested primarily 10-year-olds, but, for feasibility reasons, ended up also testing five 9-year-olds and one 11-year-old. All participants came from middle-class households. As stated in the preregistration, sample size was determined a priori using GPower 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) with the following settings: analysis of variance (ANOVA) for repeated measures, within-between interaction, .05 alpha error, .95 power to detect an effect size of $f(V) = 0.52$ (as found in pilot studies). All participants as well as children's parents gave written informed consent prior to testing. Sessions lasted about 60 min for young adults and about 75 min for children. Young adults received 10 Euro or course credit for their participation. Children were given a toy worth 5 Euro and their parents received 5 Euro to cover their travel expenses. We recruited the university students through bulletins at Goethe University Frankfurt as well as announcements in student groups on social media. We recruited the children via a database of children who participated in previous studies of our research group as well as an email distributor that disseminates information to parents of children in Grade 4 in the area of Frankfurt/Germany. Ethics approval was obtained from the ethics committee of DIPF|Leibniz Institute for Research and Information in Education.

Measures and Procedure

Overview

After written informed consent was obtained, the test session started with the tasks measuring EFs. This was followed by the numerical facts learning task, which consisted of two study–test cycles (one per condition). We performed eye tracking during both study phases. After finishing the numerical facts learning task, the group of children additionally completed the analogical reasoning task.

Numerical Facts Learning Task

We examined the effect of GLS condition (generating predictions vs. generating examples) on learning performance using a computerized experimental task, which consisted of a study phase and a test phase for each condition, respectively. Participants performed the GLS conditions successively (i.e., study phase 1—test phase 1—study phase 2—test phase 2), the order being counter-balanced across participants.

Each study phase (see Figure 1) started with three practice trials, followed by 30 numerical facts in the format "X out of 10" (e.g., "X out of 10 animal species are insects"). The format remained the same for all trials, thus not requiring varying algebraic operations from children. Experimenters made sure that all children understood the concept of "X" standing for a number. In the prediction condition and in the memory test, a visual analogue scale (VAS) was used (see below) to further aid intelligibility of the task for participants with low numeracy skills (e.g., Galesic, Garcia-Retamero, & Gigerenzer, 2009). Stimuli were presented using PsychoPy v1.8 (Peirce, 2007). Participants first performed the generative task and, after a brief delay, saw the correct number for the placeholder "X." Participants were instructed to remember the correct results for the subsequent memory test. The correct numbers ranged from "1" to "9".

The two GLS conditions only differed in the generative activity that took place at the beginning of each trial. In the prediction condition, participants indicated their expectation for the correct number on a 10-point VAS (portraying ten manikins to make the task more intuitive for children). In accordance with the definition of GLSs (Fiorella & Mayer, 2016; Wittrock, 2010), doing so required learners to activate relevant prior knowledge to come up with an informed guess. Their response was then highlighted for 1 s. In the example condition, participants had to generate an example relevant to the fact. They were instructed to click on the smiley as soon as they had found an example, or to click on the red button if they could not find any. The experimenter stressed that participants should only click on the smiley if they had actually found an example and, to ensure task compliance, mentioned that participants would be asked to

