Exposure to detectable inaccuracies makes kids more diligent fact-checkers of novel claims

Evan Orticio[1], Martin Meyer[1,2,3], and Celeste Kidd[1]

[1] Department of Psychology, University of California, Berkeley

[2] Department of Psychology, Yale University

[3] Department of Philosophy, Yale University

Corresponding Author: Evan Orticio, email: eorticio@gmail.com

Abstract

How do children decide when to believe a claim? We show that children fact-check claims more and are better able to catch misinformation when they've been exposed to detectable inaccuracies. In two experiments, 4-7-year-old children exposed to falsity (as opposed to all true information) sampled more evidence before verifying a test claim in a novel domain. Children's evidentiary standards were graded: fact-checking increased linearly with the proportion of false statements heard during exposure. A simulation suggests that children's behavior is adaptive, because increased fact-checking in more dubious environments supports the discovery of potential misinformation. Importantly, children were least diligent at fact-checking a new claim when all prior information was true, suggesting that sanitizing kids' informational environments may inadvertently dampen their natural skepticism. Instead, these findings support the counterintuitive possibility that exposing children to some nonsense may scaffold vigilance toward more subtle misinformation in the future.

Children have unprecedented access to information on their phones and computers. This fact represents a very recent shift—one of both promise and problem. The internet leaves users exposed to unprecedented amounts of misinformation. Exposure to misinformation can lead to the long-term adoption of false beliefs in both adults (Brown & Nix, 1996) and children (Fazio & Sherry, 2020). This is true even when the learner is aware of this bias (Fazio et al., 2015). Misinformation exposure is also expected to increase with the widespread adoption of generative AI models like ChatGPT and Bard. When these models produce fabricated information in their outputs, they can transmit them to users (Kidd & Birhane, 2023). Kids are likely most vulnerable because they have less world knowledge (Xu, Shtulman & Young, 2022) and are biased to trust information (Jaswal et al., 2010), particularly under conditions of uncertainty (Plate et al., 2021). Indeed, even when preschoolers directly observe data that conflicts with testimony, they rarely seek additional data and struggle to disregard the misleading information (Hermansen, Ronfard, Harris & Zambrana, 2021). Despite these unique vulnerabilities, the overwhelming majority of work on misinformation centers adults (e.g., Ecker et al., 2022).

What we know of children's media habits suggest that they are immersed. A third of American children are on at least one social media platform by age 9 (Mott Poll, 2021). A majority of American teens get their news from social media or YouTube (Common Sense Media, 2019). And children who have used ChatGPT for schoolwork report using it in place of traditional search (Common Sense Media, 2023). Thus children's media diets are rife with dubious sources. How do we best prepare children to navigate this complex informational sea?

The preeminent solution has been to shield children from misinformation via sanitized platforms. YouTube Kids, for example, curates a small selection of child-focused content through a combination of automated filters and human review (Rodriguez, 2018). This solution is limited by its reactive nature. As an example, YouTube Kids received widespread criticism when a Guardian article reported on a multitude of videos featuring themes that were not appropriate for children (e.g. violent and sexual

situations) were inaccurately labeled as child-friendly by the platform's filters, likely because they contained characters from children's movies and shows ("Elsagate", BBC Trending, 2017; Maheshwari, 2017). Efforts to sanitize content for children are resource-intensive and subjective (who decides what is age-appropriate?). Moreover, automated curation approaches are easily gamed, and human curation approaches cannot scale as rapidly as new content is produced (Kallioniemi, 2021).

Another proposed strategy for safeguarding adults from misinformation comes from inoculation theory (for reviews, see Lewandowsky & van der Linden, 2021; Compton et al., 2021; van der Linden, 2022). Inspired by an analogy to biomedical inoculation, the theory postulates that preemptively exposing learners to a weakened form of a misleading argument can confer immunity to its persuasiveness later on. This process involves 'prebunking' the argument by refuting false information in advance, and/or deconstructing misleading argumentation techniques more broadly. Researchers claim to have successfully inoculated adults against misinformation spanning many topics, including climate change, vaccination, and extremist ideology (van der Linden et al., 2017; Wong, 2016; Braddock, 2022; Roozenbeek, van der Linden, & Nygren, 2020; Basol, Roozenbeek & van der Linden, 2020). However, inoculation interventions are fragile, ephemeral, and difficult to scale (Maertens et al., 2021). Inoculation interventions are completely ineffective after 48 hours if participants don't receive an immediate post-test (Capewell et al., 2023). These interventions have also been criticized for fatal methodological weaknesses in the assessment of their efficacy (Guay et al., 2023; Williams, 2023; Chan & Albarracín, 2023). For example, a recent analysis found no evidence that inoculation improves discrimination, but rather that it induces a potentially counterproductive, negative response bias (Modirrousta-Galian & Higham, 2023).

Further, misinformation inoculation techniques are untested in young children—and there's reason to expect they may not achieve even the modest, ephemeral effects seen in adults because of differences in children's decision-making and metacognition. Children have less developed

metacognitive skills than adults, a tendency toward overconfidence (Finn & Metcalfe, 2014; Lipko et al., 2009; Salles et al., 2016), and less executive function (Best & Miller, 2010). Thus, it may be more difficult to find interventions that effectively lower overconfidence and slow decision-making about factual accuracy for kids.

