Stumped! Learning to think outside the box in 3-7-year old children

Junyi Chu
Department of Psychology
Stanford University
junyichu@stanford.edu

Misha O'Keeffe
Department of Psychology
Stanford University
mokeeffe@stanford.edu

Silvia K. Liu
Department of Psychology
College of the Holy Cross
kliu@holycross.edu

Elizabeth Bonawitz
Graduate School of Education
Harvard University
elizabeth_bonawitz@gse.harvard.edu

Tomer D. Ullman
Department of Psychology
Harvard University
tullman@fas.harvard.edu

Abstract

Many theories conceptualize thinking as search through a space of hypotheses. But what if your initial space is wrong? What cognitive skills support abandoning an ineffective hypothesis space and re-construing a problem with the correct hypothesis space? We examined the development of such abilities in n=172 children ages 3-7 years using *Stumper* riddles, which challenge respondents to explain seemingly impossible situations. We found evidence that children both learned the relevant hypothesis space for different categories of *stumpers* and generalized the cognitive strategy across riddle categories. Although older children showed greater overall accuracy, these effects of learning and meta-learning were found even for the youngest 3-5-year-olds. These results suggest a promising method for probing both flexible hypothesis search and meta-cognitive skills. We discuss ongoing plans to characterize individual differences as a way to uncover the underlying mechanisms of creative problem-solving.

Keywords: problem solving, social reasoning

Introduction

Dee Septor, the famous magician, claimed to be able to throw a ping-pong ball so that it would go a short distance, come to a dead stop, and then reverse itself. He also added that he would not bounce the ball against any object or tie anything to it. How could he perform this feat? [1]

This is a *stumper*, a riddle that challenges people to explain a seemingly impossible or even paradoxical situation (Bar-Hillel et al., 2018). In one study, only 36% of adult participants answered the opening riddle correctly (Ansburg & Dominowski, 2000).

In cognitive science, many theories conceptualize thinking as a search through a space of hypotheses (Austerweil et al., 2015; Newell & Simon, 1972). Stumpers are crafted to set up the *wrong* space: they elicit a dominant representation of the situation in which no suitable answer can be found. This characteristic makes *Stumpers* a powerful tool for studying how individuals can both get stuck in the wrong hypothesis space, and how they can learn to get unstuck by restructuring their mental representations and "thinking outside the box" (Weisberg, 2019).

Recent work has examined how adults respond to stumper riddles. For example, Bar-Hillel, Noah, and Frederick (2019) investigated the relationship between solving stumpers and performance on other cognitive tasks, specifically the Cognitive Reflection Test (CRT) and the Compound Remote Associates Test (CRAT). Their findings suggest that the ability to solve stumpers may rely on similar cognitive mechanisms as those involved in overcoming initial representations (as in the CRT) but less with making novel associations between seemingly unrelated concepts (as in the CRAT).

While the previous work sheds light on what makes stumpers tricky, here we are interested in cases of success. By construction, solving a stumper means engaging in a different kind of hypothesis search: beyond simply searching *within* a given hypothesis space, respondents must realize that their initial space is inadequate, and figure out how to reformulate and "jump" to a different hypothesis space – a kind of *meta*-hypothesis search.

Some research suggests that people can learn to make such "jumps". Like many other instances of problem-solving (e.g., Bonawitz & Griffiths, 2010; Duncker, 1945), performance on stumper riddles can be improved by exposing participants to diverse problem types, encouraging them to compare and contrast riddles, and helping them recognize when to abandon unproductive solution attempts (Ansburg & Dominowski, 2000). These findings suggest that the ability to solve stumpers might be improved through targeted interventions that foster meta-cognitive awareness.

Meta-cognitive skills – the ability to reflect on and regulate one's own thinking process (Flavell, 1979) – are rapidly developing in childhood. Yet, to date, there is no documented research on how children respond to stumper riddles, or how they engage in meta-hypothesis search more broadly. Given that performance on stumpers in adults can be improved by providing examples, there is an interesting opportunity to test whether children can also benefit from such an intervention. A positive result would be consistent with the hypothesis that children can engage in meta-hypothesis search. We set out to fill this gap.

