COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13397

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13397

Consistency and Variation in Reasoning About Physical Assembly

William P. McCarthy,[a] David Kirsh,[a] Judith E. Fan[b,c]

[a]*Department of Cognitive Science, University of California San Diego*
[b]*Department of Psychology, University of California San Diego*
[c]*Department of Psychology, Stanford University*

Received 17 January 2023; received in revised form 27 October 2023; accepted 6 December 2023

Abstract

The ability to reason about how things were made is a pervasive aspect of how humans make sense of physical objects. Such reasoning is useful for a range of everyday tasks, from assembling a piece of furniture to making a sandwich and knitting a sweater. What enables people to reason in this way even about novel objects, and how do people draw upon prior experience with an object to continually refine their understanding of how to create it? To explore these questions, we developed a virtual task environment to investigate how people come up with step-by-step procedures for recreating block towers whose composition was not readily apparent, and analyzed how the procedures they used to build them changed across repeated attempts. Specifically, participants ($N = 105$) viewed 2D silhouettes of eight unique block towers in a virtual environment simulating rigid-body physics, and aimed to reconstruct each one in less than 60 s. We found that people built each tower more accurately and quickly across repeated attempts, and that this improvement reflected both group-level convergence upon a tiny fraction of all possible viable procedures, as well as error-dependent updating across successive attempts by the same individual. Taken together, our study presents a scalable approach to measuring consistency and variation in how people infer solutions to physical assembly problems.

Keywords: Planning; Spatial reasoning; Intuitive physics; Construction; Action

Humans have populated much of the world with physical artifacts of their own design, from sand castles to skyscrapers. Taken together, these structures exemplify the human capacity to interact with the physical world in creative, yet goal-directed ways. This creative capacity also manifests in many everyday tasks, from assembling a piece of furniture to making a sandwich and knitting a sweater. In these scenarios, people rely upon their ability to not

Correspondence should be sent to Judith E. Fan, Department of Psychology, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305, USA. E-mail: jefan@stanford.edu

only judge the static properties of objects (e.g., their size, shape, weight), but also to infer the process by which objects are made (e.g., the parts they consist of and how to arrange them). What cognitive mechanisms enable people to engage in such reasoning about complex objects, and how do people draw upon prior experience with an object to continually refine their understanding of how to create it?

Perhaps the most basic requirement is a general-purpose and intuitive understanding of how material objects interact in the physical world, a suite of abilities known as intuitive physics (McCloskey, 1983). That is, even without performing formal calculations, people can make reasonably accurate predictions about how objects will behave in a variety of settings (Kubricht, Holyoak, & Lu, 2017; Smith, Battaglia, & Vul, 2018). A prominent proposal argues that generating these predictions relies on mental simulation, perhaps reflecting a noisy approximation to real-world physical dynamics (Battaglia, Hamrick, & Tenenbaum, 2013; Hegarty, 2004; Hamrick, Smith, Griffiths, & Vul, 2015; Smith & Vul, 2013; Sanborn, Mansinghka, & Griffiths, 2013; Schwartz & Black, 1999). Recent work has explored the role that simulation plays when people plan single interventions on physical scenes—for example, joining two blocks together to stabilize a block tower (Hamrick et al., 2018) or causing an object to move into a target zone (Allen, Smith, & Tenenbaum, 2020; Dasgupta, Smith, Schulz, Tenenbaum, & Gershman, 2018). However, the role of physically grounded mental simulation has yet to be fully explored in the context of the multi-step action sequences required to assemble a complex object (Kurth-Nelson et al., 2023; Kirsh, 1995; Schwartenbeck et al., 2021). This gap in knowledge at least in part reflects the methodological challenges posed by measuring behaviors as open-ended as physical assembly while maintaining a sufficient degree of experimental control (Cortesa et al., 2017, 2018; Wolfgang, Stannard, & Jones, 2001).