Here, we propose an approach that builds upon people's known capacities to use the statistical properties of information in their environment—capacities for which we have strong evidence even in infancy (Saffran, Aslin, & Newport, 1996; Fiser & Aslin, 2002; Xu & Garcia, 2008). The idea leverages the fact that children attend to statistical properties of their environments in order to form expectations which then modulate their learning and behavior (Saffran, Aslin, & Newport, 1996; Kidd, Piantadosi & Aslin, 2012). Existing empirical work shows that children wait longer in a delay-of-gratification task when given evidence that waiting will pay off (Watts, Duncan & Quan, 2018; Kidd, Palmeri & Aslin, 2013). We hypothesize that, in a similar fashion, children will use the prior reliability of information in a given context to adjust their a priori skepticism toward new claims. In two experiments and a simulation, we test whether controlled but imperfect informational environments may serve as useful scaffolding for children's abilities to detect misinformation. Exposure to detectable inaccuracies may provide critical opportunities for children to express their skepticism and practice key critical thinking skills.

Experiments

Experiments were approved by the Institutional Review Board at the University of California, Berkeley (protocol no. 2018-12-11653). Informed consent was obtained by a legal guardian of all participants before participation. All children provided verbal assent, and 7-year-old children additionally signed an assent form.

Study 1

Study 1 asks whether children use the prior reliability of information in their environment to shape their standards of evidence for a novel claim. Do children increase their evidentiary standards for a claim selectively after exposure to misinformation? To test this, children were randomly assigned to judge the veracity of a set of animal facts that were either all true (Reliable condition) or partially false (Unreliable condition). Following this, children judged a novel claim about aliens, and were given the opportunity to freely sample evidence about the claim before making their final decision. We hypothesized that children would sample more evidence before trusting the claim in the Unreliable condition.

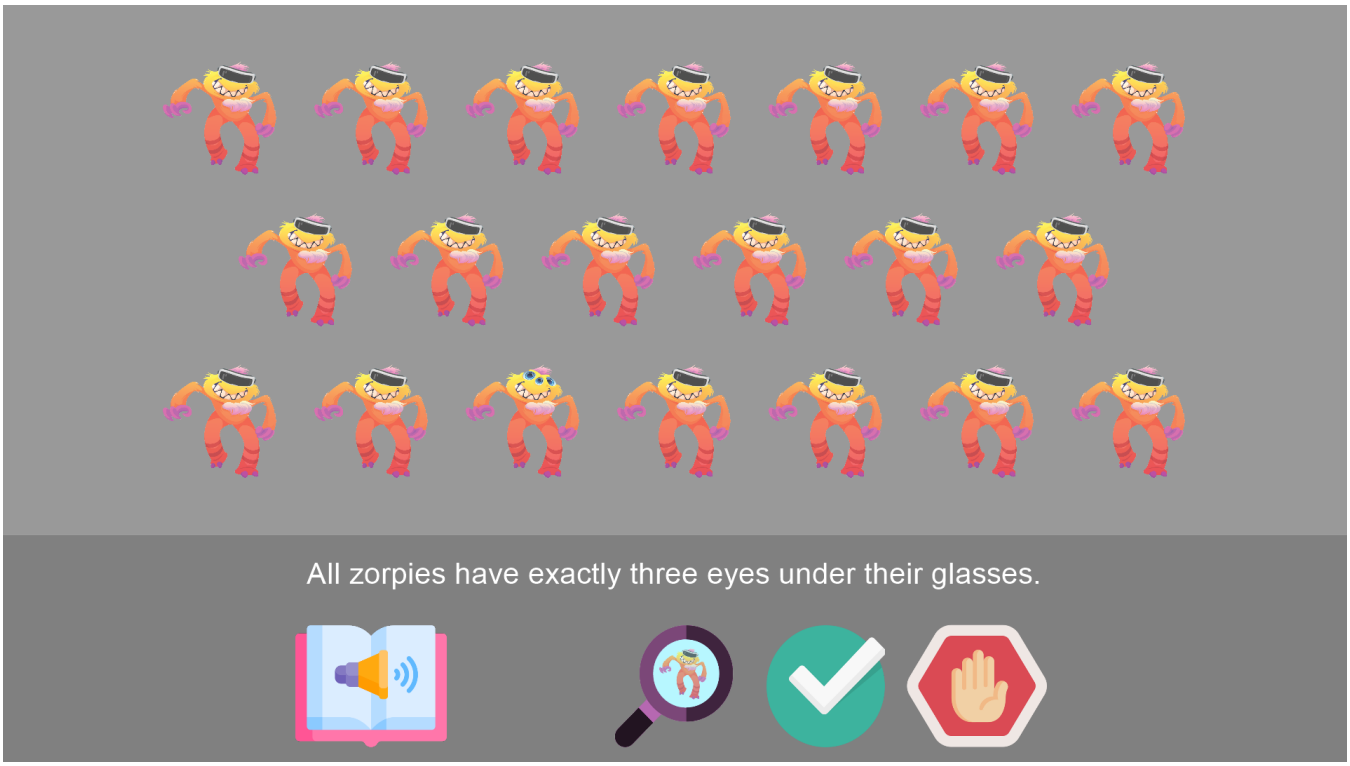

Figure 1. Test phase, identical in Studies 1 and 2. After checking a zorpie (e.g., bottom row, third from left), children could choose to accept the statement (green button), reject the statement (red button), or check another zorpie before deciding (magnifying glass).