We ask three empirical questions: (1) Can children ages 3-7 successfully produce solutions to stumper riddles? And, assuming that children will find these riddles difficult, (2) can children learn from hearing a riddle solution, demonstrated by effectively transferring a previous solution to a new riddle? Finally, (3) do children demonstrate meta-hypothesis search by showing improved performance on later riddles, across different kinds of stumpers?

[1] A solution is provided at the end of the paper.

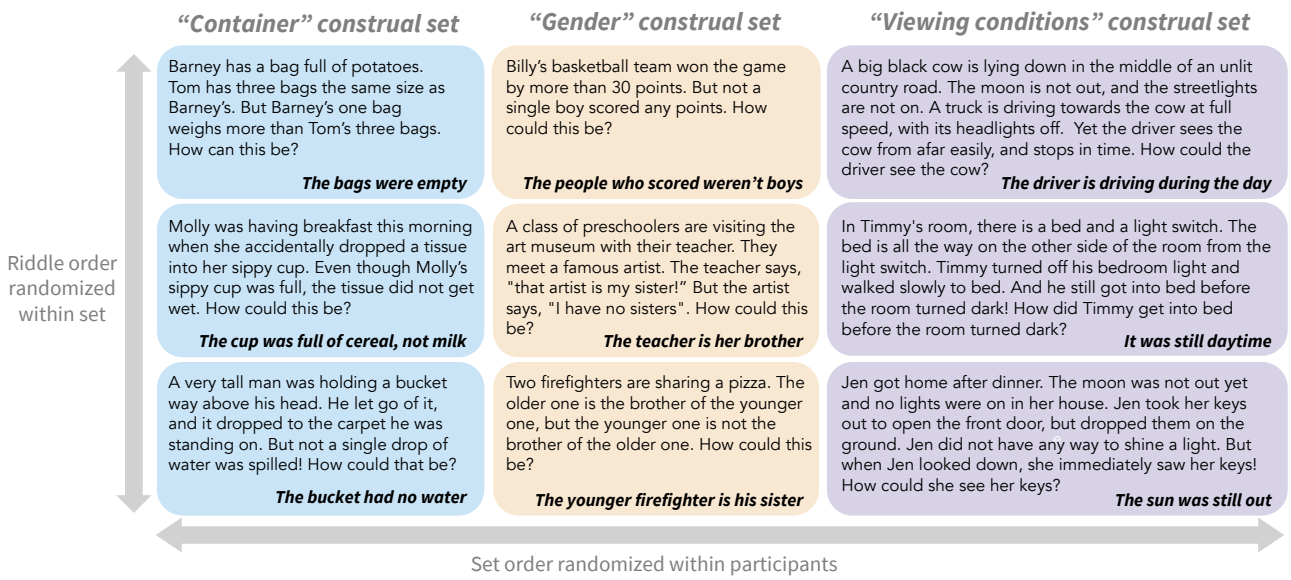

Figure 1: The 9 stumper riddles used in the current study, organized by construal set. A condensed version of the presented solution is shown in bold italics below each riddle.

We report a pre-registered study where children responded to 9 stumpers. After providing a response, children always heard the target solution before proceeding to the next riddle. Figure 1 shows the riddles and solutions used in the current study. To probe both learning and meta-learning, we designed riddles with varying amounts of similarity: some stumpers could be answered with exactly the same response (e.g., *viewing condition* items could all be answered with "it was daytime"); others with a similar target of reconstrual but in different ways (e.g., within "*gender*" set).

If children are learning narrowly from a previous solution, they might show improved performance only on very similar riddles that can be answered with the identical response (i.e., on viewing riddles). In contrast, if children are learning more generally to explore different hypothesis spaces or to construct alternative mental scenes, they may correctly answer new riddles that require quite different solutions, and show improved performance on riddles from different sets.

Methods

This study was conducted as part of Project GARDEN, an NSF-funded collaboration between multiple labs where 3- to 7-year-old children could complete up to 16 separate study modules examining different aspects of cognitive development (e.g., executive function, vocabulary, counterfactual reasoning), with a mixture of asynchronous and synchronous studies. Each module was managed independently by different research teams, with parents and caregivers providing informed consent on a module by module basis. Before completing the present study, participants had completed at least the first 8 modules over several months.