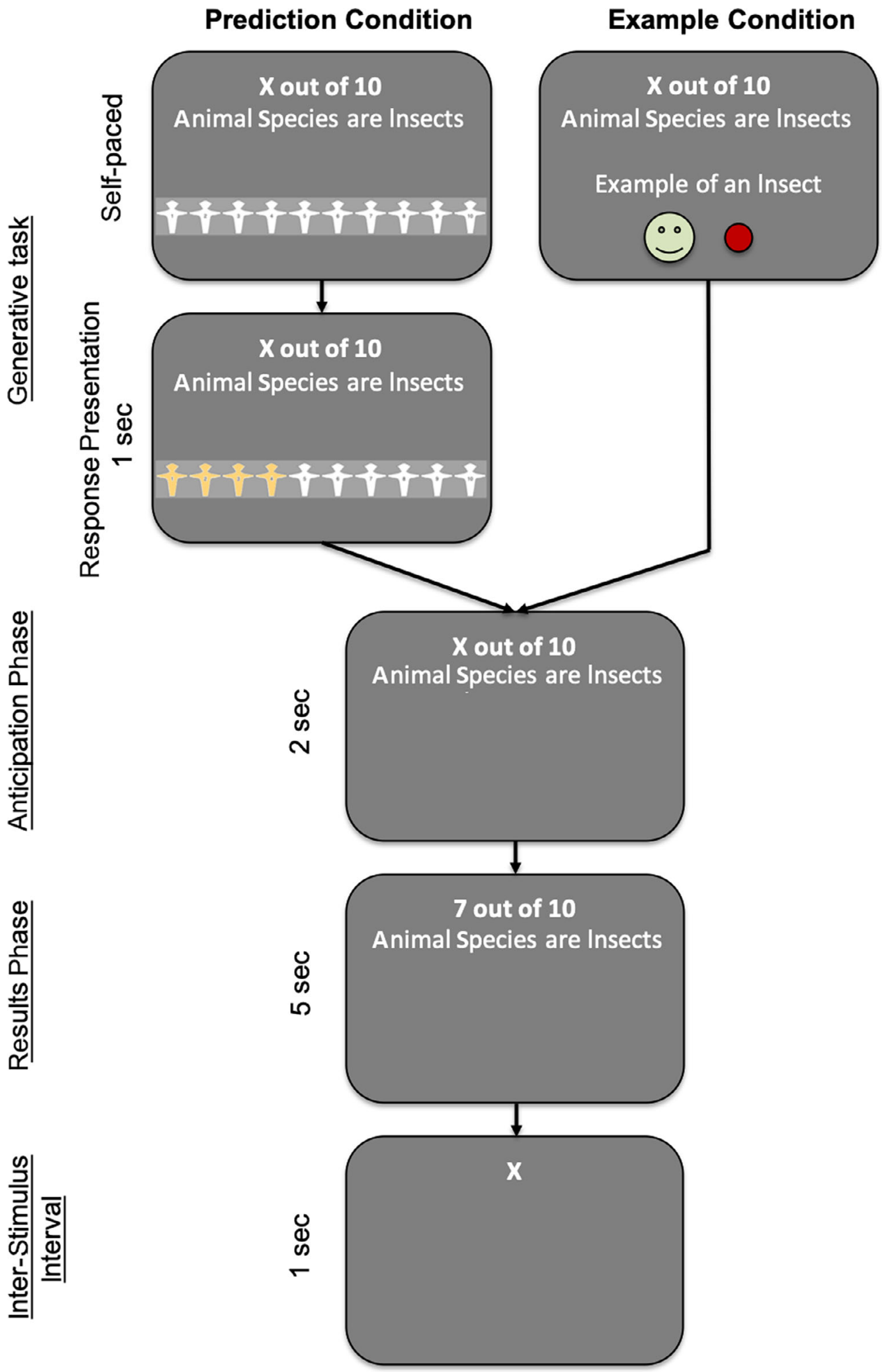

Figure 1. Schematic overview of the study phase of the numerical facts learning task. At the beginning of each trial, participants had to either generate a prediction for the correct value of "X" (prediction condition) or an example that corresponded to the fact (example condition). After a brief delay, the correct result was presented.

provide some of their examples later on. A sample of eight examples was collected after completion of the task (see Table S1).

Participants performed an unrelated digit span backwards task (ca. 1 min) upon completion of the study phase to ensure that the facts were cleared from short-term memory. During the test phases, participants saw the 30 facts again and had to indicate the correct results on the VAS. Responses were highlighted for 1 s before the start of the next trial.

Upon completion of the numerical facts learning task, young adults filled out a brief questionnaire in which they indicated (a) which facts they had known prior to the experiment, (b) which condition was more enjoyable (on a scale from 1, *clearly prediction*, to 6, *clearly example*), and (c) whether they thought they had learned more in the prediction condition or in the example condition (using the same 6-point scale), and why. Children's answers to these questions were recorded by the experimenter.

Moderators

Reasoning abilities. We measured children's reasoning abilities with the analogies subtest of the SON-R 6-40 (Tellegen, Laros, & Petermann, 2012). The test was chosen because it is highly reliable in the age range of our study (0.86–0.88 for children age 9–11). Each item of the paper-and-pencil test showed two figures next to each other, with the right one being obtained from the left one by changing one or more features. Children had to deduct the conversion rule and apply it to a new figure. The test involved an adaptive testing procedure with a maximum of 36 items.

Executive functions. We used a slightly modified version of the Hearts and Flowers Task (HFT; Wright & Diamond, 2014) to measure executive functioning. The task parameters and the testing procedure were identical to those reported by Brod, Bunge, and Shing (2017). (The task, including the exact stimuli and stimulus lists, can be found at https://osf.io/c8gbj/.) In short, the task consisted of three blocks with 20 trials each, each block requiring increasing levels of executive functioning. In the first block (congruent condition), a heart appeared on the left or right side of the screen and participants had to press a button on the same side. In the second block (incongruent condition), a flower was presented and participants had to press the button on the opposite side. In the third block (mixed condition), heart and flower trials were intermixed, and participants had to switch between the previously learned rules.

In addition to the HFT, participants performed an open-source version of the Wisconsin Card Sorting Test, called the Berg Card Sorting Test (BCST; Fox, Mueller, Gray, Raber, & Piper, 2013). We decided to focus on the HFT as our measure of EFs because the HFT is specifically tailored to the assessment of EFs, whereas the BCST is a broader measure of prefrontal cortex function (Diamond, 2013) and might, thus, potentially overlap with reasoning abilities. Scores from the HFT and BCST correlated moderately (r = .37, p = .035, one-tailed) and results were highly similar (see Supporting Information).

Eye Tracking Data Acquisition

We performed eye tracking throughout both study phases of the numerical facts learning task at a frequency of 500 Hz. The eye tracking apparatus (EyeLink 1000; SR Research, Osgoode, Ontario, Canada) was located below the computer screen. Subjects were seated about 68 cm from the screen in a dimly lit room.

Our interest lay in pupil size changes in response to the presentation of the correct number (i.e., during the results phase, see Figure 1). The anticipation phase served as a "pupil baseline" and was therefore closely matched to the luminance of the results phase. We further sought to minimize eye movements by instructing participants to pay close attention to the target position "X" when awaiting the presentation of the correct result.

Data Analyses

We analyzed data using R (R Core Team, 2014) and applied an α level of .05 throughout the analyses. We performed logistic mixed-effects regression models for the analysis of effects on memory performance and linear mixed-effects models for all other analyses that included multiple measurements per condition. We report odds ratio (OR) as an effect size measure for the logistic models and adjusted pseudo R^2 for the linear models using the R package MuMIn (Bartoń, 2018). Deviating from our preregistered analysis plan, we included not only a random intercept but the full random effects structure for participants and a random intercept for items to avoid type I errors due to an underestimation of standard errors (Brauer & Curtin, 2018). We did not include a random slope for items because inclusion would have led to nonconvergence for some models and its correlation with the random intercept was estimated at 1.00. Compared to our preregistered analysis plan, these deviations did not lead to any differences in the pattern of results or in the interpretation, however. Significance was determined by conducting likelihood-ratio tests between the full model and a reduced model without the fixed effect in question while keeping the random effects structure the same; in the presence of an interaction term, we tested main effects by converting factors to sum-coded numeric representations (Levy, 2014).

We analyzed the combined effects of GLS condition and age on performance in the numerical facts learning task as specified in our preregistered analysis plan (https://osf.io/e4h9n/?view_only=37f9109f122b42faaec678796b225037). We considered all other analyses exploratory. Specifically, we explored (a) whether the assumed performance difference between generating predictions and generating examples was linked to children's reasoning abilities or EFs, (b) whether generating predictions had the specific effect of boosting surprise, (c) whether the pupillary response was differentially predictive of learning in the two age groups and conditions, and (d) whether the strength of this link predicted the performance difference.