Method

Participants Sixty 4- to 6-year-old children (M_{age} = 5.51, SD = 0.89, 47% White) were recruited from parks in the California Bay Area. Three additional children were excluded from analysis because they had watched another child participate or were too distracted to complete the study.

Procedure Children used a touchscreen computer to play a game created in PsychoPy. In an exposure phase, the experimenter asked children to determine whether a set of statements about animals in an e-book were right or wrong. On each of 12 exposure trials, the tablet displayed a statement (e.g., "Zebras have black and white stripes") and an accompanying picture. Children first tapped a button to hear an audio recording of the statement, and then indicated whether the fact was right (by tapping a green button) or wrong (by tapping a red button). The facts varied by condition (between-subjects, pseudo-randomly assigned, N = 30 per condition). In the Reliable condition, all 12 animal statements were true. In the Unreliable condition, four of the 12 animal statements were clearly false (e.g., "Zebras have red and green stripes"). Pictures were identical across conditions, so children could judge the statements using real-world knowledge or the pictures alone. The first two trials were considered practice trials, and children were given feedback if they were wrong. No feedback was provided on the remaining 10 trials.

In the subsequent test phase (see Figure 1), children moved on to a second chapter of the e-book which was about a novel alien species called "zorpies". They were asked to evaluate a new statement about zorpies: "All zorpies have exactly three eyes under their glasses." The screen displayed the fact alongside 20 zorpies wearing opaque sunglasses. After tapping a button to hear the fact, children were told that they could tap any zorpie to remove its glasses and reveal its eyes. All zorpies were identical and had three eyes, so any evidence the child sampled supported the test statement. Once the child tapped a zorpie, they had to decide to tap a button to accept the statement as true, reject the statement as false, or to check another zorpie first. This procedure repeated such that children could check as many

zorpies as they wished (from 1 to all 20) before indicating whether the fact was right or wrong and completing the study. The task was designed to produce different information seeking behavior depending on one's level of skepticism toward the claim. A fully trusting learner might see that all the zorpies are identical and be satisfied after checking only one, while a highly skeptical learner might feel the need to check all 20 zorpies because the statement refers to "*all* zorpies".

Results and Discussion

Manipulation check Children reliably discerned between true and false statements in the exposure phase of the experiment. Children's accuracy in evaluating statements as true or false were significantly above chance in both the reliable (M = 9.40 of 10 correct, 95% CI = (8.99, 9.81), $t(29) = 21.88$, two-sided $p < .001$) and unreliable conditions (M = 8.43 of 10 correct, 95% CI = (7.62, 9.25), $t(29) = 8.64$, two-sided $p < .001$), indicating that we successfully manipulated the perceived reliability of information in the exposure phase. Nine of the 60 participants failed to achieve 80% accuracy, but their exclusion does not affect any results, so we retain their data for all future analyses. Additionally, all but three children (95%) correctly judged the test claim to be true, suggesting that children were generally tracking the evidence appropriately. Of the three participants who rejected the test claim, two were in the unreliable condition.

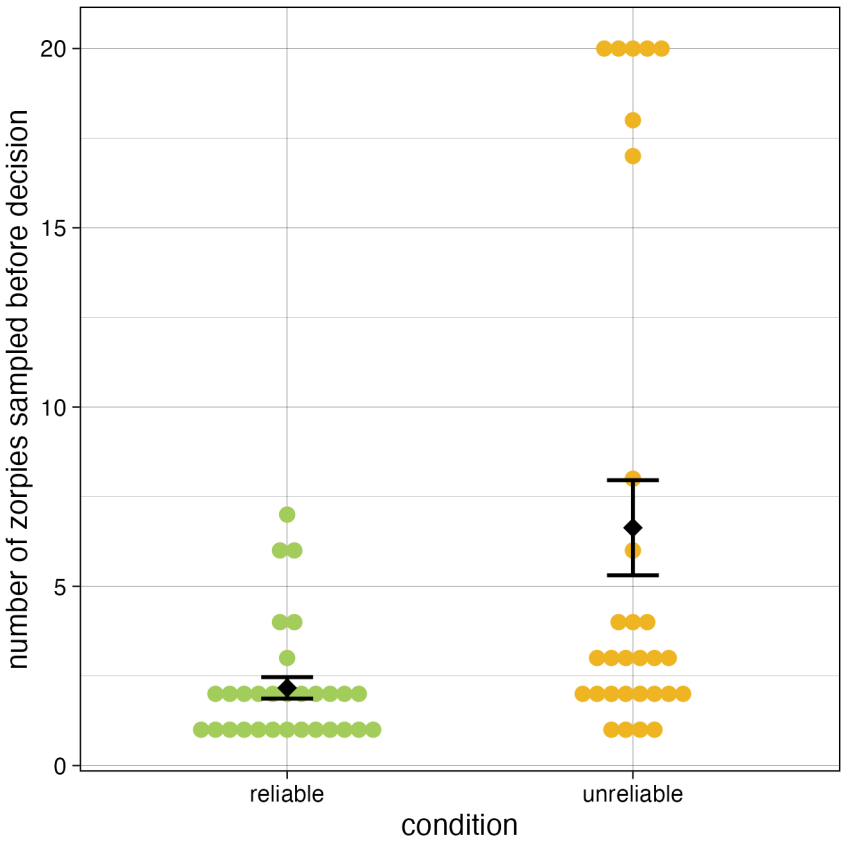

Figure 2. Children sampled more evidence in the unreliable condition. Dots are individual data points, diamonds are condition means, and error bars represent one SEM. The effect of condition remains robust after winsorization, ensuring that the highest sample values do not drive the effect.