Recent advances in the study of multi-step planning and decision-making in other settings suggest promising ways forward (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Huys et al., 2015; Solway & Botvinick, 2015, 2012). To the degree that these "grid-world" environments used often in this work sacrifice physical realism, they do so in favor of empirical and formal tractability (Van Opheusden, Galbiati, Bnaya, Li, & Ma, 2017; van Opheusden & Ma, 2019; van Opheusden et al., 2023). Nevertheless, as the state space grows, the computational cost of conducting thorough mental simulations over the full set of possibilities becomes prohibitive (Callaway et al., 2018; Hamrick et al., 2015; Huys et al., 2015; Solway & Botvinick, 2015, 2012). Prior work has found evidence that humans use a variety of strategies to reduce the cost of planning, such as pruning the search space (i.e., circumventing expensive but irrelevant action sequences; Huys et al., 2012) and learning procedural abstractions to generate hierarchically organized plans (Botvinick & Weinstein, 2014; Dezfouli & Balleine, 2013; Éltető & Dayan, 2023; Huys et al., 2015; Xia & Collins, 2020). However, it remains unknown which, if any, of these strategies are ones that humans use when attempting to solve physical assembly problems, in which transitions between states are governed by physical constraints (e.g., stability, friction) rather than arbitrary rules (Daw et al., 2011). A valuable step toward bridging this gap would be the development of experimental methods for exploring human assembly behavior in task environments with a greater degree of physical realism than those commonly used to probe multi-step decision-making.

An additional benefit of developing such methods would be the opportunity to investigate the impact of experience on assembly behavior, building on a long tradition of work investigating changes in problem-solving accompanying the acquisition of expertise (Chase & Simon, 1973; Campitelli & Gobet, 2004; Sheridan & Reingold, 2017; Van Harreveld, Wagenmakers, & Van Der Maas, 2007). For example, experience might be linked to changes in both how people encode state information and how they search over the space of possible solutions. Classic and contemporary work using board games suggests that experts display both a pronounced ability to plan further ahead in games than novices and to mentally represent the configuration of game pieces in visual memory with higher fidelity (Chase & Simon, 1973; Gobet & Simon, 1998; Sheridan & Reingold, 2017; van Opheusden et al., 2023). Moreover, prior work that used video games to impose substantial demands on rapid spatial reasoning (e.g., Tetris) has found that experience might also improve the fluency with which participants explore alternative states and determine the value of potential actions (Maglio & Kirsh, 1996). In principle, these experience-dependent changes might also apply to the domain of physical assembly, which would suggest that the underlying learning mechanisms generalize beyond the problem contexts in which they were initially proposed. On the other hand, it might be that there are important differences between reasoning domains: for example, problem-solving experience might have a stronger impact on how state information is encoded in less physically realistic game environments, such as board games, but a more modest impact in physical settings, where the mechanisms for encoding physical state are more stable across the lifespan (Baillargeon, 1995; Spelke & Kinzler, 2007).

Here, we introduce a task paradigm for investigating how people reason about physical assembly in a virtual environment that is simple enough to provide a high degree of experimental control and formal tractability, but expressive enough to engage multi-step planning and understanding of core physical concepts (e.g., stability, mass, and friction). We report our findings from an exploratory study in which participants aimed to construct a series of 2D block towers from a set of rectangular blocks of varying sizes. We restricted the set of possible actions to placements of a fixed set of parts, enabling straightforward comparison of building procedures across participants. We further investigated how practice reconstructing a tower impacts the procedures participants subsequently used to build that tower across repeated attempts. Our approach takes inspiration from recent studies in which participants were asked to build copies of actual LEGO structures from LEGO bricks (Cortesa et al., 2018; Shelton et al., 2022). Findings from this line of work suggest that people converged upon shared strategies for building these LEGO structures layer by layer, consistent with a bias toward shared layer-wise subgoals that also corresponded to physical subunits of the structures themselves (Shelton et al., 2022).

As in this prior work, we go beyond simple measures of assembly performance to characterize the action-by-action procedures people used to build each structure. However, our methodological approach differs in three key ways: First, because the current study aims to investigate the role of experience in assembly behavior, here we ask participants to build the same structures multiple times, allowing us to ask how practice influences the procedures that people use. Second, in order to put greater pressure on participants' ability to reason about how an object could be made, we presented participants with *silhouettes* of block towers

that could in fact be built in many different ways. Third, in order to support high-throughput measurement of these open-ended behaviors, we developed a virtual assembly environment embedded in a web application to enable the concurrent participation of many individuals.

1. Method

The goal of our experiment was to investigate how people's strategies for solving physical reasoning tasks shift as they gain experience. To achieve this goal, we developed a web-based environment in which people could construct various block towers under simulated rigid-body physics. To provide participants with a specific goal, we considered the space of *physical assembly* tasks—namely, those in which people must create an exact replica of a target structure given the set of components used to construct it. However, such straightforward assembly tasks typically permit only a small range of solutions and can be solved using a simple strategy of copying block for block (Cortesa et al., 2017). To explore how strategies change with experience, we needed a task that permitted a large range of solutions. Therefore, rather than display target towers as a configuration of blocks that could be copied, we showed participants *silhouettes* of target towers and asked them to create *any configuration of blocks that matched the silhouette* shown. This required participants to infer which blocks to use, where to place them, and in what order. On each trial, participants aimed to reconstruct a target tower in less than 60 s using a fixed inventory of rectangular blocks. Over the course of an experimental session, participants built each tower either two or four times, allowing us to assess whether additional practice reconstructing a specific tower led to greater improvement than general practice with the task.