Analysis of the Pupillary Data

The main goal of the pupillometry analyses was to test the hypothesis that generating predictions, unlike generating examples, elicits a surprise response during result presentation, and that this surprise response mediates the benefit of predictions on learning. Pupillary reactions to expectancy-violating events are an objective measure of surprise (e.g., Kloosterman et al., 2015; Preuschoff, 't Hart, & Einhauser, 2011) and can be observed already in infants (Jackson & Sirois, 2009). Changes in pupil diameter reflect the release of the neurotransmitter noradrenaline in the locus coeruleus (LC). LC activity regulates task-related arousal and is associated with the optimization of task performance (for an overview, see Aston-Jones & Cohen, 2005). Baseline pupil diameters change with age (Eckstein, Guerra-Carrillo, Miller Singley, & Bunge, 2017) and vary from person-to-person and trial-to-trial. Therefore, pupillary reactions to task stimuli are not measured as absolute sizes but as relative changes from baseline at the start of each trial to make them comparable between persons and conditions (cf. Sirois & Brisson, 2014).

The following preprocessing steps were performed in R to prepare the pupil data for further analyses: First, we fitted local regressions (loess) to the data, using 300 data points on each side of the regressed point to calculate the fitted value. We filtered data points more than five standard errors smaller or bigger than the fitted value. Second, we used the loess values to interpolate missing values in the time series unless the gap was longer than 100 data points (i.e., 200 ms). Third, we smoothed time series data with a moving average, using 25 samples on each side of the smoothed data point. Finally, we aligned pupil data relative to the onset of the results phase, and normalized it by subtracting the diameter at each time point from the average diameter in the interval 300 ms before the onset until 100 ms after the onset. The resulting signal change measure was unconfounded from any nonspecific effect (e.g., arousal or fatigue) that lasted longer than an individual trial. We excluded trials with less than 20% valid pupil samples from analyses (children: $M = 5.27\%$, range = 0%–31.67%; adults: $M = 1.93\%$, range = 0%–26.67%).

For each trial, we calculated the average change in pupil diameter for the time interval starting 250 ms after the onset of the results phase until 2,500 ms after the onset to obtain a marker of the surprise response. We chose this time interval based on a previous study (Brod et al., 2018) and in line with the conceptual understanding of surprise as the initial, value-neutral consequence of a perceived discrepancy (Mandler, 1990). We originally intended to explore changes in pupil diameter during the generative task as well, but refrained from doing so because of the confounding influence of differences in image content between the two conditions (Naber & Nakayama, 2013).

Results

Performance Analyses

We discarded trials that participants had known prior to the experiment (children: $M = 0.47\%$, range = 0%–5.00%; adults: $M = 0.27\%$, range = 0%–3.33%) and trials for which no example was found in the example condition (children: $M = 7.73\%$, range = 0%–23.33%; adults: $M = 1.20\%$, range = 0%–6.67%).

Study Phase Performance

On average, children predicted 10.27% of facts correctly ($SD = 5.08\%$; range = 3.33%–23.33%) and young adults predicted 14.93% of facts correctly (*SD* = 4.92%; range = 6.67%–26.67%), the difference being significant ($t(48) = 3.30$, $p = .002$). Children did not find an example in 7.73% of trials (*SD* = 7.25%; range = 0%–23.33%) and young adults could not find an example in 1.20% of trials (*SD* = 2.13%; range = 0%–6.67%). This difference was also significant ($t(48) = 4.33$, $p < .001$).

It took adults 5.61 s (*SD* = 2.14 s), on average, to generate a prediction and 5.52 s (*SD* = 2.06 s) to generate an example, whereas children needed 7.52 s (*SD* = 1.67 s) to generate a prediction and 7.97 s (*SD* = 2.81 s) to generate an example. Because response time (RT) data were positively skewed, we analyzed it using generalized estimating equations (GEE), which are appropriate for the modeling of non-normal distributions. The interpretation of results remained the same regardless of whether GEE, log-transformed data, or a linear model with raw data were used. We discarded three trials for which a RT of 0 s had been recorded. We found a significant effect of age ($\chi^2(1) = 17.10$, $p < .001$), but no significant effect of condition ($\chi^2(1) = 0.37$, $p = .540$), and no interaction ($\chi^2(1) = 0.73$, $p = .390$).

Following the study phase, participants were asked about the examples that they had generated for a subset of 8 items. On average, children and adults generated less than one example that did not fit the item (children: *M* = 0.44, range = 0–2; adults: *M* = 0.08, range = 0–1); they could not think of any example in less than one case (children: *M* = 0.68, range = 0–3; adults: *M* = 0.12, range = 0–1), and could not remember the example they had generated during the task for less than one item (children: *M* = 0.32, range = 0–2; adults: *M* = 0.28, range = 0–2). We, thus, have no indication that finding examples was excessively difficult for either age group.

We converted ratings on the questionnaire into categorical values to make them comparable between age groups since adults had rated the items on a 6-point scale, whereas children had been asked by the experimenter without mention of the response scale. We performed chi-squared tests to see if one answer was more likely within each age group and if the frequencies differed between age groups. Eighteen children ($\chi^2(1) = 4.84$, $p = .028$) and 13 adults ($\chi^2(1) = 0.04$, $p = .842$) reported that they found learning in the prediction condition more enjoyable than in the example condition. The distributions were not significantly different between age groups ($\chi^2(1) = 2.12$, $p = .145$). Ten children ($\chi^2(1) = 1.00$, $p = .317$) and 14 adults ($\chi^2(1) = 0.67$, $p = .414$) thought that they had learned more in the prediction condition. The distributions did not differ significantly ($\chi^2(1) = 1.65$, $p = .199$). 75% of adults and 56% of children were accurate in their judgment regarding which condition they learned better in. The difference between age groups was not significant ($t(46.59) = 1.40$, $p = .168$). However, while adults' judgments were above chance ($t(23) = 2.77$, $p = .011$), children's were not ($t(24) = 0.59$, $p = .559$).

