Title
Learning to build physical structures better over time

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 42(0)

Authors
McCarthy, Will
Kirsh, David
Fan, Judith

Publication Date
2020

Learning to build physical structures better over time

Will McCarthy
Department of Cognitive Science
UC San Diego
wmccarthy@ucsd.edu

David Kirsh
Department of Cognitive Science
UC San Diego
kirsh@ucsd.edu

Judith Fan
Department of Psychology
UC San Diego
jefan@ucsd.edu

Abstract

Our ability to plan and build a wide array of physical structures, from sand castles to skyscrapers, is a defining feature of modern human intelligence. What cognitive tools enable us to create such complex and varied structures? Here we investigate how practice "reverse-engineering" a set of physical structures impacts the procedures that people subsequently use to build those structures, as well as how well they build them over time. Participants (N=105) viewed 2D silhouettes of 8 unique block towers in a virtual environment simulating rigid-body physics, and aimed to reconstruct each one in less than 60 seconds. We found that people learn to build each tower more accurately and quickly across repeated attempts, and that these gains reflect both group-level convergence upon a smaller set of viable policies, as well as error-dependent updating of each individual's strategies. Taken together, our study provides novel insight into how humans learn from prior experience to discover better solutions to physical reasoning problems over time.

Keywords: planning; spatial reasoning; intuitive physics; construction; action

Our ability to plan and build a wide array of physical structures, from sand castles to skyscrapers, is a hallmark of modern human intelligence. What cognitive mechanisms enable us to create such complex and varied structures? Towards answering this question, a natural starting point is to consider how people learn to "reverse-engineer" existing structures — that is, infer an appropriate decomposition of it that can be translated into a sequence of actions to recreate it from simpler components. Such problems are likely to recruit general-purpose mechanisms for physical reasoning and planning, in addition to mechanisms for learning from prior experience. Here we investigate the role of practice in guiding how people discover better solutions to such problems over time.

Our paper builds on classic work investigating how people reason about the properties of physical objects and how they interact with one another (McCloskey, 1983), a suite of abilities known as intuitive physics. A useful proposal emerging from more recent work on intuitive physics is that people reason about how physical systems evolve over time via mental simulation, which may provide a noisy approximation to real physical dynamics (Battaglia, Hamrick, & Tenenbaum, 2013; Sanborn, Mansinghka, & Griffiths, 2013; Hegarty, 2004). While many tasks in this literature involve passive judgments about physical scenes, a promising new direction is to consider tasks that involve active interventions on physical systems to achieve various goals (Allen, Smith, & Tenenbaum, 2019; Hamrick et al., 2018). In particular, the current study takes inspiration from prior work investigating how physical interventions can be beneficial for downstream performance on various physical reasoning tasks (Dasgupta, Smith, Schulz, Tenenbaum, & Gershman, 2018; Kirsh & Maglio, 1994).

Our paper is also informed by recent advances in theories of human planning that highlight the pervasive role of mental simulation in guiding human sequential decision making (Solway & Botvinick, 2015, 2012; Daw, Gershman, Seymour, Dayan, & Dolan, 2011), combined with reasonable assumptions about the cognitive costs of conducting mental simulations (Callaway et al., 2018; Hamrick, Smith, Griffiths, & Vul, 2015). While this progress has been galvanizing, the generalizability of current theories to construction behavior is potentially limited by the historically narrow focus on tasks with relatively low state-space complexity (van Opheusden, Galbiati, Bnaya, Li, & Ma, 2017), as well as abstract action spaces and state transitions (Daw et al., 2011) that are far removed from the physical environment. Moreover, these theories do not address our core question of how people make efficient use of prior task experience to overcome inherent cognitive resource limitations, thereby learning how to plan better over time.

Here we investigate how practice "reverse-engineering" a set of physical structures in a 2D virtual environment impacts the procedures participants subsequently use to build those structures, and how well they build them across repeated attempts. Our specific approach draws most direct inspiration from prior work in developmental science (Cortesa et al., 2018; Bullock & Lütkenhaus, 1988) and AI (Bapst et al., 2019; Jones, Hager, & Khudanpur, 2019) that have examined physical construction behavior. The current study advances these prior investigations in three ways: *first*, we develop a web-based environment for physical construction, facilitating the collection of large behavioral datasets; *second*, we examine how participants build the same structure multiple times, providing quantitative insight into how people adapt their strategies based on prior performance; *third*, we show that healthy adult participants learn in a highly sample-efficient manner from previous construction attempts, providing a novel benchmark for AI construction agents to emulate and explain.