Children seek more evidence in unreliable environments Children increase their standards of evidence for new claims in an environment containing some misinformation. Figure 2 shows the number of zorpies children sampled before deciding to accept or reject the test claim by condition. We used the *rstatix* package in R to run a Wilcoxon paired signed rank test assessing the effect of condition on the amount of evidence sampled. On average, children in the Unreliable condition sampled more evidence than those in the Reliable condition (M = 6.63 vs. 2.17 zorpies, location parameter = -1, 95% CI = (-2, -1), two-sided Wilcoxon W = 233, $p < .001$, r = 0.43). When exposed to some misinformation, children sought out significantly more evidence before deciding whether to accept the test claim. In the

Unreliable condition, a number of children even opted for an exhaustive or near-exhaustive sampling strategy, checking up to 20 zorpies in a row even though all the prior evidence was identical. Children were thus able to leverage the prior quality of information in a known domain (animal facts) in order to adapt their skepticism and subsequent information search about a novel claim about which they had no prior knowledge.

The distribution of sampling behavior in the unreliable condition was bimodal, so we also winsorized the data such that the maximum value of zorpies sampled was 8 (the maximum of the other mode). The fact that an exhaustive sampling strategy leads children to sample exactly 20 zorpies is the result of a design choice, so replacing extreme values with the maximum value of the other mode provides a more stringent and design-neutral test of our hypothesis. The effect remained robust after winsorization (location parameter = -1, 95% CI = (-2, -1), two-sided Wilcoxon $W = 233$, $p < .001$, r = 0.43), suggesting that it was not driven by the subset of exhaustive samplers in the unreliable condition.

Finally, we tested whether children's sensitivity to the reliability of their informational environments changes with age. We used the *MASS* package in R to run a robust linear regression using standardized age and condition to predict the number of zorpies sampled. This analysis replicated the main effect of condition ($\beta = 1.41$, 95% CI = (0.43, 2.39), $t(56) = 2.81$, $p = .007$), but revealed no main effect of age and no interaction (p's > .05). There were no changes in how children responded to the reliability of information from ages 4 through 6 in our sample.

Study 2

The reliability of a body of information is not all-or-nothing. Do children appreciate nuances in the reliability of their broader informational environments, and adapt their level of skepticism accordingly? To address this question, Study 2 introduced five between-subjects conditions of varying reliability, ranging from 0% to 80% false statements in the exposure phase. Additionally, children likely assumed that the information in Study 1 came from a single, cohesive source. The task was framed as an

ebook, and children heard all statements in audio recordings using the same voice. Can children still make smart inferences about claims derived from a more complex environment composed of many distinct sources? In Study 2, we presented statements as individual search results, read in distinct voices, to test whether children make more abstract generalizations about their informational environment to adapt their information seeking.

Method

Participants Sixty-two 4- to 7-year-old children (M_{age} = 5.88, SD = 1.06, 52% White) were recruited from parks in the California Bay Area. Four additional participants were excluded from analysis because they had watched another child participate. None of the Study 2 participants had completed Study 1 previously.

Procedure The procedure was identical to Study 1 aside from two main changes. First, we created five between-subjects conditions such that 0%, 20%, 40%, 60%, or 80% of the 10 exposure trials were false statements. Experimental conditions were pseudo-randomly assigned and had sample sizes of N = 12, 13, 13, 12, and 12, respectively). Second, the activity was reframed so that the statements appeared to originate from distinct search engine results. The experimenter typed "Animal facts" into a search bar to generate a simulated results page in the exposure phase. The experimenter tapped on a search result to begin a trial. On each trial, the style of the picture and the voice of the audio recording were different. The test phase followed. In the test phase, the experimenter input "Alien facts" into the search bar and tapped a result to display the page of zorpies. The audio recording on the zorpie test trial featured another distinct voice.

Results and Discussion

Manipulation check Children reliably discerned between true and false statements in the exposure phase of the experiment in Study 2. Accuracy in the exposure phase was significantly above chance (*M*

= 9.24 of 10 correct, SD = 1.00, two-sided $p < .001$). Four of the 62 participants failed to achieve 80% accuracy, but their exclusion does not affect any results, so we retain their data for all future analyses. Additionally, all but four children (93.5%) correctly judged the test claim to be true. The children who rejected the test claim were in the two most unreliable conditions (three in the 80% false condition, one in the 60% false condition).