Test Phase Performance

We assessed memory performance during the test phases of each GLS condition. Our main outcome measure was the percentage of facts for which the correct number was recalled. We computed a logistic mixed-effects model with GLS condition (prediction vs. example), age group (children vs. young adults), and their interaction as fixed effects. Figure 2a depicts the percentage of correctly remembered facts, separately for the two conditions and for the two age groups. In line with our hypotheses, we found a significant interaction between age and GLS condition ($\chi^2(1) = 4.02$, $p = .045$, OR = 1.496). Post-hoc tests revealed that children benefitted significantly more from generating predictions ($\chi^2(1) = 12.36$, $p < .001$, OR = 1.75), whereas the performance difference between GLS conditions was not significant in adults ($\chi^2(1) = 0.98$, $p = .322$, OR = 1.15). Both main effects were significant as well, that is, children performed significantly worse on the memory test than adults ($\chi^2(1) = 29.15$, $p < .001$, OR = 0.29) and performance was overall better after generating a prediction than after generating an example ($\chi^2(1) = 10.77$, $p = .001$, OR = 1.15).

In addition to percentage of correctly remembered facts, we also explored the absolute difference between the recalled and the correct number as outcome measure, which provides a more fine-grained (nondichotomous) measure of learning. The pattern of results was highly similar (Figure 2b): The interaction of age and condition was significant in the expected direction ($\chi^2(1) = 4.98$, $p = .026$, $R^2 = .002$), with children showing better memory performance (i.e., smaller difference scores) in the prediction (1.12 ± 0.38) than in the example condition (1.67 ± 0.75) compared to adults (prediction: 0.53 ± 0.22, example: 0.75 ± 0.37). Here, post-hoc tests revealed that the difference between GLS conditions, although smaller than in children, was also significant in adults ($\chi^2(1) = 8.37$, $p = .003$, $R^2 = .006$). The main effects of age ($\chi^2(1) = 34.04$,

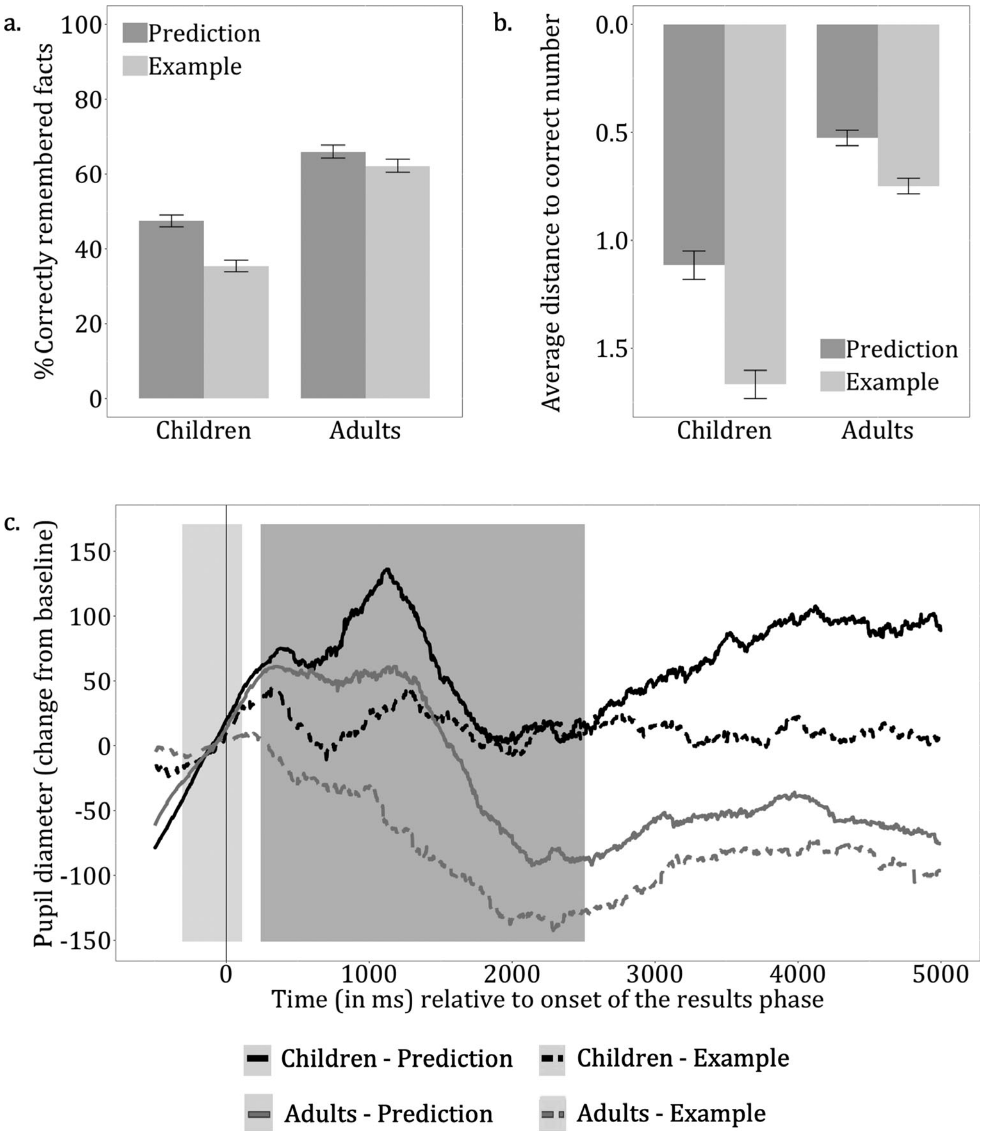

Figure 2. Results of the numerical facts learning task. (a) Children remembered a higher percentage of facts after generating a prediction rather than an example, whereas both generative learning strategy (GLS) were similarly effective in adults. (b) The pattern of results remained the same when using absolute difference scores between the recalled and the correct number as an outcome measure. Here, the difference between GLS conditions was significant in adults, albeit smaller than in children. Error bars represent within-subject standard error. (c) Full time series of the pupillary response to seeing the correct result. The average difference in pupil diameter relative to the baseline (light gray area) was calculated for the time interval from 250 to 2,500 ms after the onset of the results phase (dark gray area).

p < .001, R^2 = .012) and GLS condition ($\chi^2(1)$ = 22.99, p < .001, R^2 = .008) were significant, too.

