Overoptimistic predictions and well-calibrated expectations: Children *say* they will achieve unrealistic outcomes but are *surprised* when they do

Adani Bennett Abutto, Misha O'Keeffe & Hyowon Gweon
{aabutto, mokeeffe, gweon}@stanford.edu
Department of Psychology, Stanford University, Stanford, CA, USA

Abstract

Accurately predicting one's own performance outcomes is a crucial skill for children and adults alike. Prior research, however, has shown that young children are notoriously overoptimistic, making predictions far beyond their actual performance. Curiously, these findings contradict decades of work showing that infants and children hold reasonable expectations about the world, showing surprise—an indication of prediction error (PE)—when events violate their expectations. If children have well-calibrated expectations about their own performance, they might experience PE when they produce unrealistically good outcomes. Using a probability-based game (Experiment 1, N=48) and a memory-based game (Experiment 2, N=64), we show that preschoolers are indeed overoptimistic in their explicit predictions, but express surprise after achieving precisely those unrealistic performance outcomes they predicted. These results demonstrate an early-emerging sensitivity to prediction error about the self, revealing a striking discrepancy between what children *say* they can do and what they *think* they can do.

Keywords: cognitive development; overoptimism, prediction error; surprise

Introduction

In everyday life, we often think about how we might do on upcoming tasks: Can I win this treasure hunt? Can I make it down this double-black ski run? Can I pass this math exam? By making accurate predictions about the likely outcomes of our own actions, we can make better decisions about what to do next, whether to seek help, and how to improve our skills.

Reasoning about performance outcomes is an important skill even for young children, who frequently face novel tasks while also undergoing rapid changes in their motor and cognitive skills. A task that seems feasible at first may in fact be out of reach (e.g., opening a jar that is stuck), and what appears well beyond their capabilities may become readily achievable only days later (e.g., learning how to tie shoelaces). The current work explores the development of this foundational capacity: Can young children generate well-calibrated expectations about their own performance outcomes?

A long-standing body of work on children's overoptimism has much to say about this question: Dozens of empirical studies have demonstrated robust, persistent overoptimism in children, especially in early to middle childhood, suggesting they have a limited understanding of their own abilities (for reviews, see Leonard & Sommerville, 2025; Xia et al., 2024). Children's tendency to make overly optimistic predictions—also referred to as optimism bias—is present in a range of domains and surprisingly resistant to practice and feedback. For instance, even after predicting how far they can jump and subsequently observing that their actual performance falls well short of their initial prediction, 4- to 6-year-old children continue to make unrealistic, inflated predictions about their future performance (Schneider, 1998; Xia et al., 2022, 2023). Similarly, children this age continue to overestimate how much information they can memorize and recall even after receiving feedback on their true (poor) performance (Flavell et al., 1970; Lipko et al., 2009; Lipko-Speed, 2013; Schneider, 1998; Xia et al., 2023; Yussen & Levy, 1975). Beyond these skill-based domains, overoptimism shows up even in probability-based tasks; children consistently overestimate the probability of a desirable outcome, especially when the outcome results in a reward for themselves versus for someone else (Hennefield & Markson, 2022; Wente et al., 2020). In sum, this literature suggests that children's predictions and expectations relating to themselves are fraught with bias and a lack of calibration.

Yet, prior work on metacognitive development suggests that young children are not completely ignorant about their own performance. Even 20-month-olds selectively ask for help when they are uncertain about the location of an item (Goupil et al., 2016), and children become increasingly better at monitoring their uncertainty; by age 4, children can report their level of confidence and strategically decide whether to skip a question or exclude an answer from being evaluated (e.g., Hembacher & Ghetti, 2014; Lyons & Ghetti, 2011, 2013). Notably, some studies have found that children can be overconfident but nonetheless choose to revise their belief (Hagá & Olson, 2017) or seek additional information (Lapidow et al., 2022), suggesting a potential discrepancy between explicit predictions and other behavioral markers of uncertainty (cf. Destan & Roebers, 2015). While these findings involve judgments made during or after the task and do not provide direct evidence for children's ability to *predict* their performance outcomes, collectively this literature raises the possibility that despite overoptimistic predictions, children have a better internal sense of how they might do on a task.

Children's difficulty with predicting their own performance outcomes also stands in stark contrast to decades of prior work documenting young children's ability to make accurate predictions about the world. Infants look longer at physically impossible events (e.g., a ball appearing to pass through a wall; Spelke et al., 1992; Stahl & Feigenson, 2015) or statistically

improbable events (Xu & Garcia, 2008), suggesting that they had *a priori* expectations about the outcomes of these events. They also hold expectations about their social world, showing surprise when agents do not act efficiently with respect to their goals (e.g., Gergely & Csibra, 2003; Liu & Spelke, 2017) or approach someone mean vs. nice (Kuhlmeier et al., 2003). Such systematic responses to unexpected outcomes are consistent with prediction error (PE)—a signal that indicates a discrepancy between an actual outcome and an expected outcome (Clark, 2013; Hohwy, 2020; Niv, 2009; Pessiglione et al., 2006; Schultz et al., 1997). Beyond looking time, surprising events that elicit PE can also give rise to systematic changes in facial expressions, pupil dilation, and other physiological measures (e.g., Camras et al., 2002; Gredebäck & Melinder, 2010; Moll et al., 2016; Ni et al., 2023). These indices of PE provide evidence that humans, starting early in life, can hold rich, sophisticated expectations about the world.