1.1. Participants

Based on data from pilot studies, we estimated that between 100 and 150 participants would be sufficient to obtain reasonably precise estimates of our measures of consistency and variability. In the end, we successfully recruited 107 U.S.-based participants from Amazon Mechanical Turk. After accounting for technical issues during data acquisition (i.e., missing data), data from 105 participants were retained (49 females, mean age = 36.8 years). Participants provided informed consent in accordance with the University of California, San Diego Institutional Review Board.

1.2. Stimuli

To identify a set of block towers that were nontrivial to reconstruct, we randomly sampled a large number of stable configurations of 8–16 blocks, then manually selected eight of these that could be reconstructed in many different ways (Fig. 1B). We started with an inventory of five types of rectangular blocks that varied in their dimensions (i.e., `1x2`, `2x1`, `2x2`, `2x4`, `4x2`). To generate configurations of blocks, we partially filled an 8x8 rectilinear grid, bottom to top, by sampling random blocks in random x-locations, then randomly selected several blocks to be removed. We simulated the construction of each tower in a physics engine (Pybox2d),

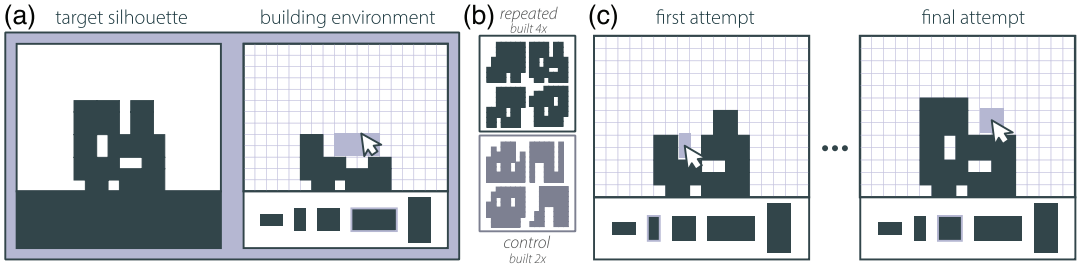

Fig. 1. (A) Schematic of task display. The left window contained a target silhouette, and the right contained a building environment with gridlines. (B) For each participant, the eight silhouettes were randomly assigned to conditions, four in repeated and four in control. (C) Repeated towers were attempted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

rejecting any tower that was unstable at any point during the construction process. To select towers that required planning ahead, we manually identified eight configurations that included holes and/or overhanging blocks, and verified that these towers could be reconstructed in many different ways (59–7128 minimum unique solutions, *mean* = 2418).

1.3. Design

To more thoroughly characterize the effects of practice on physical construction ability, we sought to distinguish improvement resulting from general task experience from improvement resulting from practice reconstructing a specific tower. For each participant, we, therefore, randomly split the eight block towers into two sets containing four towers each: a *control* set and a *repeated* set (Fig. 1B). Participants reconstructed towers over four consecutive rounds. In the *first* (first) and *final* (fourth) rounds, participants reconstructed all eight towers in a randomized order. In the middle two rounds (second and third), participants reconstructed only the four *repeated* towers, also in a randomized order. Thus, there were a total of 24 trials in each session: eight *first* attempts, two rounds of four *repeated* attempts, and eight *final* attempts. In subsequent comparisons between the first and final attempts on each tower, we combine data from both the repeated sets (built four times) and control sets (built two times). In analyses of fine-grained changes in behavior across successive attempts on the same tower, we restrict our analysis to the repeated sets.

1.4. 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 empty building environment and the inventory of blocks that was used to generate the towers.

Participants' goal was to build a tower that matched the shape of the target silhouette in less than 60 s 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 again. Blocks could be placed on any level surface in the building environment (i.e., either the floor or at least partially supported by another block). To minimize the intrusion of low-level motor noise in block placement, the location of each block "snapped" to a visible 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 s 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 participants perfectly reconstructed the target silhouette, they could earn an additional bonus for speed.