Exploratory Analyses

Pupillary Data

We first compared the pupillary response upon seeing the correct result between the GLS conditions to determine whether generating predictions boosted surprise. To this end, we subjected the differences in pupil diameter from baseline to a logistic mixed-effects regression that included GLS condition, age group, and their interaction as fixed effects. As can be seen from the time series shown in Figure 2c, the average pupillary surprise response was greater in the prediction condition than in the example condition ($\chi^2(1)$ = 13.97, p < .001, R^2 = .01), and children had overall larger pupil dilations than adults ($\chi^2(1) = 11.15$, $p = .001$, $R^2 = .00$). There was no interaction between condition and age group ($\chi^2(1) = 1.65$, $p = .199$, $R^2 = .00$), suggesting that generating a prediction elicited a surprise response in both age groups.

We next tested whether the pupillary response was differentially predictive of learning in the two age groups and conditions (cf. Figure S1). To this end, we calculated the average difference in pupil size between remembered and forgotten facts (i.e., subsequent memory effects; SMEs), separately for both GLS conditions. Data of three children and two adults were discarded because of too few trials (< 4 per cell). We then performed a mixed-design ANOVA with GLS condition as a within-subject factor and age group as a between-subject factor. The results indicated a trend for the Age × Condition interaction ($F(1, 43) = 3.10$, $p = .086$, $\eta_p = .07$), whereas there was no significant main effect of age group ($F(1, 43) = 0.17$, $p = .683$, $\eta_p = .00$), and a trend for larger SMEs in the prediction condition than in the example condition ($F(1, 43) = 2.92$, $p = .095$, $\eta_p = .06$). We performed post-hoc *t*-tests (Bonferroni-corrected), which revealed larger SMEs in the prediction condition than in the example condition in children ($t(21) = 2.59$, $p = .017$), and no condition effect in adults ($t(22) = -0.01$, $p = .994$).

Predictors of the GLS Effect

To explore potential predictors of the observed performance difference between the two GLS conditions in children, we calculated the difference in memory performance between the prediction and example condition for each child. We first tested whether differences in pupillary SME between the prediction and example condition were related to the difference in memory performance. We observed a positive correlation between the two ($r = .50$, $p = .008$): Children with a larger pupillary SME in the prediction condition relative to the example condition also showed a bigger memory boost in the prediction condition. This finding provides further evidence for the important role of surprise in mediating the beneficial effect of generating a prediction on learning in children.

We then explored why generating an example was not as effective as generating a prediction in children. We expected good analogical reasoning to be particularly important for example-based learning, resulting in smaller performance differences for children with better reasoning abilities. We thus tested whether performance in the reasoning task was linked to performance in the example condition and to the performance difference (directional tests). Scores on the SON-R as a measure of analogical reasoning abilities were obtained by adding the number of items answered correctly in each set minus the number of errors. One child's analogical reasoning score lay more than 1.5 interquartile ranges under the 25% quartile and was therefore discarded (Tukey, 1977). There was a close correlation between analogical reasoning and performance in the example condition ($r = .52$, $p = .005$). As expected, the follow-up test on the link between analogical reasoning and the performance difference between GLS conditions revealed that children with better reasoning abilities performed more similarly in the two GLS conditions ($r = -.36$, $p = .044$; see Figure 3a). Together, these findings suggest that good analogical reasoning abilities are an important prerequisite for the benefits of generating examples to occur, and thus partially explain the performance difference found in children.

We further explored EFs as a potential moderator due to their importance for learning complex material (e.g., Zaitchik, Iqbal, & Carey, 2014) and because, similar to reasoning abilities, they also display strong improvements during late childhood and adolescence (Diamond, 2013). In the HFT, the incongruent block and the mixed block assess inhibitory control and cognitive flexibility respectively. We calculated the average RT across both blocks to obtain a combined measure of EFs. We controlled for processing speed by calculating relative RT difference scores ($[x - y]/y$) through subtracting the mean RT of trials in the congruent block. To obtain a score in which higher values reflect better performance, we subtracted this value from 1. The correlation of this score with the performance difference was not significant ($r = .03$, $p = .887$).

To explore the distinctive effects of analogical reasoning and EFs on the condition effect in children, we computed a logistic mixed-effects regression analysis on memory performance with condition on Level 1, analogical reasoning and HFT score on Level 2 as well as the cross-level interactions of Condition × Analogical Reasoning and Condition × HFT. The interaction of Condition × Analogical Reasoning was significant ($\chi^2(1) = 4.00$, $p = .045$), with reasoning ability having a stronger impact on performance in the example condition (see Figure 3b). The Condition × HF interaction was not significant ($\chi^2(1) = 0.11$, $p = .743$) and there were no significant main effects except for the effect of condition ($\chi^2(1) = 6.37$, $p = .012$). These results reinforce the hypothesis that analogical reasoning is a prerequisite for good example-based learning.