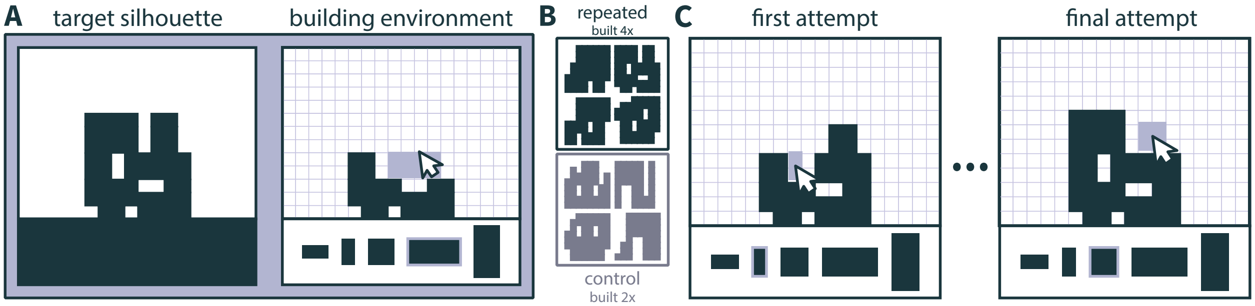

Figure 1: (A) Schematic of task display. The left window contained a target silhouette, and the right contained a building environment with gridlines. (B) For each participant the 8 silhouettes were randomly assigned to conditions, 4 in repeated and 4 in control. (C) Repeated towers were attempted 4 times, interleaved among other towers. Control towers were attempted twice, once at the beginning and once at the end of each session.

Methods

The goal of our experiment was to understand how practice "reverse-engineering" a set of physical structures impacts how people build them over time. To accomplish this, we developed a web-based environment in which people could construct various block towers under simulated rigid-body physics. On each trial, participants aimed to reconstruct a specific target tower in less than 60 seconds using a fixed inventory of rectangular blocks. What made this task challenging is that only silhouettes of these towers were provided, requiring participants to infer which blocks to use, where to place them, and in what order. Over the course of an experimental session, participants built each tower either two or four times, allowing us to measure the degree to which additional task experience led to greater improvement.

Participants

107 U.S.-based participants were recruited from Amazon Mechanical Turk. After accounting for technical issues during data acquisition (i.e., missing data), data from 105 participants were retained (49 female, mean age = 36.8 years). Participants provided informed consent in accordance with the UC San Diego IRB.

Stimuli

To identify a set of block towers of similar complexity, we randomly sampled a large number of stable configurations of 8-16 blocks, then manually selected 8 of these that could be reconstructed in different ways (Fig. 3A).

Task Procedure

On each trial, participants were presented with two adjacent display windows: On the left, a target block tower was presented as a silhouette centered on the floor in a 18x13 rectilinear grid environment (Fig. 1A); on the right, they were provided with an open building environment and a fixed inventory of five types of rectangular blocks that varied in their dimensions (i.e., `1x2`, `2x1`, `2x2`, `2x4`, `4x2`).

Participants' goal was to build a tower that matched the shape of the target silhouette in less than 60 seconds using any combination of the blocks provided. To select a specific block type, they clicked on its image in the block inventory. Then, by hovering the mouse cursor over the building environment, a translucent block would appear, showing where the block would be placed when they clicked the mouse again. Blocks could be placed on any level surface in the building environment (i.e., either the floor or on top of another block). To minimize the intrusion of low-level motor noise in block placement, the locations of each block 'snapped' to grid.

After the placement of each block, participants' towers became subject to gravity, simulated using `Matter.js`. Thus, if their tower was not sufficiently stable, single blocks or even the entire tower could fall over. After 60 seconds had elapsed or if any block fell, the trial immediately ended and participants moved onto the next tower. We truncated trials on which any block fell for two main reasons: first, to ensure that all recorded block placements could in principle form part of a forward plan to build the target silhouette, rather than reflect online corrections for error; and second, to strongly incentivize the production of stable towers. Participants were rewarded for both accuracy and speed: the more accurate their reconstructions, the larger the monetary bonus they received. If they perfectly reconstructed the target silhouette, they could earn an additional bonus for speed.