Collectively, these findings raise an intriguing possibility: Even though young children make overoptimistic predictions, they might still have reasonable expectations about their own performance. While extensive work has documented how children respond to events that violate their expectations about the external world, little work has examined violations of expectations about *the self*—i.e., surprising outcomes resulting from their own choices or actions. Insofar as PE serves as a domain-general learning signal, it might manifest not only when children passively observe surprising outcomes, but also when they actively produced those surprising outcomes. Granted, it is not always easy to distinguish surprise about the world vs. surprise about ourselves, as the efficacy of our own actions often depends critically on the properties of the external world. Yet, at least in the sensory motor domain, prior work has identified evidence for 'self-PE'—error signals in the brain that plausibly pertain to the outcomes of one's own actions and their sensory consequences (e.g., Blakemore et al., 1998; Knolle et al., 2013; Wolpert et al., 1998). Furthermore, recent work suggests that unexpected success increases exploratory play in young children (Doan et al., 2020), suggesting a potential role of self-PE.

Inspired by these findings, we hypothesize that although children make overoptimistic verbal predictions about the outcomes of their own actions, they might report being surprised after having achieved those unrealistic outcomes. This hypothesis is rather counterintuitive: We predict children will be surprised even when the observed outcome is *consistent* with their explicit predictions. If children are surprised by the unrealistic outcomes they themselves predicted to achieve, that constitutes evidence that children can generate internal representations of predicted outcomes that are better-calibrated to their actual competence than their explicit predictions.

To test our prediction, we created an experimental paradigm in which children experienced the same successful outcome under different conditions: In the *Uncued* condition, which served as our main condition of interest, children performed either a highly challenging probability-based (Exp. 1) or a memory task (Exp. 2) without any visible aids, similar to setups in prior work. In the *Cued* condition, which served as our control condition, children performed the same tasks with clearly visible cues such that they were guaranteed to succeed. This design allows us to directly compare children's surprise responses across two conditions that vary in the likelihood of a desirable outcome while holding the outcome itself constant. Importantly, the successful outcome they experience under this setup is one that is aligned with their explicit overoptimistic predictions. We conduct our initial investigation with 4- to 6-year-olds, tapping the earliest ages at which prior studies have examined children's explicit predictions about their performance outcomes (Leonard & Sommerville, 2025; Xia et al., 2024). See https://osf.io/9bcjw/ for data and code.

Experiment 1

In Exp. 1, children played a simple probability-based card game and achieved a desirable outcome. We asked whether children express more surprise when the desirable outcome was highly improbable (Uncued) than when it was fully expected (Cued). Critically, drawing on prior work on children's overoptimism in probability-based predictions (Hennefield & Markson, 2022; Wente et al., 2020), we also sought evidence for overoptimism in children's explicit predictions.

Methods

Participants Forty-eight 4- and 5-year-olds ($M_{age}(SD)$ = 4.92 (0.68) years, Range: 4.02-6.00 years; 54% girls; 52% white) were recruited online via Children Helping Science (CHS; Scott and Schulz, 2017) and randomly assigned to either Uncued or Cued conditions (n = 24 in each). An additional 22 participants were excluded due to incompletion (n = 6) and interference/distraction (n = 11), or selecting non-target cards in the Cued condition (n = 5; see Procedure).[1] No child failed the attention check. All families were compensated with a $5 gift card for their time.

Procedure See Figure 1 for an overview of the procedure. Children were first introduced to a set of 15 cards arranged in three rows of five. Three cards (20%) had a star, while the remaining 12 (80%) did not. The distribution of star and non-star cards was highlighted using flashing animations and verbal emphasis in the accompanying voiceover. The goal of the game was to find all three stars; children answered an attention check question with feedback, ensuring that they understood this goal. The cards were then flipped over and shuffled on screen, and the game unfolded in four phases.

In the **Choice** phase, children picked one card at a time, yielding a total of 3 chosen cards. Each time they chose a card, their choice was marked and held for later reveal. After their final pick came the **Prediction** phase, where children were asked, 'Do think you got no stars, 1 star, 2 stars, or 3 stars?' (increasing/decreasing order of card numbers counterbalanced). Children responded by clicking an image cor-

[1] This exclusion rate is typical in asynchronous online studies.