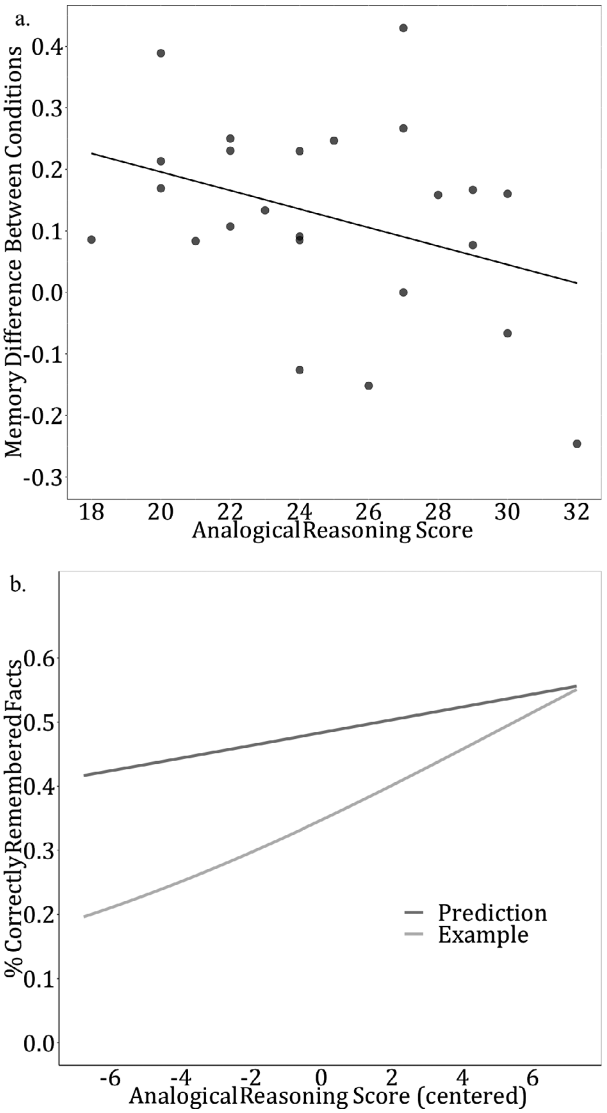

Figure 3. Children's analogical reasoning ability predicts the difference in effectiveness between the learning strategies. (a) Children's performance difference between generative learning strategy conditions was negatively correlated with their analogical reasoning abilities ($r = -.36$, $p = .044$). (b) Analogical reasoning predicted performance differences between the two conditions, also when controlling for executive functions. Estimated values indicate that reasoning had a stronger impact on performance in the example condition. Children with higher reasoning abilities, thus, performed more similarly in the two conditions than children with lower reasoning abilities.

Discussion

This study revealed an Age × Learning Strategy Type interaction for the effectiveness of GLSs. As hypothesized, children remembered more facts after generating predictions than after generating examples, and this difference was significantly greater than in adults. Furthermore, compared to generating examples, generating predictions was associated with a larger pupillary surprise response upon seeing the correct result. This pupillary response was a better predictor of children's learning in the prediction condition than in the example condition, and this effect correlated with the behavioral performance difference. Taken together, these findings suggest that prediction-induced surprise promoted children's learning. Also in line with our hypotheses, we found that children's reasoning abilities were positively related to their performance in the example condition and negatively with the performance difference between the prediction and example condition. That is, the better (i.e., more mature) their reasoning abilities, the more do children resemble adults in that generating examples is similarly effective than generating predictions. In summary, the results support our hypothesis that there are distinct cognitive prerequisites for generating predictions and generating examples, which result in different degrees of effectiveness of these strategies for different age groups.

Our findings provide a proof of concept for the importance of cognitive prerequisites for understanding individual differences as well as age-related trends in the effectiveness of GLSs. Our study bridges the gap between research comparing the effectiveness of different GLSs within the same age group (e.g., Brod et al., 2018; Ritchie & Volkl, 2010; Yeo & Fazio, 2019) and research comparing the effectiveness of the same GLS between different age groups (e.g., Gurlitt & Renkl, 2008). The former approach has identified specific mechanisms underlying different GLSs (e.g., Brod et al., 2018), whereas the latter approach has revealed the need for instructional adaptations to support younger learners (Gurlitt & Renkl, 2008). Only the combination of the two approaches, however, allows to investigate why different GLSs might be differentially effective for learners of various ages and with varying cognitive abilities.

Our findings regarding the mechanisms underlying the effectiveness of generating predictions in children correlate well with previous research in adults. In a recent study, Brod et al. (2018) used pupillometry data to demonstrate that predictions trigger surprise for expectancy-violating events and that, on a between-subject level, the strength of this surprise effect relates to better learning. Our pupillary results are consistent with these findings in that pupil dilations upon seeing the correct results were larger after generating predictions than after generating examples. Our findings extend the ones by Brod et al. (2018) by demonstrating a similar pupillary surprise response in children, which supports the notion that surprise is an age-invariant mechanism (Schützwohl & Reisenzein, 1999) that can be triggered by generating predictions. Moreover, the link between children's surprise response and subsequent memory predicted children's performance difference. These results are in line with research that suggests that surprise increases attention to task-relevant information (Fazio & Marsh, 2009; Stahl & Feigenson, 2019), thereby enhancing learning. In a different experiment (Brod & Breitwieser, 2019), we were able to show that generating a prediction further increases attention to one's knowledge gap, which leads to increased curiosity for the correct answer. Taken together, the results of our experiments suggest that the effectiveness of generating predictions is at least partially mediated by enhanced attention to the new information.

Generating predictions is arguably one of the simplest possible GLSs. The results of the present study attest to the notion that it is this simplicity that makes generating predictions particularly effective in children, whose limited attentional and cognitive control functions impede the effectiveness of more complex strategies. To improve our understanding of this strategy's underlying mechanisms, future studies should address the link between prediction errors, surprise, and learning more explicitly as generating predictions does not necessarily require prior knowledge activation. Its effectiveness likely relies on where on the continuum from mere guessing to effortful prior knowledge retrieval the generated prediction lies, as prediction errors made with greater confidence should increase the experience of surprise, which should in turn enhance learning. A previous study by Brod, Breitwieser, Hasselhorn, and Bunge (2019) suggests that children are not always able to leverage the benefit of surprise for learning, however. Metacognitive skills might need to be in place to override a previous held belief with the correct information (Brod et al., 2019). In the current study, the accuracy of children's retrospective judgments of task performance was not significantly above chance, in contrast with the adults'. This finding fits our knowledge of the protracted development of procedural metamemory (i.e., monitoring and regulation of memory performance; Fritz, Howie, & Kleitman, 2010; Schneider, 2008). Future studies should look into the mediating role of procedural metamemory for the effect of surprise on learning more closely by measuring metacognitive judgments during task performance.