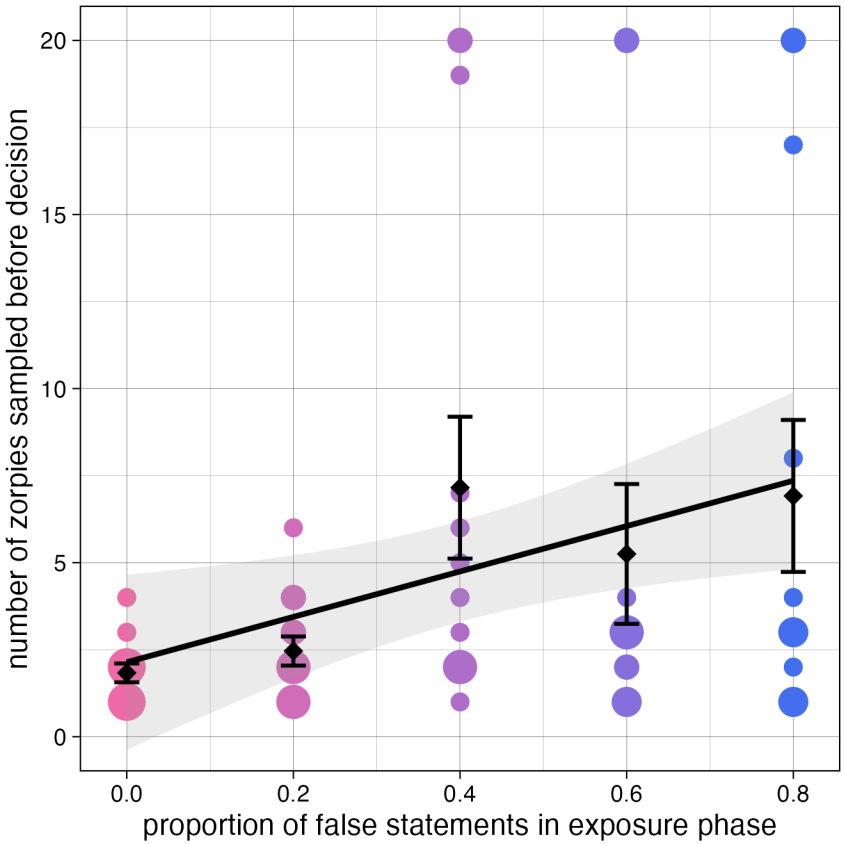

Figure 3. Children sampled more evidence as the reliability of their environments decreased. Amount of evidence sampled (out of a possible 20 zorpies) vs. proportion of false statements in exposure phase in Experiment 2. The size of the dot represents the number of data points. Diamonds are conditional means and error bars represent 1 SEM. Line is the linear regression fit with a 95% confidence interval.

Graded sensitivity to reliability Figure 3 shows the number of zorpies children sampled before deciding to accept or reject the test claim as a function of the proportion of false statements encountered

in the exposure phase. A linear regression revealed that the proportion of false statements in the exposure phase was a significant linear predictor of the number of zorpies sampled ($\beta = 6.52$, 95% CI = (1.34, 11.70), $t(60) = 2.52$, $p = .014$). Skepticism increased linearly with increases in the number of false statements in the exposure phase, manifesting in more extensive information search in the test phase. Children are thus able to make sophisticated, graded judgments about the reliability of their current informational environments, and use that to guide future learning. Note that this sensitivity was observed in a simulated search engine context composed of distinct sources–each statement was heard from a different voice. This suggests that children went beyond speaker-based heuristics and tracked the cumulative quality of information throughout the exposure phase.

Because the data was bimodal, we replicated this result with a robust linear regression, which is less sensitive to influential observations. The proportion of false statements remained a significant linear predictor ($\beta = 1.95$, 95% CI = (0.05, 3.84), $t(60) = 2.01$, $p = .049$). These findings suggest that, at a group level, children linearly scaled up their evidentiary standards as the reliability of their informational environments decreased.

While we observed a linear relationship between reliability and information sampling in the present data (as predicted), it is unclear whether this would generalize across all tasks. In our task, the information to be gained by additional sampling was maximally transparent. The outcome was binary and directly related to the claim in question (the next zorpie is three-eyed or not), and the full space of available evidence was clearly delineated. Other studies with low-risk exploration have also found linear associations between low certainty and information seeking in children (Coughlin et al., 2014) and adults (Desender et al., 2018).

On the other hand, some evidence suggests environments characterized by variable expected information gain induce a U-shaped relationship between curiosity and information seeking (Baranes, Oudeyer, & Gottlieb, 2014; Wang et al., 2021). This pattern of results is consistent with a dual-process

account of metacognition, in which information seeking is guided not only by certainty, but also by an appraisal of the potential information gain afforded by the environment (Goupil & Proust, 2023; Baer & Kidd, 2022). Speculatively, the bimodal distribution of sampling strategies even in the most unreliable conditions of Study 2 may represent two distinct interpretations of the environment. Some of the children who checked only a few zorpies in highly unreliable environments may have been highly skeptical, but doubted that the available evidence would provide accurate information in the first place. The linear effect we observe may therefore be the combination of two patterns of responses: an even stronger linear effect dampened by a group exhibiting a U-shaped pattern.