Study methods and analyses were preregistered on the Open Science Framework at `https://osf.io/bzkm6`. Study materials, anonymized data, and analyses can be found at `https://osf.io/reaz5/`.

Participants

Our analysis includes 1479 responses from 172 children (89 female) aged 3 to 7 years (M = 65.24 months, SD = 13.96 months, range 39–92 months) who were recruited and tested via Children Helping Science (Scott & Schulz, 2017). Caregivers reported their children's race/ethnicity as White (80.2%), Asian (23.3%), Hispanic/Latino (10.5%), Middle Eastern (4.1%), Black (2.3%), and Pacific Islander (1.2%). Most participants were monolingual English speakers (70.3%), while 29.1% children were exposed to two or more languages, with the most commonly reported languages being Somali (n = 12), Spanish (n = 7), and French (n = 6). All participants received a $5 gift card for participating. The study was approved by the Institutional Review Board at Harvard University.

Following our pre-registered exclusion criteria, data from another 44 sessions were collected but deemed unusable due to technical difficulties (12), repeat participation (11), incomplete sessions (11), distraction (4), language (2), being present during a sibling's session (2), interference from siblings or caregivers (1), or were outside the age range (1). Note that some of these sessions had partially usable data, but we excluded any participant who provided fewer than 6 valid trials.

Materials

We designed 9 stumper riddles organized into three sets (see Figure 1 for the full stimuli text). These riddles were borrowed from prior work (Bar-Hillel et al., 2018; Bar-Hillel, 2021) but adapted to be more age-appropriate, such as us-

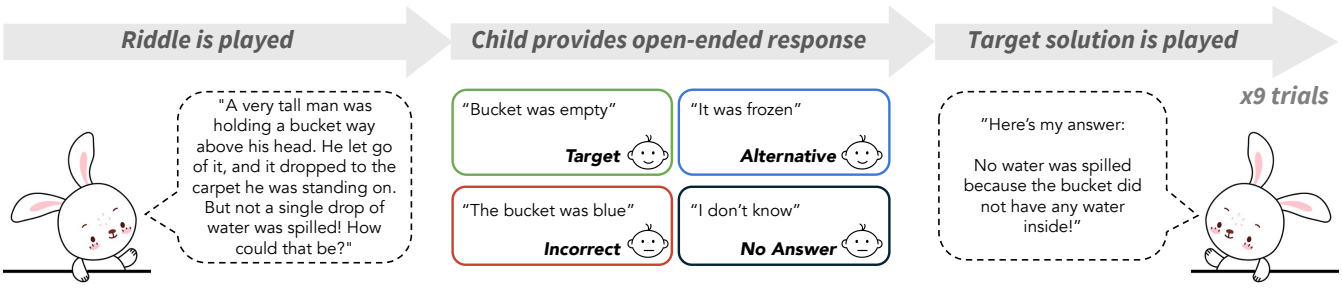

Figure 2: Procedure of a single trial, from left to right. First, participants heard the riddle, and were allowed to replay it as many times as they liked. Participants then respond verbally while being recorded on videotape. After submitting their response, participants hear a solution for the riddle which provides a target construal. This procedure repeated for all 9 riddles, with a different animal character posing each riddle.

ing more familiar objects (e.g., replacing 'coffee mug' with 'sippy cup') or situations (e.g., a child going to bed), and using simpler vocabulary throughout.

Riddles from the same set targeted a similar default construal, and therefore could be solved using a similar reconstrual strategy. In the *container* set, riddles can be solved by re-considering whether a container is full or empty. In the *gender* set, riddles leverage common gender-related beliefs (e.g., that firefighters are typically male, or that players on sports teams tend to be of the same gender). These riddles can be solved by re-considering the gender of the protagonists. Finally, riddles in the *viewing conditions* set worked by suggesting that a the situation happened in the dark (e.g., that it was bed time or after dinner); these riddles could be solved by imagining that the scene could have happened in daylight.