1.5. Statistical analysis procedure

Our primary statistical approach involved fitting linear mixed-effects models mirroring, as close as possible, the structure of the experimental design. This included fixed effects for round and condition, as well as their interaction, and random intercepts for participant and tower. We then compared this full model to a series of nested models that had some of the predictors removed, typically starting with the interaction term, then the effect of condition. To select a model, we calculated the Akaike Information Criterion (AIC) for each model, selecting the most complex model for which AIC substantially dropped relative to the subsequent simpler model. Full parameter estimates for selected models are reported in the Supplementary Materials. For statistics outside of the models, we report confidence intervals generated using bootstrap resampling over 1000 iterations. In each bootstrap iteration, we resampled participants with replacement from the entire sample, including all data from each participant every time they were sampled.

2. Results

2.1. Change in reconstruction accuracy across attempts

We first needed a measure of reconstruction accuracy that tracked how well the towers participants built matched the silhouette they were attempting to reconstruct. Reconstructions are accurate insofar as they coincide with the same region as the target silhouette, while not extending beyond it. We, therefore, selected a metric that takes into account both *recall* (i.e., the proportion of the target silhouette that coincided with the participants' reconstruction) and *precision* (i.e., the proportion of participants' reconstruction that coincided with the target silhouette). As stable towers existed in a gridworld, we could compute precision and recall directly by comparing the bitmaps of squares occupied by the target silhouette and reconstruction. The F_1 score takes the harmonic mean of these values to provide a measure

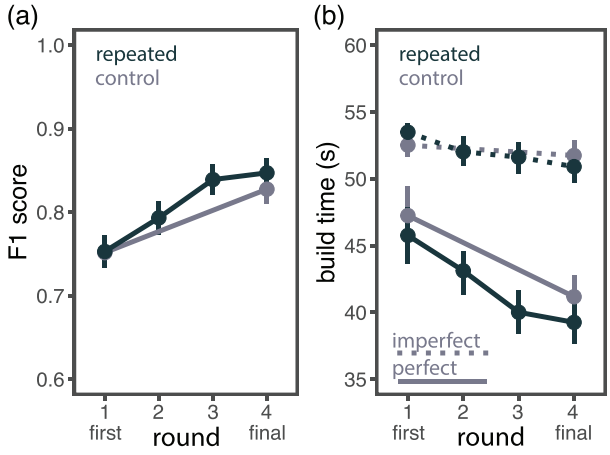

Fig. 2. (A) Reconstruction accuracy across all four rounds. Control towers were built in only the first and final rounds. (B) Build time across attempts, separated by perfect ($F_1 = 1$) and imperfect reconstructions. Error bars represent 95% CI.

that lies in the range [0,1] and reflects the degree to which the participants' reconstruction coincided with the target silhouette:

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

In their first attempts, 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]). 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 (Table S1). We found a main effect of attempt ($b = 0.0759$, $t = 6.99$, $p < .001$), showing that participants' reconstruction accuracy reliably improved between their first and final attempts (Fig. 2A). We found no reliable effect of condition ($b = 0.00803$, $t = 0.737$, $p = .461$), and no evidence of an interaction between attempt and condition ($b = 0.0182$, $t = 1.19$, $p = .235$), suggesting that these improvements were at least in part explained by general effects of task practice.

In particular, participants may have learned how to more consistently place blocks that are fully contained within the silhouettes, 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

Fig. 3. (A) Eight 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 placed a block in that location.

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: $[-1.08, -0.209]$, $p = .012$).

2.2. Change in reconstruction fluency across attempts

In addition to placing blocks more precisely, participants may have also produced more accurate reconstructions by improving their ability 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 (Tables S2 and S3). This analysis revealed a strong main effect of attempt ($b = 1.19$, $t = 7.41$, $p < .001$), showing that participants were able to consistently use more blocks in their final attempt. There was no evidence of an effect of condition ($b = 0.0425$, $t = 0.264$, $p = .792$) nor of an interaction between attempt number and condition ($b = 0.167$, $t = 0.735$, $p = .463$).

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 ($b = -1.34$, $t = -13.548$, $p < .001$; Table S4) and preparation time decreased ($b = -2.24$, $t = -8.64$, $p < .001$; Table S5) between first and final attempts, suggesting that participants' improved accuracy may reflect changes in both.