Sensitivity moderately increases with age Does children's skepticism become more finely attuned to the reliability of their informational environment as they age? We ran a robust linear regression using standardized age and the standardized proportion of false statements in exposure to predict the number of zorpies sampled. First, we replicated the main effect of environmental reliability (i.e., proportion of false statements in exposure phase, $\beta = 1.29$, 95% CI = (0.74, 1.85), $t(58) = 4.61$, $p < .001$), demonstrating that this effect is robust using an analysis that is less sensitive to influential observations. In addition, this analysis revealed a significant main effect of age ($\beta = 1.49$, 95% CI = (0.94, 2.04), $t(58) = 5.29$, $p < .001$) and a significant interaction ($\beta = 1.22$, 95% CI = (0.62, 1.82), $t(58) = 3.99$, $p < .001$). The main effect of age suggests that older children sought out more evidence than younger children overall. The reliability by age interaction suggests older children were more sensitive than younger children to variation in environmental reliability. Older children in our sample, and particularly the 7-year-olds, were more likely than younger children to sample a high number of zorpies when they had encountered a high proportion of false information in the past.

Study 3: Simulation

In Studies 1 and 2, the test claims were true. Yet, the selective skepticism that children exhibit in these studies is theoretically adaptive because increased information sampling facilitates the discovery

of counterevidence. If the test claim was actually false, what kind of environment would best prepare children to discover that? We ran a simulation to determine whether experience learning in an unreliable environment enables children to identify misinformation more easily. In each of 400 simulation runs, we randomly sampled a proportion of the available zorpies to serve as hypothetical counterevidence to the test claim (i.e., zorpies *without* three eyes). Then, using participants' real sampling behavior from Study 1, we calculated the percentage of participants in each condition who revealed one or more of the counterevidence zorpies before making a decision about the test claim. Thus, each run of the simulation represents a hypothetical experiment with potential counterevidence randomly distributed among the 20 zorpies. We simulated four different proportions of counterevidence in the sample space with 100 simulation runs each, such that either 10%, 20%, 30%, or 40% of the available zorpies provided evidence against the test claim. The results of the simulation are in Figure 4.

To test whether children in the unreliable condition were more likely to discover counterevidence during sampling, we used the *betareg* package in R to run a beta regression using experimental condition (reliable vs. unreliable) to predict the percentage of Study 1 participants who would have discovered one or more pieces of counterevidence during sampling. This analysis revealed a main effect of condition ($\beta = 0.86$, $p < .001$), and the effect holds after controlling for the simulated amount of counterevidence ($\beta = 0.93$, $p < .001$). These results support the commonsense conclusion that children in the unreliable condition, who sampled more evidence bearing on the test claim, would have been more likely to discover counterevidence had it been available. Unreliable learning environments elicit increased skepticism and thus enable children to debunk misinformation more readily.

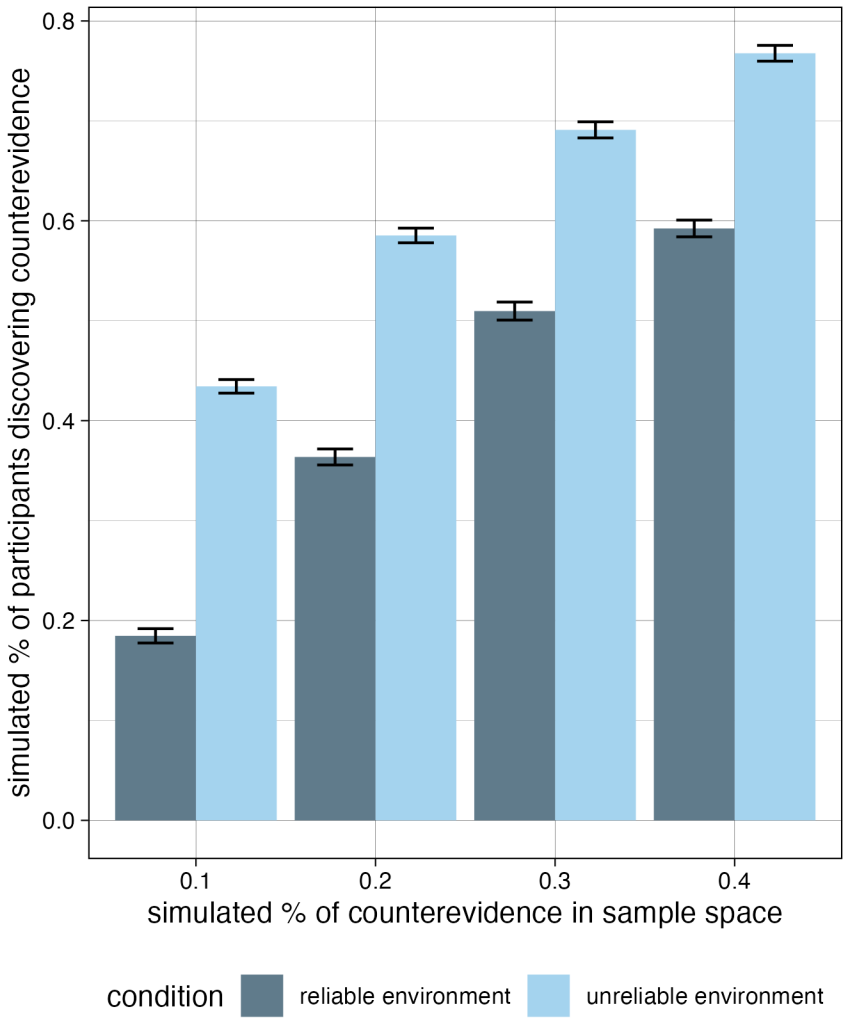

Figure 4. Simulation results reveal that children in the unreliable condition of Study 1 would have been more likely to discover counterevidence than those in the reliable condition. This pattern holds when counterevidence is both rare and relatively common. Error bars represent 1 SEM.