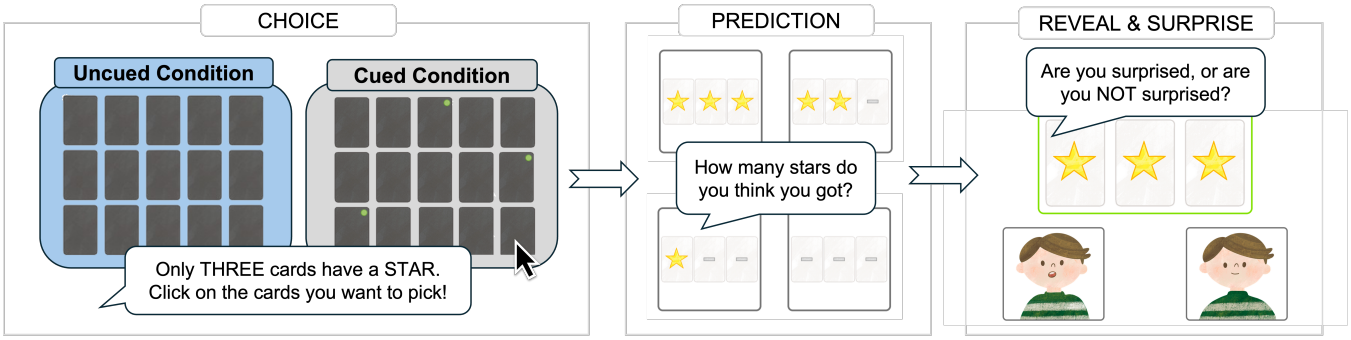

Figure 1: Overview of task procedure in Exp. 1 (probability-based game). The game involved choosing 3 cards from a deck of 15 cards (face-down), only 3 of which had a star. In the Uncued condition, the cards were shuffled and children had to guess; in the Cued condition, the cards had visible cues (green dots) identifying the star cards. After choosing 3 cards (Choice phase), children indicated how many star cards they think they found (Prediction Phase). Once the outcome was revealed, which always showed 3 star cards (Reveal Phase), children indicated their level of surprise (Surprise Phase). In Exp. 2, children played a variant of this game where they had to remember the locations of star cards (see Exp. 2 Procedure).

responding to their desired option. In the **Reveal** phase, the outcome of children's choices was revealed: It was rigged such that children in both conditions saw 3 stars regardless of their choice in the Choice phase. In the final **Surprise Response** phase, children were asked: 'Are you surprised or are you not surprised that all 3 cards you chose are stars?' and responded by clicking an image of a child with a surprised or neutral expression (order counterbalanced). Those who responded 'surprised' were prompted to indicate their level of surprise using a similar pictorial scale: 'Are you a little surprised, pretty surprised, or very surprised that all 3 cards you chose are stars?' (order of options counterbalanced).

The procedure was identical across the two conditions, except for one crucial difference: During the Choice phase, children in the Uncued condition saw identical images of cards, forcing them to choose randomly, whereas those in the Cued condition were told that the star cards have a salient cue in the back, allowing them to know which ones are star cards.[2]

Results and Discussion

In the Cued condition, children almost certainly *knew* they would get 3 stars, and all 24 children appropriately predicted such; these responses are irrelevant to overoptimism. In the Uncued condition, however, these responses would indicate overoptimism. Replicating prior work, a majority of children in the Uncued condition said they would find 3 stars (75%, 18 of 24). While some 'hedged' their predictions (i.e., other responses included 2 (n = 1), 1 (n = 1), and 0 (n = 4) stars), the average number of stars ($M(SD)$ = 2.38 (1.17)) far exceeded the expected value assuming random choice (0.6 stars): $t(23)$ = 7.42, $p < .001$ (one sample t-test). See Figure 2A.

Next, we turned to our key analysis: Given the same positive outcome (3 stars), were children more likely to express surprise when the outcome was highly improbable (Uncued condition) than when it was anticipated (Cued condition)? We fit a logistic regression model with condition as a predictor and found a significant effect: Children in the Uncued condition were more likely to report surprise than children in the Cued condition (Uncued vs. Cued: 66.7% vs. 37.5%, β = 1.20, p = .046). Children's reported level of surprise mirrored this pattern (Uncued vs. Cued ($M(SD)$): 1.75(1.36) vs. 0.71(1.12)): β = 1.44, p = .013, ordinal logistic regression. No effect of age was observed in either measure. Importantly, the effect of condition was reduced but largely remained even after excluding children in the Uncued condition who predicted fewer than 3 stars (n = 6); it was marginally significant in the binary surprise measure (β = 1.20, p = .066) and remained significant in children's level of surprise (β = 1.42, p = .022). Thus, the condition difference was not driven by the few children in the Uncued condition who made more realistic predictions and were reasonably surprised by the improbable outcome.

In sum, these findings reveal a striking discrepancy between children's explicit predictions and their self-reported surprise about their own action outcomes. While children in the Uncued condition *said* they would achieve a highly improbable outcome—showing the well-documented overoptimism effect—their self-reported surprise suggested otherwise: After observing the outcome, they reported feeling more *surprised* than children in the Cued condition who achieved the same (but expected) outcome. Notably, these results held even when matching children's predictions across conditions.

In our task design, the observed outcome was held constant across conditions; the main difference was whether children's choice was uncued or cued, which affected the likelihood of success, and presumably, their subjective levels of uncertainty during the Choice phase (guessing vs. following cues). The effect of condition in children's surprise responses suggest that they were sensitive to this difference. This is particularly strik-

[2] All but n = 5 children in the Cued condition followed the cues and chose the 3 marked cards. Note that children who chose unmarked cards observed an outcome that deviated from the rest; these participants were thus excluded from analyses; see Participants.