Design

For each participant, the 8 block towers were randomly split into 2 sets containing 4 towers each: a *repeated* set and a *control* set (Fig. 1 B). Repeated towers were attempted 4 times, randomly interleaved among other towers. Control towers were attempted twice, once near the beginning and once near the end of each session, randomly interleaved among other towers. Thus there were a total of 24 trials in each session, including 8 first attempts and 8 final attempts of each tower. In subsequent comparisons between the first and final attempts on each tower, we combine data from both repeated and control sets. In analyses of fine-grained changes in behavior across successive attempts on the same tower, we restrict our analysis to the control sets.

Results

Change in reconstruction accuracy across attempts

We used the F_1 score as our primary measure of reconstruction accuracy, which reflects the degree to which the shape of participants' reconstruction coincided with the target silhouette, and lies in the range $[0,1]$, where higher scores indicate higher accuracy. It is computed by taking the harmonic mean of the *precision* (i.e., the proportion of participants' reconstruction that coincided with the target silhouette) and *recall* (i.e., the proportion of the target silhouette that coincided with the participants' reconstruction):

$$F_1 = \frac{2}{(recall^{-1} + precision^{-1})}$$

In their first attempt, participants' reconstructions were moderately accurate, suggesting that they were engaged with the task but not at ceiling performance (control: $F_1 = 0.790$, 95% CI: $[0.776, 0.803]$; repeated: $F_1 = 0.800$, 95% CI: $[0.786, 0.814]$; confidence intervals generated via bootstrap resampling). To evaluate changes in reconstruction accuracy over time, we fit a linear mixed-effects model predicting F_1 score from attempt (first, final) and condition (repeated, control) as fixed effects, including random intercepts for participant and tower. We found a main effect of attempt ($b = 0.0759$, $t = 6.99$, $p < 0.001$), showing that participants' reconstruction accuracy reliably improved between their first and final attempts. We found no reliable effect of condition ($b = 0.00803$, $t = 0.737$, $p = 0.461$), and no evidence of an interaction between attempt and condition ($b = 0.0182$, $t = 1.19$, $p = 0.235$), suggesting that these improvements may primarily reflect a combination of task-general and tower-specific learning.

In particular, participants may have learned how to more consistently place blocks that are fully contained within the silhouette, resulting in fewer 'off-by-one' errors. To explore this possibility, we visualized the spatial distribution of block placements by constructing a heatmap of block placements, averaged across participants (Fig. 3). This heatmap suggested that participants did place a greater proportion of blocks outside of target locations in their first attempts than in their final attempts. To evaluate this possibility, we defined the spatial error for a given tower on a given attempt as the root-mean-squared cityblock distance between each location in the heatmap and the edge of the target silhouette (zero if within the silhouette), weighted by the value at each location in the heatmap. We then computed the mean change in spatial error between their first and final attempts, which revealed that participants generally made fewer and less extreme errors in their final attempts than in their first attempts ($m = -0.625$, 95% CI: $[-1.08, -0.209]$, $p = 0.012$).

Change in reconstruction fluency across attempts

In addition to placing blocks more precisely, participants may have also learned to produce more accurate towers by being better able to place more blocks within the time available on each trial. To evaluate this possibility, we modeled the change in the number of blocks used between the first and final attempts using a linear mixed-effects model otherwise identical in structure to that previously used to analyze accuracy, however we excluded trials which were truncated due to blocks falling. This analysis revealed a strong main effect of attempt ($b = 1.19$, $t = 7.41$, $p < 0.001$), showing that participants were able to consistently use more blocks in their final attempt. There was no evidence of an effect of condition ($b = 0.0425$, $t = 0.264$, $p = 0.792$) nor of an interaction between attempt number and condition ($b = 0.167$, $t = 0.735$, $p = 0.463$).

There are at least two potential explanations for how participants were able to place more blocks in their final attempt: *first*, their fluency with the construction task interface may have improved, allowing them to select and place more blocks per unit of time; *second*, they may have been able to recall previously used procedures for building a given tower, and thus required less preparation time to devise an action plan prior to placing their first block. We estimated task fluency by computing the mean time between successive block placements within a single trial. We estimated preparation time by computing the time between trial onset and the placement of the first block. We found that task fluency increased ($b = -1.30$, $t = -9.306$, $p < 0.001$) and preparation time decreased ($b = -2.24$, $t = -8.64$, $p < 0.001$) between first and final attempts, suggesting that participants' improved accuracy may reflect changes in both.