We also found evidence that generating examples can be an effective learning strategy. Adults utilized examples almost as effectively as predictions. The effectiveness of generating examples in children was linked to their analogical reasoning abilities: Children with better reasoning abilities showed a smaller performance decrease in the example condition (i.e., an adult-like performance pattern). Our results, thus, suggest that analogical reasoning abilities are an important and distinct prerequisite for the benefits of example generation to occur. The late development of analogical reasoning abilities (e.g., Richland et al., 2006) can then explain why, on average, children did not meet the cognitive requirements to utilize examples as effectively as adults. Although it has been argued before that analogical reasoning might underlie the successful use of examples (Zamary & Rawson, 2018), to our knowledge, this is the first study that explicitly tested the moderating effect of analogical reasoning abilities for the effectiveness of example-based learning. While the correlational results clearly do not allow to infer a causal relation between analogical reasoning and example-based learning, they are a good starting point for future research to test their causal relation via intervention studies.

Unlike previous research, the present study tested example generation not as a means of learning declarative concepts (e.g., Rawson & Dunlosky, 2016; Zamary & Rawson, 2018), but as a means of learning isolated facts. Thus, in our study, the primary purpose of the examples was not to make sense of an abstract concept but to associate the new information with a self-generated retrieval cue, which has been shown to aid recall (Greenwald & Banaji, 1989). It proved similarly effective to generating predictions in the university students. Nevertheless, one might argue that the task design was not ideal for the benefits of example generation to occur. Specifically, unlike the predictions, the examples did not directly relate to the to-be-learned information (i.e., the numbers). While example generation still required elaboration of task-relevant information, the appropriateness of the examples was not determined by the correct number. It remains an open question to what extent this has dampened the effectiveness of generating examples.

Since the present study did not include a control condition without any learning strategy prompt, we cannot make inferences about how much performance was boosted by generating predictions or generating examples compared to "baseline" performance. Such a baseline condition would be difficult to implement because university students can be expected to use effective learning strategies without being prompted to—in accordance with their knowledge of strategy effectiveness (Justice & Weaver-McDougall, 1989). Elementary school children, in contrast, are unlikely to spontaneously use effective learning strategies, and large interindividual differences are to be expected due to the ongoing development of metamemory abilities (Bjorklund, 2010; Bjorklund & Coyle, 1995). We, thus, deemed baseline performance to be difficult to compare between age groups and decided to not include a baseline condition in our within-subjects design. Having said that, however, an interesting question for future research could be to elucidate which strategies children and adults spontaneously use to learn facts, whether there are interactions between spontaneous and instructed strategy use (i.e., switching of strategies during the experiment), and how this impacts learning success.

Another limitation of the current study is the sample size, which was determined a priori to be sufficient for testing the hypothesized Age × GLS interaction effect as well as pupillary within-subject condition differences. These preregistered parts of the study were based on a pilot study with 26 children and 18 adults and can, thus, be considered a successful replication. The sample size is too small, however, to ensure reliable estimates of the between-subject correlational analyses. We see these exploratory analyses as a first step toward exploring the specific mechanisms and cognitive requirements of different GLSs to explain age-related performance differences. Conceptual replications are required to test the generalizability of the results to other populations, strategies, and task types. Including additional measures of cognitive abilities would further allow to test the specificity of the effects found in the present study. Finally, the time span between the study and test phases was rather short. We are therefore unable to draw conclusions about the long term effects of generating predictions and generating examples.

In closing, the present study demonstrates the importance of considering learners' cognitive prerequisites for selecting the most effective GLS. Leveraging an age-group comparison and knowledge of typical developmental trajectories of specific cognitive abilities, this study suggests that different GLS can differ strongly in effectiveness depending on maturity of these abilities. Thus, to achieve the goal of selecting optimal learning strategies for a particular group of learners or even for individual learners, knowledge of their cognitive abilities as well as knowledge of the cognitive prerequisites of a specific GLS are needed. While this goal seems distant still, it is becoming clear that, rather than proclaiming the most effective learning strategy for all learners, the effectiveness of any learning strategy will critically depend on the fit between its cognitive requirements and the cognitive prerequisites of the learner.

References

Alberini, C. M., & LeDoux, J. E. (2018). Memory reconsolidation. *Current Topics in Behavioral Neurosciences*, *37*, 151–176. https://doi.org/10.1007/7854_2016_463

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*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Bartoń, K. (2018). MuMIn: Multi-model inference. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn

Bjork, R. A. (1994). Memory and metamemory considerations in training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–206). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.

Bjorklund, D. F. (2010). Remembering on their own: The development of strategic memory. In H.-P. Trolldenier, W. Lenhard, & P. Marx (Eds.), *Das Gedächtnis entwickeln: Entwicklungs- und pädagogisch-psychologische Forschungen zum Gedächtnis (The development of memory: Research on memory in the fields of developmental and educational psychology)* (pp. 171–189). Göttingen, Germany: Hogrefe Verlag.

Bjorklund, D. F., & Coyle, T. R. (1995). Utilization deficiencies in the development of memory strategies. In E. F. Weinert & W. Schneider (Eds.), *Memory performance and competencies: issues in growth and development* (pp. 161–180). Mahwah, NJ: Erlbaum.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*, 389–411. https://doi.org/10.1037/met0000159

Brod, G., & Breitwieser, J. (2019). Lighting the wick in the candle of learning: generating a prediction stimulates curiosity. *NPJ Science of Learning*, *4*. https://doi.org/10.1038/s41539-019-0056-y

Brod, G., Breitwieser, J., Hasselhorn, M., & Bunge, S. A. (2019). Being proven wrong elicits learning in children—But only in those with higher executive function skills. *Developmental Science*. https://doi.org/10.1111/desc.12916

Brod, G., Bunge, S. A., & Shing, Y. L. (2017). Does one year of schooling improve children's cognitive control and alter associated brain activation? *Psychological Science*, *28*, 967–978. https://doi.org/10.1177/0956797617699838

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

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1491–1494. https://doi.org/10.1037/0278-7393.27.6.1491

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

Chularut, P., & DeBacker, T. K. (2004). The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language. *Contemporary Educational Psychology*, *29*, 248–263. https://doi.org/10.1016/j.cedpsych.2003.09.001

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750.Executive

Duit, R. (1991). On the role of analogies and metaphors in learning science. *Science Education*, *75*, 649–672. https://doi.org/10.1002/sce.3730750606

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques. *Psychological Science in the Public Interest*, *14*, 4–58. https://doi.org/10.1177/1529100612453266