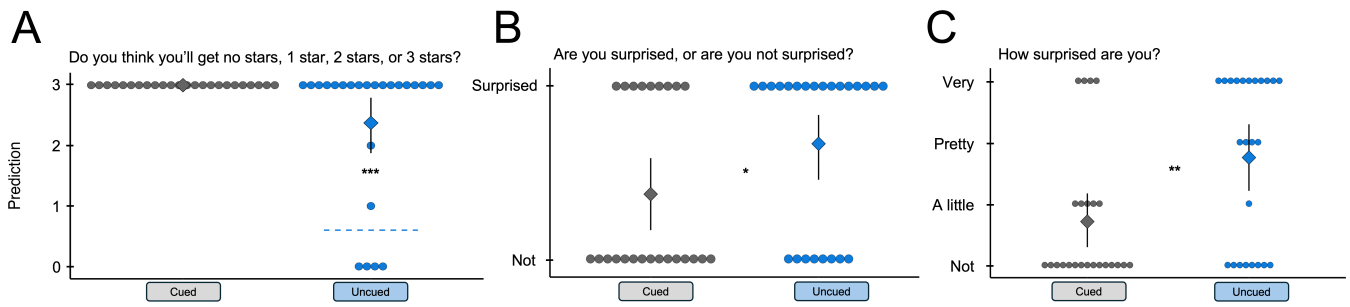

Figure 2: Results for Exp. 1 (probability-based game). Error bars represent bootstrapped 95% CIs. (A) Prediction; dashed line indicates expected value when guessing (.6 stars). (B) Binary surprise. (C) Graded surprise. $^{***}p \leq .001$; $^{**}p \leq .01$; $^{*}p \leq .05$.

ing because all children observed the outcome they said they would achieve, which should, in principle, be unsurprising.

However, one might question whether these findings truly reflect children's prediction error about their own performance outcomes. Given that this was a probability-based game, it is possible that children's surprise purely reflected a response to the statistical nature of the outcome, rather than anything about their own actions or choices. If so, we may not find the same effect in domains where outcomes are mainly driven by the child's competence rather than chance. Furthermore, this initial experiment was unpreregistered with a relatively small sample (N = 48), raising questions about its replicability.

To test the generalizability of our findings, we sought to conceptually replicate the key effects in a skill-based domain: memory. We chose this domain for two reasons. First, the classic effect of overoptimistic performance predictions was first documented using a memory-based task (Flavell, 1970) and was replicated many times (Lipko et al., 2009; Lipko-Speed, 2013; Schneider, 1998; Xia et al., 2022). Second, it was possible to design a memory task very closely matched to the card task used in Exp. 1, minimizing potential concerns about procedural differences between experiments.

Experiment 2

In Exp. 2, we adapted our probability-based task to create a task in a different domain: a memory-based task. As in Exp. 1, we examined children's verbal surprise alongside their explicit, a priori predictions. This study was preregistered (https://aspredicted.org/jcmm-z553.pdf).

Methods

Participants Sixty-four 4- and 5-year-olds were recruited via CHS and randomly assigned to either the Cued or Uncued condition (n = 32 each); $M_{age}(SD)$ = 5.02 (.59) years, Range: 4.02-6.00 years; 44% girls; 61% white. An additional n = 36 children were excluded due to incompletion or missing data (n = 4), interference/distraction (n = 20), failing attention checks (n = 7) and selecting non-target cards in the Cued condition (n = 5; see Footnote 2).[3] Families received a $5 gift card.

Procedure The procedure was similar to Exp.1 (see Fig. 1) except for two major changes. First, we expanded the card set from 15 to 50 cards, displayed in five rows of ten. Second, to tie children's card game performance to skill rather than chance, we adapted the game to (purportedly) rely on memory.

Children were first introduced to the set of 50 cards. Children were then prompted to remember the location of star cards, and then the cards were flipped over. Unlike Exp.1 where the cards were also shuffled, the spatial arrangement of cards remained the same. The game then unfolded in the same four phases: Choice, Prediction, Reveal, and Surprise phases (see Exp.1 Procedure for details). As in Exp.1, the procedure was identical across the two conditions, except for one crucial difference: During the Choice phase, once the cards were flipped over, children in the Uncued condition had to rely on memory to locate the star cards, whereas children in the Cued condition could use a salient cue in the back of the star cards, making them trivially easy to find.

Results and Discussion

Given that this was meant to be a skill-based task, we first sought to confirm that children made genuine attempts at locating the star cards and that our manipulation of success expectancy worked as intended. For children's *incorrect* choices in the Uncued condition, the majority were within 1-card distance from an actual star card (Manhattan distance across 3 card choices: $M(SD)$ = 1.40 (.65). Despite the high level of engagement, most children located either 0 star cards (n = 16) or just 1 star card (n = 14); only 2 children located 2 star cards, and no child remembered all 3 locations. In sum, even though children expended genuine effort, this memory game proved to be quite difficult for them.