Procedure

Participants completed the web-based study asynchronously from home, with responses captured through webcam recordings and clicks. Caregivers were asked to stay and help children navigate through the task, although they were not always visible in the video recording. At the start of the study, caregivers first watched an instructional video that introduced the study procedure and provided examples of ways to encourage a response from their children without providing clues.

Participants then watched an introductory video that presented the task as a "question game". In the video, 9 animal characters were introduced as friends who would take turns to ask them questions. Children would hear the 9 riddles one by one, organized into three blocks. Riddles in each block belonged to the same riddle category and could thus be solved using a similar re-interpretation of the situation. There are no pauses between the blocks, so participants experienced the task as a continuous sequence of 9 trials, without explicit signal that there are 3 blocks of riddles. Block order and trial order within each block were randomized.

Each trial followed the same sequence of events. First, an animal character would present the situation and pose the riddle by asking, "How can that be?" Then, the children would provide their responses verbally. Participants could replay the question as many times as needed. After submitting a response, participants would hear the target solution from the character, who announced, "Here's my answer: ...". To encourage caregivers to sit through the entire study, we asked them to judge at the end of each trial if their child's answer was similar to or different from the provided solution, or if their child did not answer. These judgments are available in the dataset but not included for further analyses.

Response coding

Trained research assistants transcribed and coded children's responses while being naive to trial order. Our primary dependent variable is trial-level accuracy, i.e., whether children referred to the same concept as our target solutions (**Target** responses). We also identified **Alternative** responses that provided a valid solution to the riddle, but did not involve the same reconstrual. For example, an alternative solution to the Bucket riddle could be that the bucket was full of ice. Responses were coded as **Incorrect** if they included fantastical or impossible situations (e.g., that an object turned invisible), rejected the premise of the riddle (e.g., suggesting that a character was lying), or were otherwise irrelevant or nonsensical. Finally, we coded responses as **No answer** if the child expressed ignorance ("I don't know") or otherwise refused to answer. Note that we excluded trials where participants immediately proceeded to the solution phase without pause.

To assess coding reliability, a second rater coded 20% of all trials (n=296). We found high agreement when determining whether participants gave the target response or not (agreement = 97%). When considering all four possible response categories, agreement was 88% (kappa = 0.8, 95% CI [.74, .86].

We note that all data are open access and freely available at: `https://github.com/jchu10/stumpers_cogsci2025`. We include coded responses for reproducibility, as well as de-identified transcriptions (combed for removing any personal information) for supporting future researchers' novel exploratory analyses.

Results

We first verify that these riddles "stump" children by examining performance on the very first trial, before children are exposed to any correct answers or re-construals. Most children provided incorrect responses (63.6%) or no response