To quantify how quickly participants completed their reconstructions, we measured the amount of time elapsed between the start of each trial and the final block placement on that trial, again omitting trials which were truncated due to falling blocks. In their first attempt, participants used nearly all of the time allotted (control: 51.8s, 95% CI: $[51.1, 52.7]$; repeated: 52.2s, 95% CI: $[51.6, 52.8]$),

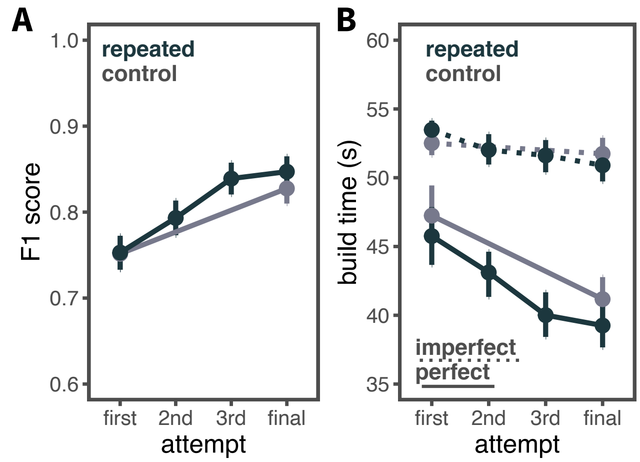

Figure 2: (A) Reconstruction accuracy across build attempts. (B) Build time across attempts, separated by perfect ($F_1 = 1$) and imperfect reconstructions. Error bars represent 95% CI.

Figure 3: (A) 8 target silhouettes used in the experiment. (B,C) Heatmap representations of the spatial distribution of block placements for each tower, for first and final attempts. The intensity of each cell reflects the proportion of participants who had placed a block in that location.

and appeared to use less time to build each tower across attempts (Fig. 2B). To evaluate changes in build time between the first and final attempt, we fit a linear mixed-effects model including attempt (first, final) and condition (repeated, control) as fixed effects, including random intercepts for participant and tower. This analysis revealed a main effect of attempt ($b = -1.92$, $t = -4.25$, $p < 0.001$) but not of condition ($b = -0.704$, $t = -1.80$, $p = 0.0725$). In exploratory analyses, we discovered that 22.4% of all trials contained perfect reconstructions (i.e., $F_1 = 1$) of the target silhouette. When we included an additional binary variable in our regression model indicating whether a trial contained a perfect reconstruction, we discovered that these 'perfect' reconstructions took reliably less time than imperfect reconstructions ($b = -3.81, t = -4.47, p < 0.001$). Moreover, a reliable interaction between attempt number and this binary variable revealed that decreases in build time from first to final attempts were greater for perfect reconstructions ($b = -5.04, t = -5.10, p < 0.001$). Together, these findings suggest that the greatest increases in speed occurred once participants had discovered a way of producing a perfect reconstruction.

Change in reconstruction procedure across attempts

While an increase in speed and decrease in preparation time are consistent with the possibility that participants were reusing successful procedures they had previously used to build each tower, these measures only indirectly bear on this question. To directly assess the extent to which participants reused previously used procedures across attempts, we derived a measure of the similarity between two procedures, which evaluates how similar the individual actions comprising each procedure are.

Each *action* consists of an individual block placement, represented by a 4-vector $[x, y, w, h]$, where $0 \leq x \leq 15$, $0 \leq y \leq 13$ represents the coordinates of the bottom-left corner of the current block and where $(w, h) \in \{(1,2), (2,1), (2,2), (2,4), (4,2)\}$ represent its width and height, respectively. Each procedure consists of the full *sequence* of such actions performed on a given reconstruction attempt. For any *pair* of action sequences, we define the "raw action dissimilarity" as the mean Euclidean distance between corresponding pairs of $[x, y, w, h]$ action vectors (Fig. 4A, light). When two sequences are of different lengths, we evaluate this metric over the first k actions in both, where k represents the length of the shorter sequence. As this measure compares the dissimilarity of sequences on an action-by-action basis, it assumes that when a 'similar' plan is executed again, that similar actions are performed in exactly the same order.