Despite the difficulty of the task, did children in the Uncued condition still expect to remember the locations of all 3 stars? Consistent with Exp. 1 and previous literature, 62.5% of children in the Uncued condition (20 of 32) predicted to obtain 3 stars. Although some hedged their bets (6 children predicted 0 stars, n = 4 predicted 1 star, and 2 children predicted 2 stars), children's prediction average (M = 2.13, SD = 1.24 stars) sub-

[3] We deviated from our preregistration which erroneously omitted two important exclusion criteria: attention checks, and selection of non-target cards in the Cued condition. This allowed us to ensure children understood the task and maintain consistency with Exp.1.

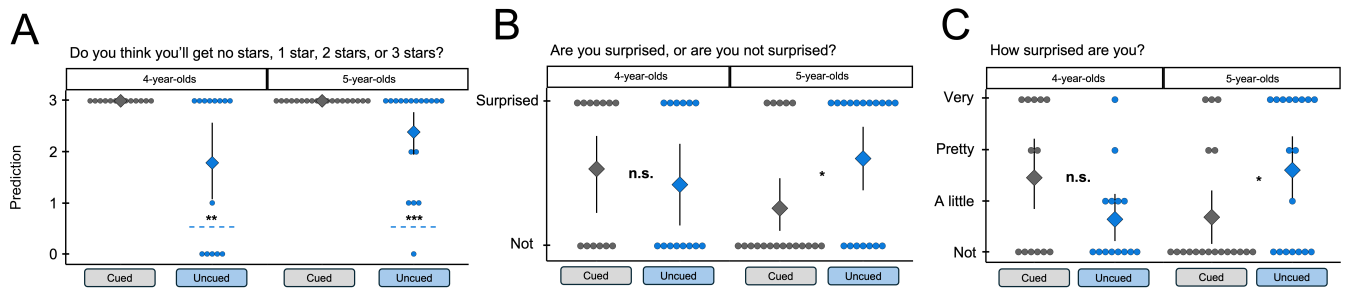

Figure 3: Results for Exp. 2 (memory game). Error bars represent bootstrapped 95% CIs. (A) Prediction; dashed line indicates children's true performance (M = .56 stars). (B) Binary surprise. (C) Graded surprise. $^{***}p \leq .001$; $^{**}p \leq .01$; $^{*}p \leq .05$.

stantially exceeded their true average performance (M = .56 stars, SD = .62): $t(31) = 7.14$, $p < .001$ (one sample t-test, preregistered). All children in the Cued condition appropriately predicted 3 star cards. See Figure 3.

We then asked whether children reported more surprise when they had to rely on their memory versus a visual cue guaranteeing success. Given the absence of any age effect in Exp.1, our main preregistered analysis was to use logistic regression with condition as the main predictor. However, condition alone was not a significant predictor of children's surprise, either in the binary measure (Uncued vs. Cued; 37.5% vs. 53%; β = .64, p = .211, binomial logistic regression), or level of surprise (M = 1.19 (SD = 1.31) vs M = 1.00 (SD = 1.34); β = .41, p = .402, ordinal logistic regression).

We then turned to our preregistered exploratory analysis to examine the effect of age. Including age (in years) with an interaction term[4] yielded an interaction between condition and age (marginal in the binary measure, β = 1.92, p = .068; significant in the level of surprise, $\beta = 2.47$, $p = .015$). Follow-up analysis found the predicted effect of condition in 5-year-olds (n = 37), who reported more surprise in the Uncued vs. Cued (binary: β = 1.48 p = .037, level of surprise: β = 1.44, p = .035), whereas 4-year-olds (n = 27) did not (ps > .165).

Similar to Exp.1, this difference was not driven exclusively by the children in the Uncued condition who made more realistic predictions and were (reasonably) surprised by the outcome. Given the smaller sample size, the effect of condition among 5-year-olds was weaker but remained marginal even after excluding those who predicted fewer than 3 stars in the Uncued condition (n = 6), in both binary surprise (β = 1.37, p = .081) and level of surprise: β = 1.31, p = .081).

To complement these findings,we examined whether the Manhattan Distance measure predicted self-reported surprise. If children reported more surprise as a function of the distance between their chosen locations and the true locations (representing greater implausibility that the chosen card is actually a star card), we should see the latter predicting the former. A pair of regression analyses showed no effect of distance on surprise (ps > .128 for binary surprise and level of surprise), suggesting that children who made wildly inaccurate guesses were not necessarily more surprised than those who made effortful guesses.

In sum, the results of our first experiment replicated among the 5-year-old children in our sample, confirming that while children showed the widely documented overoptimism in their predictions, their surprise uncovered well-calibrated expectations over their performance outcomes. When success expectancy was low, the older (but not younger) children were significantly more likely to report surprise upon succeeding than when success expectancy was high. Children's surprise responses were not driven purely by violation-of-expectation over specific card locations: The present game allowed us to precisely track which locations children expected to hold target cards (card choices measured by click inputs), and the distance between chosen location and true location was unrelated to their later surprise.

Put together, these findings suggest that children's sensitivity to expectation-outcome discrepancies applies not only to a probability-based task but also to a skill-based task. By around five years of age, children's self-reported surprise is sensitive to success expectancy across both domains, and is largely independent of superficial features of the task.