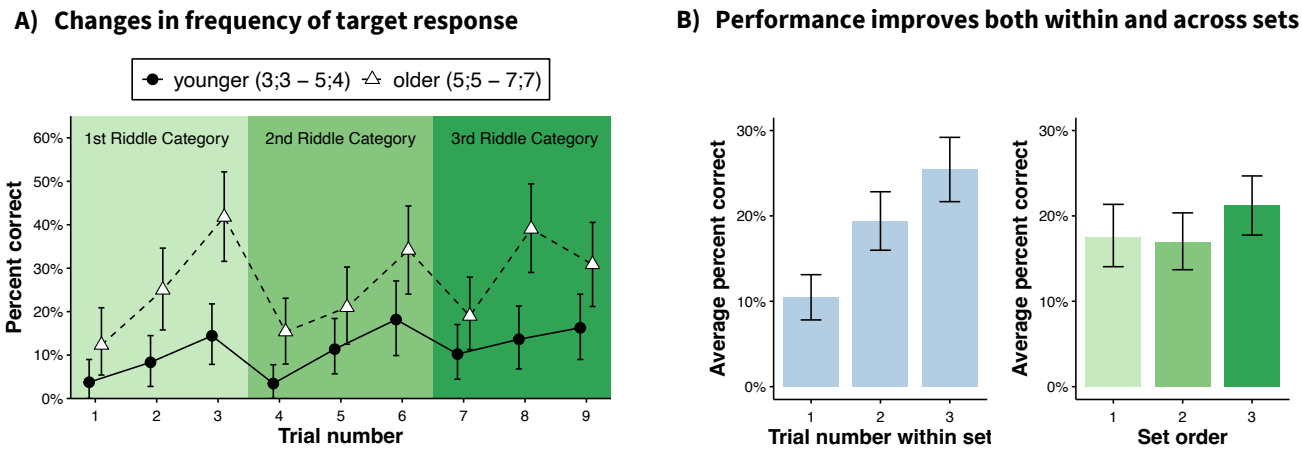

Figure 3: Performance on stumper riddles. (A) Changes in riddle accuracy by younger children (3;3-5;4 years; n=90; solid line) and older children (5;5-7;7 years; n=82, dashed line). Points show proportion of children who produce the target solution on each trial; error bars show bootstrapped 95% confidence intervals. (B) Participants in both age groups show improved performance on later trials within each riddle category, with reduced performance when receiving the first riddle of a new category.

(21.0%). A small proportion of children provided the target solution (8.2%) or an alternative but valid response (7.3%).

Performance on the first trial also varied by riddle set ($\chi^2(2) = 93$, $p < .001$). Children were most likely to provide target solutions on the first trial for *viewing condition* riddles (11.1%, 95% CI [2.63%, 22.0%]), followed by *container* riddles (8.3% [1.9%, 17.0%]), and *gender* riddles (3.9% [0%, 12.5%]).

Aggregating across trials, children produced the target solution on an average of 1.51 riddles (SD = 1.47) out of 9, with older children producing more target solutions (Pearson's $R(170) = .45$, $p < .001$; age in months). Figure 4 shows overall age-related improvements.

Performance improved within riddle sets

Our main question is whether providing examples of reconstruals and riddle solutions would support children in generating appropriate solutions to later riddles, when controlling for age and individual variability across participants and riddles. To do so, we fit a logistic mixed effects model predicting accuracy (1=target solution, 0=other) from within-set trial order (1-3), set order (1-3), their interaction, and age (in months, z-scored). We also included random intercepts for participant and riddle:

```
glmer(accuracy ~ trialOrder * setOrder + ageMonths
+ (1|childID)+ (1|riddleID), family=binomial)
```

We found strong evidence of improvement across riddles from the same set. Within each set, children were significantly more likely to provide the target solution on later trials ($OR_{TrialOrder} = 2.56$, 95% CI [1.84, 3.55], $p < .001$), with the proportion of target responses averaging 10.5% on the first riddle of each set to 25.5% on the third riddle (Figure 3B).

Examining each riddle type separately, we found the largest improvements within gender riddles (from 5.1% to 21.1%, or 4.13x), followed by viewing condition riddles (from 13.1% to 35.9%, or 2.74x), and container riddles (from 13.8% to 18.9%, or 1.37x).

Performance improved across riddle sets

We also found evidence for improvement across riddle sets ($OR_{SetOrder} = 1.67$, 95% CI [1.19, 2.45], $p = 0.003$), with average accuracy increasing from 17.5% in the first set to 21.2% in the third set. Looking at performance on the first riddle in each block, we found that overall accuracy improved from 7.6% (SD = 26.6) on trial 1, to 9.1% (SD = 28.8) on trial 4 and 14.4% (SD = 35.2) on trial 7.

Next, we examined whether improvement across trials would vary between earlier and later sets. Model comparison using a likelihood ratio test found a significant interaction between trial order and set order ($\chi^2(1) = 4.37$, $p = 0.037$), with trial-by-trial improvement declining later in the study. Comparing accuracy on the first and last trial in each block of riddles, we found that improvement rate decreased from 3.58x in the first block to 2.83x in block two and 1.63x in block three.

Developmental changes in stumper performance

Older children produced target solutions more often (Pearson's $r(170) = .45$, 95% CI [.32, .56], $p < .001$). However, we found no evidence for developmental change in the rate of learning across trials: model fit did not significantly improve when including either an age by trial order interaction ($\chi^2(2) = 0.84$, $p = 0.36$) or an age by set order interaction ($\chi^2(2) = 0.06$, $p = 0.8$).