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using GPower 3.1: Tests for correlation and regression analyses. *Behavior Repsearch Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

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

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*, 717–741. https://doi.org/10.1007/s10648-015-9348-9

Fox, C. J., Mueller, S. T., Gray, H. M., Raber, J., & Piper, B. J. (2013). Evaluation of a short-form of the Berg Card Sorting Test. *PLoS One*, *8*, 6–9. https://doi.org/10.1371/journal.pone.0063885

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 1–7. https://doi.org/10.3389/fpsyg.2012.00130

Fritz, K., Howie, P., & Kleitman, S. (2010). "How do I remember when I got my dog?" The structure and development of children's metamemory. *Metacognition and Learning*, *5*, 207–228. https://doi.org/10.1007/s11409-010-9058-0

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28*, 210–216. https://doi.org/10.1037/a0014474

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277–300. https://doi.org/10.1016/S0364-0213(86)80019-2

Gorrell, J., Tricou, C., & Graham, A. (1991). Children's short-and long-term retention of science concepts via self-generated examples. *Journal of Research in Childhood Education*, *5*, 100–108. https://doi.org/10.1080/02568549109594807

Greenwald, A. G., & Banaji, M. R. (1989). The self as a memory system: Powerful, but ordinary. *Journal of Personality and Social Psychology*, *57*, 41–54. https://doi.org/10.1037/0022-3514.57.1.41

Gurlitt, J., & Renkl, A. (2008). Are high-coherent concept maps better for prior knowledge activation? Differential effects of concept mapping tasks on high school vs. university students. *Journal of Computer Assisted Learning*, *24*, 407–419. https://doi.org/10.1111/j.1365-2729.2008.00277.x

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory and Cognition*, *40*, 514–527. https://doi.org/10.3758/s13421-011-0167-z

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, *12*, 670–679. https://doi.org/10.1111/j.1467-7687.2008.00805.x

Justice, E. M., & Weaver-McDougall, R. G. (1989). Adults' knowledge about memory: Awareness and use of memory strategies across tasks. *Journal of Educational Psychology*, *81*, 214–219. https://doi.org/10.1037/0022-0663.81.2.214

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. https://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*, 1068–1078. https://doi.org/10.1111/ejn.12859

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*, 609–639. https://doi.org/10.1037/0033-295X.100.4.609

Lee, H., Lim, K. Y., & Grabowski, B. L. (2008). Generative learning: Principles and implications for making meaning. In J. Spector, D. M. Merrill, J. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research and educational* *communications and technology* (3rd ed., pp. 111–124). New York, NY: Taylor & Francis Group.

Levy, R. (2014). Using R formulae to test for main effects in the presence of higher-order interactions. (pp. 1–7). Retrieved from http://arxiv.org/abs/1405.2094

Li, S., Lindenberger, L., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*, 155–162. https://doi.org/10.1111/j.0956-7976.2004.01503003.x

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: Erlbaum.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin and Review*, *14*, 225–229. https://doi.org/10.3758/BF03194056

Naber, M., & Nakayama, K. (2013). Pupil responses to high-level image content. *Journal of Vision*, *13*, 1–8. https://doi.org/10.1167/13.6.7.doi

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

Piaget, J. (1926). *The language and thought of the child*. London, UK: Rutledge and Kegan Paul.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667. https://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*, 1–12. https://doi.org/10.3389/fnins.2011.00115

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rawson, K. A., & Dunlosky, J. (2016). How effective is example generation for learning declarative concepts? *Educational Psychology Review*, *28*, 649–672. https://doi.org/10.1007/s10648-016-9377-z

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*, 243–257. https://doi.org/10.1037/a0016496

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*, 249–273. https://doi.org/10.1016/j.jecp.2006.02.002

Ritchie, D., & Volkl, C. (2010). Effectiveness of two generative learning strategies in the science classroom. *School Science and Mathematics*, *100*, 83–89. https://doi.org/10.1111/j.1949-8594.2000.tb17240.x

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*, 114–121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

Schützwohl, A., & Reisenzein, R. (1999). Children's and adults' reactions to a schema-discrepant event: A developmental analysis of surprise. *International Journal of Behavioral Development*, *23*, 37–62. https://doi.org/10.1080/016502599383991

Sinclair, A. H., & Barense, M. D. (2018). Surprise and destabilize: Prediction error influences episodic memory reconsolidation. *Learning and Memory*, *25*, 369–381. https://doi.org/10.1101/lm.046912.117

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*, 679–692. https://doi.org/10.1002/wcs.1323

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

Stahl, A. E., & Feigenson, L. (2019). Violations of core knowledge shape early learning. *Topics in Cognitive Science*, *11*, 136–153. https://doi.org/10.1111/tops.12389

Tellegen, P. J., Laros, J. A., & Petermann, F. (2012). *Snijders-Oomen Nonverbal Intelligence Test (SON-R 6–40)*. Göttingen, Germany: Hogrefe.

Tukey, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science*, *198*, 679–684. https://doi.org/10.1126/science.333584

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.). Cambridge, MA: Harvard University Press.

Wittrock, M. C. (2010). Learning as a generative process. *Educational Psychologist*, *45*, 40–45. https://doi.org/10.1080/00461520903433554

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology*, *5*, 1–9. https://doi.org/10.3389/fpsyg.2014.00213

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology*, *111*, 73–90. https://doi.org/10.1037/edu0000268

Zaitchik, D., Iqbal, Y., & Carey, S. (2014). The effect of executive function on biological reasoning in young children: An individual differences study. *Child Development*, *85*, 160–175. https://doi.org/10.1111/cdev.12145

Zamary, A., & Rawson, K. A. (2018). Which technique is most effective for learning declarative concepts—Provided examples, generated examples, or both? *Educational Psychology Review*, *30*, 275–301. https://doi.org/10.1007/s10648-016-9396-9

Supporting Information

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

Figure S1. Pupillary Response to Seeing the Correct Result, Separately for Later Remembered and Forgotten Facts

Table S1. Sample of Items Used in the Numerical Facts Learning Task (Translated From German)