General Discussion

Do children hold reasonable expectations about their performance outcomes? Although children tend to *say* they will achieve unrealistic performance outcomes, we find that at least by 5 years of age, their *surprise* suggests otherwise. In a probability-based task (Exp.1) and a skill-based memory task (Exp.2), children consistently predicted to achieve a desirable outcome, but after having achieved that outcome, they were more likely to express surprise when the outcome was unlikely than when it was guaranteed.

Drawing on prior work on overoptimism (Leonard & Sommerville, 2025; Xia et al., 2024) and uncertainty monitoring (Goupil et al., 2016; Hembacher & Ghetti, 2014; Lyons & Ghetti, 2011), these results demonstrate a striking discrepancy between children's explicit (and wrong) predictions and their internal (and more reasonable) expectations within the same individual participants. Going beyond prior findings suggesting that children are aware of an internal signal that indicates

[4]While we did not specify how age would be included in the preregistration, we find similar results with a continuous age variable.

subjective uncertainty during a task, the current work raises the possibility that children are also capable of generating reasonable expectations about their future performance. Consistent with the idea that surprise reflects prediction error—a discrepancy between expected vs. observed outcomes—these findings also extend prior work on children's ability to predict the outcomes in the external world; just as observing unexpected physical and social outcomes elicits surprise, children in our study reported being surprised when the outcomes they themselves generated violated their expectations.

Recent work on metacognitive development has demonstrated similar discrepancies in different ways. For instance, 4- and 5-year-olds claim that they know what color 'chartreuse' is, but are more likely than older children to adopt a peer's answer when asked for the chartreuse crayon (Hagá & Olson, 2017). However, this finding comes from a comparison between age groups, suggesting older children—perhaps rather surprisingly—held onto their choices even when they were uncertain. Thus, it leaves open questions about the underlying reason for the selective divergence in younger children. Another study with 4- and 5-year-olds showed that children claim overconfidence in a perceptual decision task (e.g., reporting they know the shape of an object behind partially or fully occluded windows), but given a choice to look behind the window, the vast majority chose either the fully occluded (64%) or partially occluded (24%) window over the clear window. This dissociation, however, was shown across two separate experiments rather than within participants. Furthermore, in both studies, children were likely incentivized by their desire to find the correct answer (Hagá & Olson, 2017), which may have been driven by their uncertainty.

The present work adds to these findings by demonstrating a striking discrepancy between an explicit measure of prediction and self-reported surprise as an indirect measure of prediction. By tapping into children's beliefs about the outcomes their actions can produce in the world, we were able to measure this discrepancy as prediction about, and reaction to, the outcome within the same participants. Yet, one limitation of the current work is its reliance on self-reported surprise. Although prior work has demonstrated converging evidence between self-reported surprise and real-time changes in facial expressions at the time of receiving surprising information (e.g., Chuey et al., 2026), the current work cannot rule out the possibility that this measure reflects the uncertainty children accumulated during the challenging task, rather than their a priori expectations.

From this perspective, leveraging additional measures to find converging evidence for prediction error—such as pupillometry (e.g., Colantonio et al., 2023), gaze data using head-mounted cameras (Long et al., 2024), facial coding (Chuey et al., 2026; Ni et al., 2023), as well as physiological (e.g., skin conductance) or neural measures (e.g., OPM-MEG)—is an exciting direction for future work. Broadening the range and nature of measures will enable the study of pre-verbal populations, assessment of convergent validity of measures, and checking for biases in children's verbal self-reports (e.g., a discrepancy between implicit and explicit surprise).

Given that we did not find an effect of age in Exp. 1, the difference between 4- and 5-year-olds in Exp. 2 was rather unexpected. While we remain cautious about interpreting this finding, it is possible that younger children found it more challenging to estimate the difficulty of the memory task (Exp. 2) than the probability-based task (Exp. 1), or struggled to represent their memory skills; both are crucial for generating well-calibrated expectations about their performance outcomes.

What might explain the discrepancy between children's explicit prediction and surprise? While the current study does not provide an answer, we can entertain at least two possibilities. First, this might reflect the well-documented discrepancy between 'implicit' measures (e.g., looking time) and more explicit, verbal responses. Examples of divergence between these types of measures have been found in the domain of physical reasoning (Hood, 1995; Hood, Carey, & Prasada, 2000; Hood, Santos, & Fieselman, 2000), probabilistic reasoning (Doan et al., 2023; Xu & Garcia, 2008) and Theory of Mind (Onishi et al., 2007; Rakoczy, 2022, but see Powell et al., 2018; Schuwerk et al., 2021). Another possibility is that children's overoptimistic predictions reflect their desire to appear confident (e.g., due to its perceived social benefits; see Birch et al., 2020) or their tendency to verbalize the outcomes they *want* to see (Bernard et al., 2016; Wente et al., 2020) rather than their true expectations or beliefs. These explanations are neither mutually exclusive nor exhaustive, and further investigating the mechanism behind this discrepancy remains an important area for future work.