To obtain a measure of similarity between procedures that is robust to differences in the exact order in which actions are performed, we also derived a "transformed" measure of dissimilarity between *sets* of actions. We used the Kuhn-Munkres algorithm to identify the one-to-one mapping between each pair of action sequences minimizing the Euclidean distance between them (Fig. 4A, dark). This "transformed" measure has the advantage of being sensitive to correspondences between similar actions performed in different attempts, even when they were performed in a different order. We fit both raw and transformed action dissimilarities with a linear mixed-effects model including fixed effects for attempt pair, the accuracy of the previous attempt, and the dissimilarity type (raw or transformed), as well as random intercepts for tower and participant. We found that Euclidean distance is negatively related to attempt pair ($b = -0.186, t = -7.40, p < 0.001$; Fig. 4A), suggesting that participants became increasingly consistent in the procedure they used to reconstruct each tower across repeated attempts. We also found that transformed dissimilarities were smaller than raw ones ($b = -0.482$, $t = -2.96$, $p = 0.00315$), suggesting that when participants did reuse actions from prior attempts, they could perform these in a somewhat different order to achieve similar outcomes.

One potential explanation for the increase in internal consistency in procedures across attempts is that

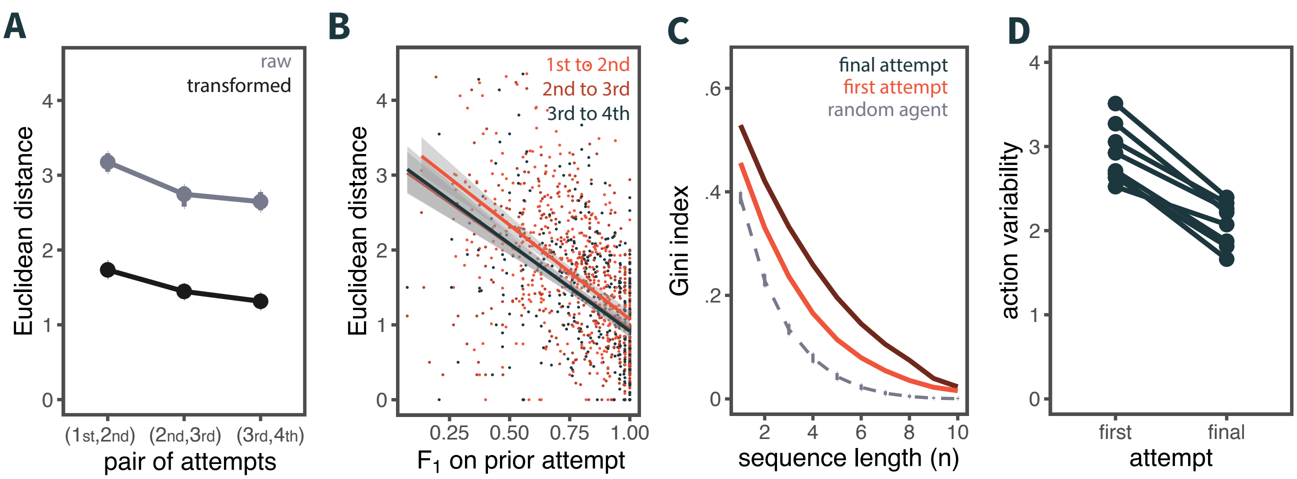

Figure 4: (A) Magnitude of change in action sequences (raw) and sets of actions (transformed) across successive build attempts. (B) Magnitude of change in sets of actions as a function of accuracy (F_1) on previous attempt, for each pair of successive attempts of a given tower. (C) Gini index for n-grams of action sequences in first and final attempts, compared to those of a random-policy agent. Higher Gini index reflects a smaller number of frequently appearing action sequences. (D) Variability between sets of actions performed by different participants on first and final attempts. Each line segment represents a different tower.

participants improved their ability to produce more accurate reconstructions across attempts, and thus did not need to update their procedure as dramatically in later attempts. To the extent that accuracy on prior attempts is related to how much participants alter their procedure in subsequent attempts, we would predict that more successful procedures are more likely to be reused than unsuccessful ones. Consistent with this prediction, we found a strong negative relationship between accuracy on the most recent attempt and how much they changed their procedure ($b = -0.6426$, $t = -4.054$, $p < 0.001$; Fig. 4B), such that participants updated their procedure to a greater extent when their previous attempt was less successful. Taken together, these results suggest that people can reason flexibly about these physical construction problems, making efficient use of prior experience to update their procedures accordingly.

