Stop, Children What's That Sound: Multi-Modal Inference Through Mental Simulation

Joseph Outa*, Xi Jia Zhou*, Hyowon Gweon, Tobias Gerstenberg
Stanford University
*joint first authors

Social Learning Lab
Stanford University

Background

- Young children can successfully use auditory evidence to infer the cause of an event.[1,2]
- Prior work argues such physical inferences are driven by an approximate mental simulation of the physical world.[3,4]
- However, children struggle to integrate visual and haptic information until late childhood.[5] whereas adults can successfully combine visual and auditory evidence via simulation to figure out what happened.[4]
- The developmental emergence of multi-modal integration remains an open question.
- Question: How do children use visual and auditory evidence to reason about a past event?

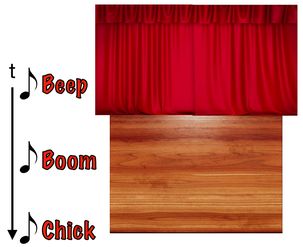
The plinko task

[N = 64, 3 – 8 years, 9 trials]

1. Elmo on 'plinko box'

2. Box covered

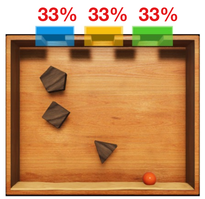
3. Elmo covered; ball dropped



4. "In which hole did Elmo drop the ball?"

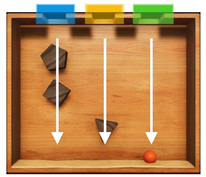
Computational Models

guessing model



- Children might randomly select among the holes.
- Random guessing formalized by assigning equal probability to each of the holes.
- $P_{hole1} = P_{hole2} = P_{hole3} = 33\%$

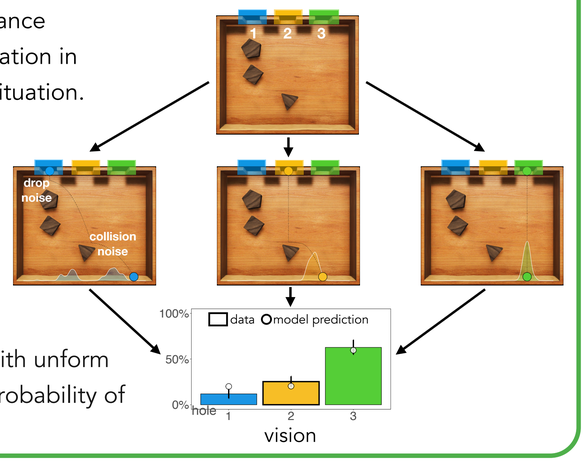
matching model



- Children might use simple 'matching' heuristic, formalized by assigning probability to each hole proportional to its distance from ball.
- The closer the ball to center of hole along x-axis, the higher probability assigned to hole.

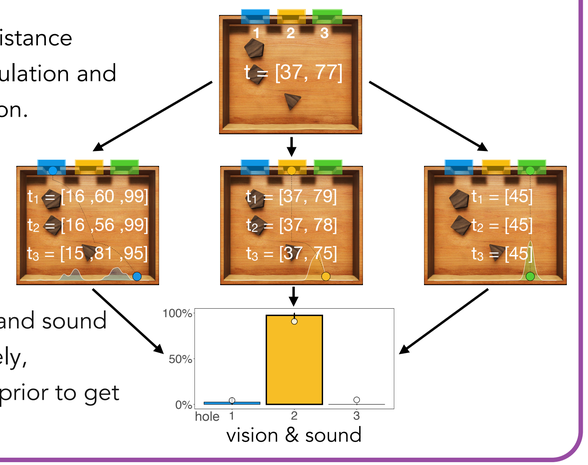
simulation (vision) model

- Noisy physical simulations of outcome if ball dropped in each hole.[4]
- Visual likelihood = distance between ball's final location in simulation and actual situation.
- Only considers visual information.
- Likelihood combined with unform prior to get posterior probability of each hole.



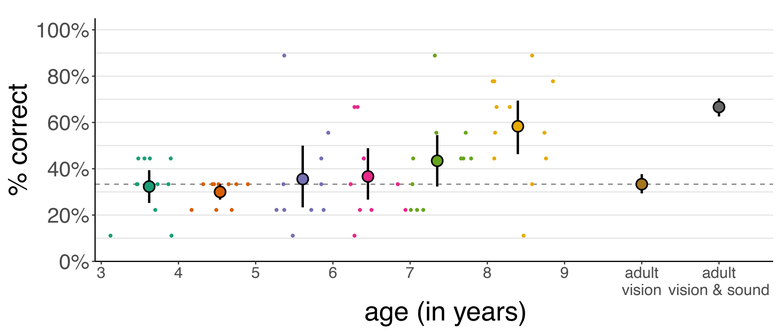
simulation (vision and sound) model

- Auditory evidence encoded as vector of collision time-points.[4]
- Auditory likelihood = distance between vectors in simulation and vectors in actual situation.
- Likelihoods from vision and sound integrated multiplicatively, combined with uniform prior to get posterior probabilities.



Results

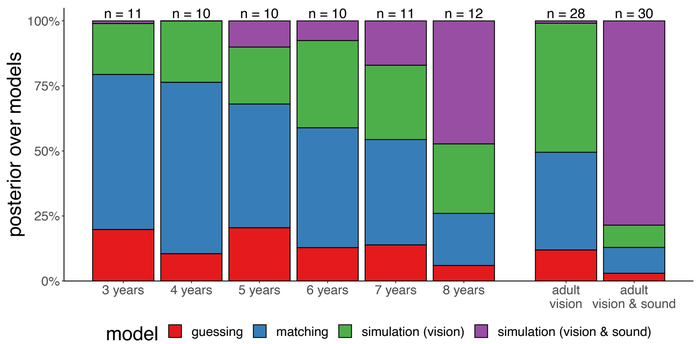
accuracy with respect to ground truth



- Overall increase in accuracy with age $\beta = 0.22$ [0.11, 0.33]
- Model-based analysis reveals how strategies change with age.

model posteriors

posterior probability of each model given the responses, aggregated by age



- Gradual transition from mostly relying on matching to relying on simulation, first by only considering visual evidence then considering both visual and auditory evidence.

Discussion and Future Directions

- Even young children's reasoning about past events may show signatures of simulation that consider not only visual but also auditory information, and this tendency increases throughout early childhood.
- Why was the task challenging? Children's knowledge of real-world physics may not yet be robust enough to run accurate simulations; task demands on working memory.
- Future work may explore more simplified versions that minimize gravity bias & task demands.

References

1. Siegel et al. (2021) *Nature Communications*, 2. Agrawal & Schachner (2022) *Developmental Science*, 3. Smith & Vul (2012) *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 4. Gerstenberg et al. (2021) *PsyArxiv*, 5. Gori et al. (2008) *Nature Precedings*.