Our findings also open up new questions. For instance, what are the inductive constraints on children's expectations over their abilities in various domains? Quantitative modeling can be useful for explicitly formalizing the role of children's priors over their competence, and their understanding of the task difficulty, as well as how children might integrate their experience with the task to update these representations. In light of findings on children's tendency to explore after observing surprising outcomes (Doan et al., 2020; Stahl & Feigenson, 2015), another exciting question is whether self-PE would lead to an increase in exploration, specifically to test hypotheses about their competence. It is also interesting to think about how children's emotional and behavioral responses might differ in cases of unexpected failure (negative prediction error) as compared to unexpected success (positive prediction error). While negative prediction error might still promote curiosity (why did I fail?) and even humility (how can I get better?), it might lead to different behavioral consequences, such as refusing to look at the outcome or reduced exploration.

The current work is only a first step in understanding how young learners respond to and learn from the outcomes of their own actions. By synthesizing research on prediction error, metacognition, and active learning, further work in this domain can help shed light on what makes humans remarkable learners not only in learning about the external world but also in learning about their own abilities.

Acknowledgments

We thank Claudia Lewis, Hannah Kang, and Emilia de Jesus for help with stimuli design and data collection, the Social Learning Lab for helpful feedback, and families and children for participating. This work was funded by the Stanford Accelerator for Learning and the James S. McDonnell Foundation.

References

Bernard, S., Clément, F., & Mercier, H. (2016). Wishful thinking in preschoolers. *Journal of Experimental Child Psychology*, *141*, 267–274. https://doi.org/10.1016/j.jecp.2015.07.018

Birch, S. A. J., Severson, R. L., & Baimel, A. (2020). Children's understanding of when a person's confidence and hesitancy is a cue to their credibility. *PLoS ONE*, *15*(1), e0227026. https://doi.org/10.1371/journal.pone.0227026

Blakemore, S.-j., Rees, G., & Frith, C. D. (1998). How do we predict the consequences of our actions? a functional imaging study. *Neuropsychologia*, *36*(6), 521–529. https://doi.org/10.1016/S0028-3932(97)00145-0

Camras, L. A., Meng, Z., Ujiie, T., Dharamsi, S., Miyake, K., Oster, H., Wang, L., Cruz, J., Murdoch, A., & Campos, J. (2002). Observing emotion in infants: Facial expression, body behavior, and rater judgments of responses to an expectancy-violating event. *Emotion*, *2*(2), 179–193. https://doi.org/10.1037/1528-3542.2.2.179

Chuey, A., Jara-Ettinger, J., & Gweon, H. (2026). Young children understand how social connections affect what people know about each other. *Proceedings of the National Academy of Sciences*, *123*(12), e2525150123.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Colantonio, J., Bascandziev, I., Theobald, M., Brod, G., & Bonawitz, E. (2023). Seeing the Error in My "Bayes": A Quantified Degree of Belief Change Correlates with Children's Pupillary Surprise Responses Following Explicit Predictions. *Entropy*, *25*(2), 211. https://doi.org/10.3390/e25020211

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*(3), 347–374. https://doi.org/10.1007/s11409-014-9133-z

Doan, T., Castro, A., Bonawitz, E., & Denison, S. (2020). "Wow, I did it!": Unexpected success increases preschoolers' exploratory play on a later task. *Cognitive Development*, *55*, 100925. https://doi.org/10.1016/j.cogdev.2020.100925

Doan, T., Friedman, O., & Denison, S. (2023). Calculated Feelings: How Children Use Probability to Infer Emotions. *Open Mind*, *7*, 879–893. https://doi.org/10.1162/opmi_a_00111

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive Psychology*, *1*(4), 324–340. https://doi.org/10.1016/0010-0285(70)90019-8

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292. https://doi.org/10.1016/S1364-6613(03)00128-1

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences*, *113*(13), 3492–3496.

Gredebäck, G., & Melinder, A. (2010). Infants' understanding of everyday social interactions: A dual process account. *Cognition*, *114*(2), 197–206. https://doi.org/10.1016/j.cognition.2009.09.004

Hagá, S., & Olson, K. R. (2017). Knowing-it-all but still learning: Perceptions of one's own knowledge and belief revision. *Developmental Psychology*, *53*(12), 2319.

Hembacher, E., & Ghetti, S. (2014). Don't Look at My Answer: Subjective Uncertainty Underlies Preschoolers' Exclusion of Their Least Accurate Memories. *Psychological Science*, *25*(9), 1768–1776. https://doi.org/10.1177/0956797614542273

Hennefield, L., & Markson, L. (2022). The development of optimistic expectations in young children. *Cognitive Development*, *63*, 101201. https://doi.org/10.1016/j.cogdev.2022.101201

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, *35*(2), 209–223. https://doi.org/10.1111/mila.12281

Hood, B. (1995). Gravity rules for 2- to 4-year olds? *Cognitive Development*, *10*(4), 577–598. https://doi.org/10.1016/0885-2014(95)90027-6

Hood, B., Carey, S., & Prasada, S. (2000). Predicting the Outcomes of Physical Events: Two-Year-Olds Fail to Reveal Knowledge of Solidity and Support [https://srcd.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8624.00247]. *Child Development*, *71*(6), 1540–1554. https://doi.org/10.1111/1467-8624.00247