Consistency and variability in procedures across individuals

Our results so far show that participants employ increasingly accurate and internally consistent procedures for reconstructing each tower, raising a natural question concerning the degree to which procedures used by different participants coincide with one another. We visualized the distribution of procedures used by different participants by constructing a map of *trajectories* over intermediate states visited between the start and end of their reconstruction (Fig. 5), where each *state* is defined by the shape of the reconstruction up to that point. Under this definition, reconstructions that are composed of different blocks but share the same shape are treated as occupying the same state.

Even in their first attempt, many participants appeared to traverse the same states when reconstructing each target silhouette (Fig. 5A), hinting at broad consistency in the procedures humans use to perform this task. Additional simulations suggested that at most 2.2% of the total number of possible solutions to each tower were represented in our dataset (i.e., 435 unique trajectories across all towers out of 19,677 discovered so far via random sampling). To rigorously quantify participants' systematic biases toward certain states, we computed the Gini index (G) over the frequency of visits to each state across all participants:

$$G = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right| * \left(2\sum_{i=1}^{n}\sum_{j=1}^{n} x_j\right)^{-1}$$

where n is the number of total states visited and x_i and x_j represents the number of times states i and j were visited, respectively. G is largest when there are a small number of frequently visited states, and lies in the range $[0,1]$. To estimate how strongly human policies concentrate on the same sequences of states at different timescales, we next extracted n-gram representations for all state trajectories, each defined by n successive states, for $1 \leq n \leq 10$, then calculated G_n for each of these n-gram frequency distributions. To provide a baseline, we also constructed a random-policy agent that samples blocks and viable locations (i.e., within silhouette, maintaining stability) with equal probability. We used this random-policy agent to generate a null distribution of 1000 Gini values, each computed from 105 random-policy agents identified by unique random seeds. When comparing the mean observed G for human trajectories to this null distribution, we found that human state trajectories were reliably more concentrated on fewer n-grams than the random-policy agents, across n-grams of all lengths, for both first attempts (Z-score = 21.6) and final ones (mean Z-score = 42.7; Fig. 4C). These results show that a policy of selecting random viable actions is insufficient to explain patterns of human action selection in this task.

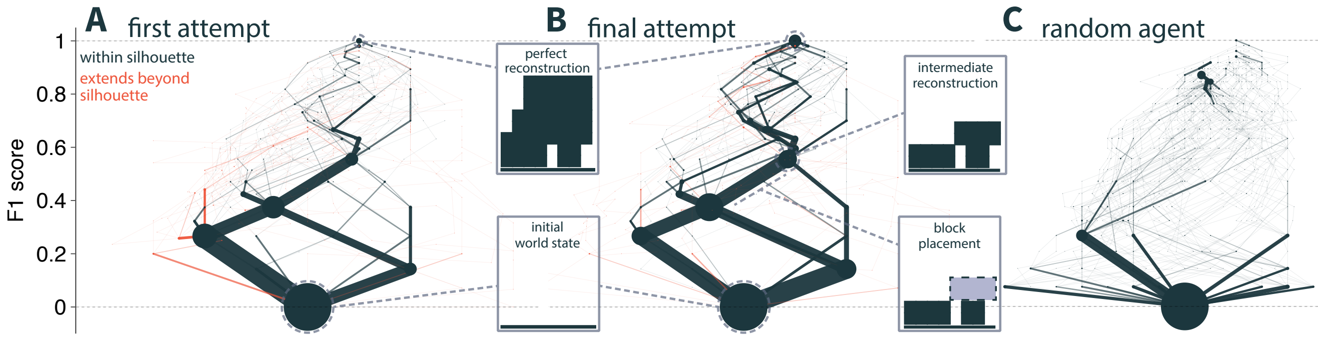

Figure 5: Distribution of state trajectories for first attempts, final attempts, and an artificial agent employing a random action-selection policy to reconstruction an example tower. Each trajectory consists of a sequence of states (nodes) connected by actions (edges), beginning from the initial world state ($F_1 = 0$) and directed upwards toward complete reconstructions ($F_1 = 1$). Node size represents the number of times a state was visited. Edge thickness represents the number of times a state-state transition was traversed.