General Discussion

In order to learn accurately and efficiently, children must have an adaptive policy for deciding which claims to trust on the spot, and which to seek more evidence for. In two experiments, we investigated whether children use the reliability of their informational environment to make rational inferences about whether a new claim warrants fact-checking. Study 1 demonstrated that children seek more evidence for a novel claim about aliens that arises in a context containing some misinformation

about animals. In Study 2, children flexibly adapted their evidentiary standards according to the prior reliability of their environment. The proportion of false animal statements kids heard in the exposure phase positively predicted the amount of evidence they sampled before verifying a test claim about aliens. This was true even in a search engine context in which each statement derived from a distinct source. Thus, children made smart inferences about their context, beyond speaker-specific cues, to decipher how much to trust new information. They made fine-grained assessments of the reliability of incoming information in a known domain, inferred that this reliability would generalize to another domain, and flexibly chose a graded evidentiary standard corresponding to that reliability. Finally, we showed with a simulation (Study 3) that this behavior is adaptive: learners have the greatest opportunity to discover counterevidence and debunk misinformation in the most unreliable environments, where misinformation is most likely to be present.

Children's ability to calibrate their evidentiary standards to the reliability of their environments helps them confront the challenge of balancing speed and accuracy during learning. Children wasted little time verifying a claim within an environment with established reliability in a known domain. Instead, they reserved more extensive information seeking for more questionable informational contexts, calibrating their evidentiary standards according to nuanced changes in reliability. While this strategy is certainly not infallible, it gives children a sensible policy for information seeking in line with resource-rational decision making (Bhui, Lai & Gershman, 2021). Even when they lack domain-relevant knowledge to judge a claim's content, they leverage sophisticated attributes of their context to guide their skepticism and exploration selectively. However, it is important to note that the scope of children's inferences–what exactly constitutes the environment that children assessed–is unclear. Future research should clarify the conceptual and temporal constraints on which experiences affect children's expectations about incoming information.

Our experiments used an open information sampling task in order to capture a graded sense of children's level of skepticism or evidentiary standards. This continuous measure allowed us to capture the quantity of evidence children searched for, which corresponded to *degrees* of belief in a given claim. This approach provides direct insight into the strength of children's belief, which is what guides their future learning and behavior. The information sampling measure we employed is also implicit, which makes it more suitable for use in younger children than explicit reports (Goupil & Kouider, 2019). This work thus builds upon literature that demonstrates that infants' and children's information seeking behavior is sensitive to uncertainty (Langenhoff, Engelmann & Srinivasan, 2023; Lapidow, Killeen & Walker, 2022; Goupil, Romand-Monnier, & Kouider, 2016). We show that information seeking is sensitive to *environmental* certainty, as well as content-specific certainty.

We demonstrate that the knowledge that novel claims require evidence is early emerging and context-sensitive. This suggests that the most fruitful avenue of intervention may not be on skepticism itself, but on children's more specific capacities to know where to look for relevant evidence in a given domain, and to evaluate how different kinds of evidence bear on complex claims. Indeed, research suggests that children aren't sensitive to the relative strengths of explanations until early school age (Danovitch, Mills, Sands & Williams, 2021). Future research should address how, and when over the course of development, children adapt their evidentiary standards in terms of the *quality* of evidence that they seek.

A central insight of this work is that children's approach toward novel information is shaped by expectations that are formed through experience with their informational environment. This suggests that efforts to expose children only to curated informational environments may be misguided. Early experiences with overly sanitized environments may lead children to develop overly trusting priors and rob them of opportunities to develop critical thinking skills. By the same token, early exposure to more heterogeneous informational environments may allow children to "flex their skepticism muscles" and

build upon their existing capacities for adaptive information seeking. This idea is consistent with evidence that exposing adults to blatant misinformation makes them less susceptible to more subtle misinformation compared to a control condition (Loftus, 1979), although recent direct and conceptual replications failed to find this effect (O'Donnell & Chan, 2023). More broadly, intervention efforts should focus on helping children develop a broad skill set for evaluating information, rather than attempting to control their information diets.

Data Availability

Data for all experiments are available at https://osf.io/7hxkt/.

Code Availability

Code for the simulation and analyses is available at https://osf.io/7hxkt/.

Acknowledgements

We thank CiCi Jiang and Jolie Witkowski for their assistance with data collection, and Steve Piantadosi and members of the Kidd Lab for helpful discussions. This work was supported by the Walton Family Foundation, Jacobs Foundation, John Templeton Foundation, and Berkeley Center for New Media. Last but not least, we are grateful to the children and families who made this research possible.

References

Baer, C., & Kidd, C. (2022). Learning with certainty in childhood. *Trends in Cognitive Sciences,* 26(10), 887–896.