Hood, B., Santos, L., & Fieselman, S. (2000). Two-year-olds' naîve predictions for horizontal trajectories [https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-7687.00127]. *Developmental Science*, *3*(3), 328–332. https://doi.org/10.1111/1467-7687.00127

Knolle, F., Schröger, E., & Kotz, S. A. (2013). Prediction errors in self- and externally-generated deviants. *Biological Psychology*, *92*(2), 410–416. https://doi.org/10.1016/j.biopsycho.2012.11.017

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of Dispositional States by 12-Month-Olds. *Psychological Science*, *14*(5), 402–408. https://doi.org/10.1111/1467-9280.01454

Lapidow, E., Killeen, I., & Walker, C. M. (2022). Learning to recognize uncertainty vs. recognizing uncertainty to learn: Confidence judgments and exploration decisions in preschoolers. *Developmental science*, *25*(2), e13178.

Leonard, J. A., & Sommerville, J. A. (2025). A unified account of why optimism declines in childhood. *Nature Reviews Psychology*, *4*(1), 35–48. https://doi.org/10.1038/s44159-024-00384-z

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology*, *103*(2), 152–166. https://doi.org/10.1016/j.jecp.2008.10.002

Lipko-Speed, A. R. (2013). Can young children be more accurate predictors of their recall performance? *Journal of Experimental Child Psychology*, *114*(2), 357–363. https://doi.org/10.1016/j.jecp.2012.09.012

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42. https://doi.org/10.1016/j.cognition.2016.12.007

Long, B., Xiang, V., Stojanov, S., Sparks, R. Z., Yin, Z., Keene, G. E., Tan, A. W. M., Feng, S. Y., Zhuang, C., Marchman, V. A., Yamins, D. L. K., & Frank, M. C. (2024). The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences.

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child development*, *82*(6), 1778–1787.

Lyons, K. E., & Ghetti, S. (2013). I don't want to pick! introspection on uncertainty supports early strategic behavior. *Child development*, *84*(2), 726–736.

Moll, H., Kane, S., & McGowan, L. (2016). Three-year-olds express suspense when an agent approaches a scene with a false belief [https://onlinelibrary.wiley.com/doi/pdf/10.1111/desc.12310]. *Developmental Science*, *19*(2), 208–220. https://doi.org/10.1111/desc.12310

Ni, Q., Shoyer, J., Bautista, Z., Raport, A., & Moll, H. (2023). Toddlers' expressions indicate that they track agent–object interactions but do not detect false object representations. *Journal of Experimental Child Psychology*, *231*, 105639. https://doi.org/10.1016/j.jecp.2023.105639

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005

Onishi, K. H., Baillargeon, R., & Leslie, A. M. (2007). 15-month-old infants detect violations in pretend scenarios. *Acta Psychologica*, *124*(1), 106–128. https://doi.org/10.1016/j.actpsy.2006.09.009

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*, 40–50.

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, *1*(4), 223–235. https://doi.org/10.1038/s44159-022-00037-z

Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking [https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-7687.00044]. *Developmental Science*, *1*(2), 291–297. https://doi.org/10.1111/1467-7687.00044

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schuwerk, T., Kampis, D., Alessandroni, N., Altvater-Mackensen, N., Arias-Trejo, N., Axelsson, E., Baillargeon, R., Frank, M. C., Rakoczy, H., et al. (2021, February). Action anticipation based on an agent's epistemic state in toddlers and adults. https://doi.org/10.31234/osf.io/x4jbm

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A New Online Platform for Developmental Research. *Open Mind*, *1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological review*, *99*(4), 605.

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

Wente, A. O., Goddu, M. K., Garcia, T., Posner, E., Fernández Flecha, M., & Gopnik, A. (2020). Young Children Are Wishful Thinkers: The Development of Wishful Thinking in 3- to 10-Year-Old Children [https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.13299]. *Child Development*, *91*(4), 1166–1182. https://doi.org/10.1111/cdev.13299

Wolpert, D. M., Miall, R., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Sciences*, *2*(9), 338–347. https://doi.org/10.1016/S1364-6613(98)01221-2

Xia, M., Poorthuis, A. M. G., & Thomaes, S. (2023). Why do young children overestimate their task performance? A cross-cultural experiment. *Journal of Experimental Child Psychology*, *226*, 105551. https://doi.org/10.1016/j.jecp.2022.105551

Xia, M., Poorthuis, A. M. G., & Thomaes, S. (2024). Children's overestimation of performance across age, task, and historical time: A meta-analysis [https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.14042]. *Child Development*, *95*(3), 1001–1022. https://doi.org/10.1111/cdev.14042

Xia, M., Poorthuis, A. M. G., Zhou, Q., & Thomaes, S. (2022). Young children's overestimation of performance: A cross-cultural comparison. *Child Development*, *93*(2), e207–e221. https://doi.org/10.1111/cdev.13709

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sci-*

ences, *105*(13), 5012–5015. https://doi.org/10.1073/pnas.0704450105

Yussen, S. R., & Levy, V. M. (1975). Developmental changes in predicting one's own span of short-term memory. *Journal of Experimental Child Psychology*, *19*(3), 502–508. https://doi.org/10.1016/0022-0965(75)90079-X