Since there was significant variation in accuracy across riddles and sets, we also examined the relative improvement from the first to third trial within each set, comparing younger and older children (median split on age). For container riddles, younger children showed a 2.9x improvement (3.4% to 10.0%) while older children showed a 2.8x improvement (10.3% to 29.4%). For gender riddles, younger children improved 1.7x (5.6% to 9.5%) while older children improved

from 0% to 24.1%. Viewing riddles showed the strongest improvements, with younger children improving 7.7x (4.3% to 33.3%) and older children improving 2.7x (20% to 54.6%).

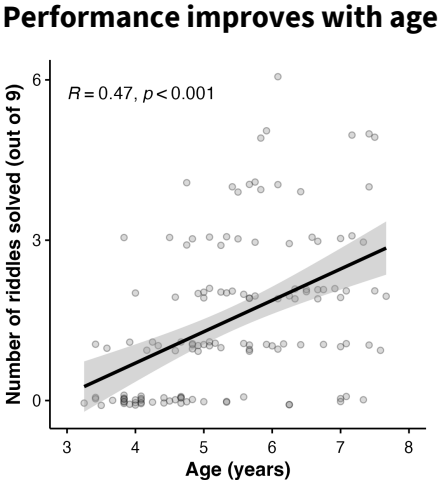

Figure 4: Performance on stumper riddles improves with age. Each point represents an individual participants' overall accuracy.

Changes in response types

The preceding analyses show that the proportion of target solutions increased over trials, consistent with our hypothesis that providing example reconstruals and solutions would support children's production of target solutions on later riddles. However, these results are also compatible with alternative explanations. For example, children might be warming up to the task and simply responding more often later in the study. This would predict a decrease in the proportion of "no response" over trials. Alternatively, children might maintain the same rate of responding, but become less likely to produce incorrect or alternative responses. This would be consistent with a shift towards a hypothesis space containing the target response as opposed to other suitable solutions.

We ran an exploratory analysis examining these possibilities. To investigate how the proportion of response categories changed across trials, we fit a multinomial logistic regression model predicting response type (no-response, target solution, alternative solution, or incorrect response) from within-set trial order (1-3), set order (1-3), their interaction, and age in months. `multinom(response_category ~ trialOrder * setOrder + age_months, family=multinomial)`

Relative to target responses, the proportion of no-responses decreased significantly throughout the study ($OR_{TrialOrder}$ = 0.35 [0.24, 0.51], $p < .001$; $OR_{SetOrder}$ = 0.56 [0.38, 0.82], p = .003). The proportion of incorrect responses also decreased relative to target responses ($OR_{TrialOrder}$ = 0.45 [0.33, 0.62], p ¡ .001; $OR_{SetOrder}$ = 0.64 [0.46, 0.88], p = .007). Finally, the proportion of Alternative solutions declined within each set ($OR_{TrialOrder}$ = 0.58 [0.37, 0.91], p = .016) but not across sets ($OR_{SetOrder}$ = 0.79 [0.50, 1.24], p = .3).

Discussion

We reported first evidence that 3- to 7-year-old children can produce alternative problem construals in the context of a verbal riddle-solving task. While initial accuracy on the first riddle was low, simply hearing a correct solution improved children's performance on subsequent riddles, both for similar riddles that shared a hypothesis space (i.e., the same riddle set) and for riddles requiring very different solutions (i.e., different riddle sets). Although older children showed greater overall accuracy, these patterns of learning and meta-learning were found even for the youngest 3-5-year-olds.

Studying how people solve riddles follows a long tradition in cognitive science examining problem solving, creativity, and mental fixation. However, less work has examined how people establish an initial set of possible hypotheses in the first place (though see Bonawitz & Griffiths, 2010; Brockbank & Walker, 2022). The present results suggest a promising method for probing flexible meta-hypothesis search among possible problem construals.

However, we note some limitations of the current study. Due to the verbal nature of stumper riddles, performance necessarily relies on children's developing language abilities, which may underestimate children's ability to engage in flexible meta-hypothesis search. Also, our analyses focused on children's final response. Future work may probe more deeply into children's reasoning process, such as by analyzing how response times vary with increased riddle exposure or by using interactive think-aloud protocols to examine what ideas children explored before arriving at their final response.