Baranes, A. F., Oudeyer, P. Y., & Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neuroscience*, *8*, 317.

Basol, M., Roozenbeek, J., & Van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1).

BBC Trending (2017, March 27). *The disturbing YouTube videos that are tricking children.* BBC. https://www.bbc.com/news/blogs-trending-39381889

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*(6), 1641-1660.

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences,* 41, 15-21.

Braddock, K. (2022). Vaccinating against hate: Using attitudinal inoculation to confer resistance to persuasion by extremist propaganda. *Terrorism and Political Violence*, *34*(2), 240-262.

Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1088.

Capewell, G., Maertens, R., Linden, S., & Roozenbeek, J. (2023, July 24). Misinformation interventions decay rapidly without an immediate post-test. PsyArXiv preprint. https://doi.org/10.31234/osf.io/93ujx

Chan, M. P. S., & Albarracín, D. (2023). A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour*, 1-12.

Common Sense Media (2019). *New Survey Reveals Teens Get Their News from Social Media and YouTube.* Common Sense Media.

Common Sense Media (2023). *New Poll Finds Parents Lag Behind Kids on AI and Want Rules and Reliable Information to Help Them.* Common Sense Media.

Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post‐truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602.

Coughlin, C., Hembacher, E., Lyons, K. E., & Ghetti, S. (2015). Introspection on uncertainty and judicious help‐seeking during the preschool years. *Developmental Science*, *18*(6), 957-971.

Danovitch, J. H., Mills, C. M., Sands, K. R., & Williams, A. J. (2021). Mind the gap: How incomplete explanations influence children's interest and learning behaviors. *Cognitive Psychology,* 130, 101421.

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761-778.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29.

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993.

Fazio, L. K., & Sherry, C. L. (2020). The effect of repetition on truth judgments across development. *Psychological Science*, *31*(9), 1150-1160.

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction*, *32*, 1-9.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822-15826.

Goupil, L., & Kouider, S. (2019). Developing a reflective mind: From core metacognition to explicit self-reflection. *Current Directions in Psychological Science,* 28(4), 403-408.

Goupil, L., & Proust, J. (2023). Curiosity as a metacognitive feeling. *Cognition*, *231*, 105325.

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.

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, 1-3.

Hermansen, T. K., Ronfard, S., Harris, P. L., & Zambrana, I. M. (2021). Preschool children rarely seek empirical data that could help them complete a task when observation and testimony conflict. *Child Development*, *92*(6), 2546-2562.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, *21*(10), 1541-1547.

Kallioniemi, P. (2021). The Role of Human Curation at the Age of Algorithms. *Journal of Digital Media & Interaction*, *4*(10), 7-20.

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. *Science*, *380*(6651), 1222-1223.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, *126*(1), 109-114.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Langenhoff, A. F., Engelmann, J. M., & Srinivasan, M. (2023). Children's developing ability to adjust their beliefs reasonably in light of disagreement. *Child Development,* 94(1), 44-59.

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.

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348-384.

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied*, *15*(4), 307.

Loftus, E. F. (1979). Reactions to blatantly contradictory information. *Memory & Cognition*, *7*(5), 368-374.

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1.

Maheshwari, S. (2017, November 4). *On YouTube Kids, Startling Videos Slip Past Filters.* The New York Times. https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html/

Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*.

Mott Poll (2021). *Sharing Too Soon? Children and Social Media Apps.* C.S. Mott Children's Hospital. https://mottpoll.org/reports/sharing-too-soon-children-and-social-media-apps/

O'Donnell, R., & Chan, J. C. (2023). Does blatantly contradictory information reduce the misinformation effect? A Registered Report replication of Loftus (1979). *Legal and Criminological Psychology*.

Plate, R. C., Shutts, K., Cochrane, A., Green, C. S., & Pollak, S. D. (2021). Testimony bias lingers across development under uncertainty. *Developmental Psychology*, *57*(12), 2150.

Rodriguez, A. (2018, April 26). *YouTube Kids is giving parents more control over what their kids watch.* Quartz. https://qz.com/1262977/youtube-kids-is-launching-a-mode-curated-by-humans-not-just-algorithms/

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. Harvard Kennedy School Misinformation Review, 1(2). https://doi.org/10.37016//mr-2020-008

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Salles, A., Ais, J., Semelman, M., Sigman, M., & Calero, C. I. (2016). The metacognitive abilities of children and adults. *Cognitive Development*, *40*, 101-110.

van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460-467.

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008.

Wang, J., Yang, Y., Macias, C., & Bonawitz, E. (2021). Children with more uncertainty in their intuitive theories seek domain-relevant information. *Psychological Science*, *32*(7), 1147-1156.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, *29*(7), 1159-1177.

Williams (2023, June 7). *The Fake News about Fake News*. Boston Review. https://www.bostonreview.net/articles/the-fake-news-about-fake-news/

Wong, N. C. H. (2016). "Vaccinations are safe and effective": Inoculating positive HPV vaccine attitudes against antivaccination attack messages. *Communication Reports,* 29(3), 127–138.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015.

Xu, S., Shtulman, A., & Young, A. G. (2022). Can Children Detect Fake News?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).