To measure 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 attempts, participants used nearly all of the time allotted (control: 51.8*s*, 95% CI: [51.1, 52.7]; repeated: 52.2*s*, 95% CI: [51.6,52.8]), 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 (Table S6). This analysis revealed a main effect of attempt ($b = -1.92$, $t = -4.25$, $p < .001$) but not of condition ($b = -0.704$, $t = -1.80$, $p = .0725$). 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 ($b = -3.81$, $t = -4.47$, $p < .001$). 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 ($b = -5.04$, $t = -5.10$, $p < .001$). Together, these findings suggest that the greatest increases in speed occurred once participants had discovered a way of producing a perfect reconstruction.

2.3. Change in reconstruction procedures across attempts

Having established that participants build more accurately and quickly across successive attempts, we then investigated the changes to participants' construction procedures that underlie this improved performance. An increase in speed and decrease in preparation time are consistent with the possibility that participants reused previous procedures to successfully reconstruct each tower; however, these holistic measures only indirectly bear on this question. We, therefore, derived two measures of similarity between the actions performed across different building attempts (Fig. 4A).

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. We define the "sequence dissimilarity" between any *pair* of action sequences as the mean Euclidean distance between corresponding pairs of $[x, y, w, h]$ action vectors (Fig. 4A, top). 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. This *sequence* measure compares the dissimilarity of procedures on an action-by-action basis, and hence assumes that when "similar" actions are executed, they are performed in exactly the same order. However, we might also consider procedures to be "similar" when they involve similar shaped blocks placed in similar locations, even when the order of these block placements varies. To obtain a measure of similarity between procedures that is robust to differences in the order in which actions are performed, we also derived a measure of dissimilarity between the *sets* of actions performed, using the Kuhn–Munkres algorithm to identify the one-to-one mapping between actions from each attempt that minimizes the mean Euclidean distance between them (Fig. 4A, bottom). This "set dissimilarity" measure has the

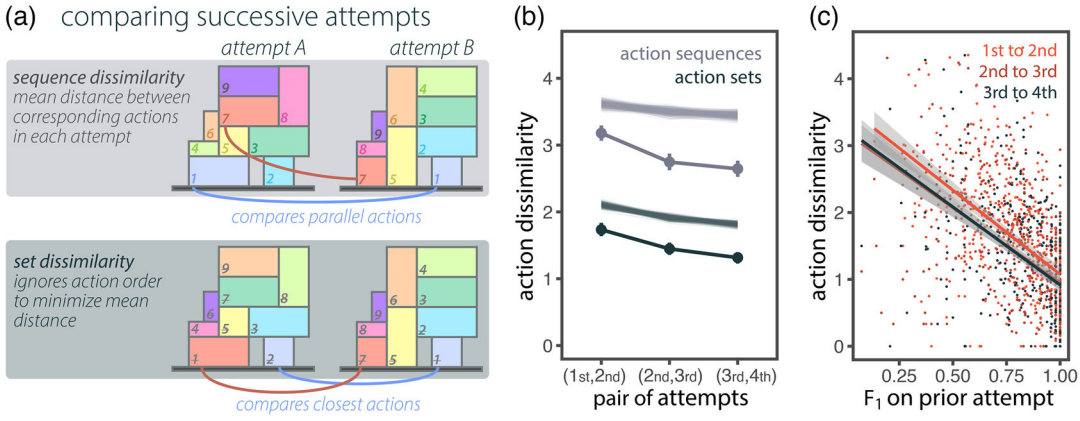

Fig. 4. (A) Example comparison between building procedures on successive attempts. Numbers on blocks indicate the order in which they were placed. Sequence dissimilarity (top) compares building procedures on an action-by-action basis (i.e., block n to block n). Actions involving different sized blocks placed further apart are judged as more distant. Set dissimilarity (bottom) minimizes the mean distance between actions by ignoring order and pairing similar actions together. (B) Magnitude of change in sequences of actions (gray) and sets of actions (dark green) across successive build attempts. Shaded area represents baseline distributions. (C) 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.

advantage of being sensitive to correspondences between similar actions performed in different attempts, even when they were performed in a different order.

We first sought to determine whether participants reused aspects of their own prior attempts when reconstructing towers. We calculated the sequence and set dissimilarities between individual participants' consecutive attempts at each tower (Fig. 4B, solid). To estimate the expected dissimilarity between attempts regardless of individual, we generated a baseline distribution of dissimilarity values comparing each participant's second, third, and fourth attempts at a each tower with prior attempts (i.e., first, second, and third) from a different, randomly sampled participant. We repeated this process 1000 times, independently and randomly pairing participants for each tower (Fig. 4B, shaded). We found that participants' procedures were more similar to their own prior attempts than to other participants' ($p < .001$ for each pair of consecutive repetitions, for both *sequence* and *set* dissimilarity), suggesting that participants did reuse aspects of their own prior solutions to reconstruct each tower.