Insofar as participants are biased to discover similar solutions over time, we may expect the Gini index to grow between the first and final attempts. To evaluate this possibility, we fit human Gini values with a linear mixed-effects model including attempt number, linear and quadratic terms for n, as well as random intercepts for target towers and participants. This analysis revealed a positive effect of attempt number ($b = 0.112$, $t = 6.02$, $p < 0.001$), suggesting that participants tended to converge on a smaller set of procedures across attempts, and this convergence applied to n-grams over action sequences of all lengths (Fig. 4C). Although such convergence is one signature of using similar procedures, the above analysis is insensitive to cases where two participants reconstruct a silhouette by placing the same blocks in the same locations, yet only have first and final world states in common. To address this limitation, we examined the distribution of dissimilarities between the *sets of actions* performed by different participants, and found that the variance of this distribution was smaller on final attempts than on first attempts, for all target towers ($t(7) = 10.603$, $p < 0.001$; Fig. 4D). Taken together, these results indicate that despite the relatively high state-space complexity of this task, people share systematic biases toward similar solutions even on their first attempt, and tend to update their strategies across repeated attempts in similar ways, converging on a smaller set of solutions over time.

Discussion

In this paper, we investigated how people reason about challenging physical construction problems and update their strategies across repeated attempts. We developed a novel task requiring participants to "reverse-engineer" various block towers, which could be reconstructed in many different ways. We found that people can substantially improve the accuracy and speed with which they reconstruct these towers, even after one or a few prior attempts. Moreover, our data indicate that changes in task fluency are insufficient to explain the observed patterns of improvement, as the degree to which participants altered the procedure they used was highly dependent on how successful their prior attempt had been. We also found that different individuals often discovered similar solutions to the same problem, suggestive of shared biases.

A key question raised by this paper concerns the source of the systematicity we see in human strategies. It is possible that shared prior experience with other physical reasoning and planning tasks may play a crucial role, and understanding how humans transfer such broad experiences to new tasks may be critical for developing AI agents that learn as flexibly as humans do. In future work, we plan to evaluate state-of-the-art AI construction agents (Bapst et al., 2019) on a version of the same task with a larger number of towers, using the same metrics. Such evaluations will be critical to expose the extent to which current algorithms for physical reasoning and planning emulate human behavior in this domain, as well as potential gaps for future algorithms to fill.

Another important direction for future work is to investigate how mental simulation and physical experience interact to support effective physical reasoning. Future experiments could manipulate participants' ability to apply physical interventions when solving these problems and measure the consequences on how quickly they learn from multiple attempts using mental simulation alone vs. physical practice. Such studies may help to shed additional light on how people spontaneously decide when to think more and when to act during problem solving (Dasgupta et al., 2018; Kirsh & Maglio, 1994).

In sum, our paper presents novel benchmarks for computational theories of physical reasoning and planning to explain, as well as a strong candidate for task domain that can be fruitfully studied in both cognitive psychology and AI. Such strong alignment between empirical studies of human behavior and algorithms development may lead to both more robust AI and a deeper understanding of human cognition.

Acknowledgments

Thanks to members of the Cognitive Tools Lab at UC San Diego for helpful discussion.

All code and materials available at:
`https://github.com/cogtoolslab/block_construction`

References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2019). The tools challenge: Rapid trial-and-error learning in physical problem solving. *arXiv preprint arXiv:1907.09620*.

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K. L., Kohli, P., Battaglia, P. W., & Hamrick, J. B. (2019). Structured agents for physical construction. *arXiv preprint arXiv:1904.03177*.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bullock, M., & Lütkenhaus, P. (1988). The development of volitional behavior in the toddler years. *Child Development*, 664–674.

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. In *CogSci.*

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., Landau, B., & Shelton, A. L. (2018). Constraints and development in children's block construction. In *CogSci.*

Dasgupta, I., Smith, K. A., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2018). Learning to act by integrating mental simulations and physical experiments. *BioRxiv*, 321497.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., & Battaglia, P. W. (2018). Relational inductive bias for physical construction in humans and machines. *arXiv preprint arXiv:1806.01203*.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? the amount of mental simulation tracks uncertainty in the outcome. In *Cogsci.*

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Jones, J., Hager, G. D., & Khudanpur, S. (2019). Toward computer vision systems that understand real-world assembly processes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 426–434).

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*(4), 513–549.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–131.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological Review*, *119*(1), 120.

Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*(37), 11708–11713.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. In *Cogsci.*