While these results are consistent with the idea that children can both recognize when they are stuck and learn to search in more appropriate hypothesis spaces, there are many other factors that plausibly impact children's performance. For example, stumpers are designed to elicit salient default construals that form an inadequate hypothesis space (Bar-Hillel et al., 2018), however, children may interpret these riddles differently than adults. In particular, children's understanding of gender roles and the reciprocal nature of kinship terms are still developing in the sampled age range (e.g., Mitchell & Jordan, 2021). Such conceptual difficulties might partly explain why riddles in the *gender* set were especially tricky for children on the first trial.

Other developmental and individual differences in cognitive skills may also impact children's baseline performance, as well as their readiness to learn from example solutions. Therefore, we recruited a large sample of participants to both enhance the generalizability of our findings and to support planned analyses of such factors. As part of Project GARDEN, participants completed a suite of other cognitive and behavioral measures. These measures include vocabulary, executive function, and counterfactual reasoning. We also obtained measures of children's performance on the Cognitive Reflection Task (Shtulman & Young, 2023), which is predictive of stumper performance in adults (Bar-Hillel et al., 2019). We are currently assessing correlations between these

tasks and children's stumper performance to better allow us to test specific hypotheses about the cognitive mechanisms supporting learning, hypothesis search, and meta-hypothesis search. For example, finding that counterfactual reasoning skills predict performance would support the hypothesis that successfully solving a stumper requires children to generate alternative problem construals.

Looking ahead, these findings point to the value of using riddles to measure flexible hypothesis search and generation in young children. Continuing to probe the mechanisms behind learning and transfer, as well as characterizing the structure of individual differences, is critical for advancing theoretical understanding of early reasoning skills and designing ways to scaffold creative problem solving.

Solution to opening riddle: He could throw it straight up.

Acknowledgments

We are grateful to all of participating families and Children Helping Science (powered by Lookit) for making this work possible. We thank Julio Caggiano, Kacper Malinowski, Janie Ro, Jiayi Song, and Cecilia Zhou for help with data coding. Special thanks to Herrissa Lamothe, members of the Harvard CoCoDev Labs, and the Stanford SLL for helpful discussions. This work was supported in part by the National Science Foundation under Grant No. 2042489 (EB), the James S. McDonnell Foundation (EB), the John Templeton Foundation (TU, JC), and the Jacobs Foundation (TU).

References

Ansburg, P. I., & Dominowski, R. I. (2000). Promoting insightful problem solving. *The Journal of Creative Behavior*, *34*(1), 30–60.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in bayesian models of cognition. *Oxford handbook of computational and mathematical psychology*, 187–208.

Bar-Hillel, M. (2021). Stumpers: an annotated compendium. *Thinking & Reasoning*, *27*(4), 536–566.

Bar-Hillel, M., Noah, T., & Frederick, S. (2018). Learning psychology from riddles: The case of stumpers. *Judgment and Decision Making*, *13*(1), 112–122.

Bar-Hillel, M., Noah, T., & Frederick, S. (2019). Solving stumpers, crt and crat: Are the abilities related? *Judgment and Decision Making*, *14*(5), 620–623.

Bonawitz, E. B., & Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in bayesian models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Brockbank, E., & Walker, C. M. (2022, August). Explanation impacts hypothesis generation, but not evaluation, during learning. *Cognition*, *225*, 105100. Retrieved from http://dx.doi.org/10.1016/j.cognition.2022.105100 doi: 10.1016/j.cognition.2022.105100

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, *58*(5), i–113. Retrieved from http://dx.doi.org/10.1037/h0093599 doi: 10.1037/h0093599

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Mitchell, A., & Jordan, F. M. (2021). The ontogeny of kinship categorization. *Journal of Cognition and Culture*, *21*(1-2), 152–177.

Newell, A., & Simon, H. A. (1972). Human problem solving. *Upper Saddle River/Prentive Hall*.

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14.

Shtulman, A., & Young, A. G. (2023). The development of cognitive reflection. *Child Development Perspectives*, *17*(1), 59–66.

Weisberg, R. W. (2019). Toward an integrated theory of insight in problem solving. In *Insight and creativity in problem solving* (pp. 5–39). Routledge.