To assess whether participants used increasingly similar procedures across consecutive attempts, we fit both *sequence* and *set* action dissimilarities with a linear mixed-effects model including fixed effects for attempt pair, the accuracy of the previous attempt, and the dissimilarity type (*sequence* or *set*), as well as random intercepts for tower and participant (Table S7). We found that attempt pair was negatively related to dissimilarity for both dissimilarity measures ($b = -0.186, t = -7.40, p < .001$; Fig. 4B), suggesting that participants became increasingly consistent in the procedures they used to reconstruct each tower across repeated attempts. In other words, actions in participants' later attempts (i.e., attempts 3 and

4) were more similar to each other than actions in earlier attempts (i.e., 1 and 2). As this result holds for set as well as sequence dissimilarity, it suggests a genuine increase in the consistency between the actions taken by participants, regardless of the specific order in which they performed.

A potential explanation for this convergence in procedures is that, as participants uncover increasingly successful procedures for recreating a tower, they may be less likely to dramatically change their strategy in later attempts. To the extent that accuracy on prior attempts is related to how much participants alter their procedure in subsequent attempts, we expect more successful procedures to be 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 ($b = -0.6426$, $t = -4.054$, $p < .001$; Fig. 4C), 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 make efficient use of prior experience to update their approach to solving assembly problems accordingly.

2.4. 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. While the analyses above suggest some variation in the actions that participants performed, they do not reveal whether participants were biased toward a small set of solutions for each tower, or whether they instead discovered a wide variety of completely different solutions. We therefore visualized the distribution of procedures used by all 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 (silhouette) are treated as occupying the same state, but are reached by taking distinct trajectories.

Even on their first attempts, many participants appeared to traverse the same states when reconstructing each target silhouette (Fig. 6), hinting at broad consistency in the procedures people 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 via random sampling). To estimate how strongly participants were biased toward the same of subsequences, we computed the Gini index (G) over the number of traversals of each sequence of states 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 states and x_i and x_j represents the number of times states i and j were visited, respectively. G can be thought of as the average difference in the number of

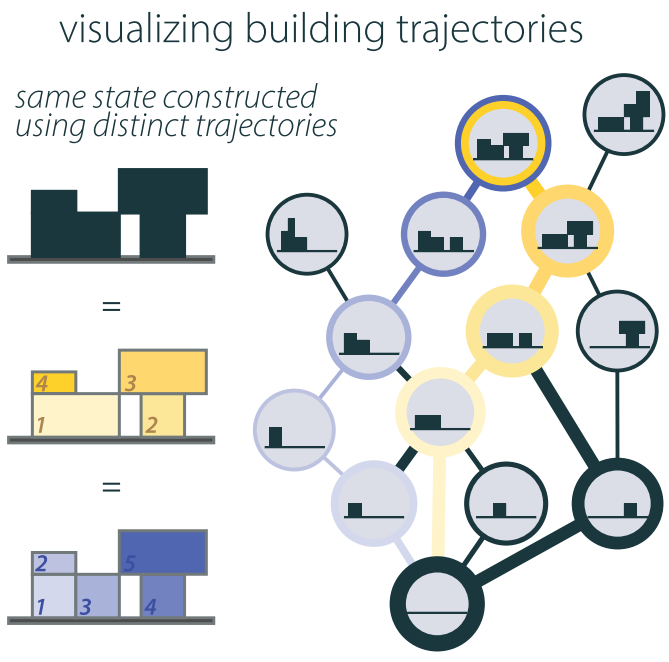

Fig. 5. To visualize the set of trajectories taken by participants, we constructed graphs of the intermediate states visited during a reconstruction attempt. Larger nodes indicate a greater number of participants constructing that intermediate state, and thicker edges indicate a greater number of participants who transitioned between two world states with a single block placement. Intermediate states are defined by their outline shape and are independent of the underlying blocks used to create them. Two distinct trajectories leading to the same state are highlighted in blue and yellow.

times each subsequence was traversed, normalized by the total number of sequences of that length (summed twice to account for differences in both directions) to lie in the range [0,1]. It is largest when there are a small number of frequently traversed subsequences and smallest when all subsequences were traversed an equal number of times.

To estimate how strongly human procedures 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 (Fig. 7A). 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. 7B). These results show that a policy of selecting random viable actions is insufficient to explain patterns of human action selection in this task.

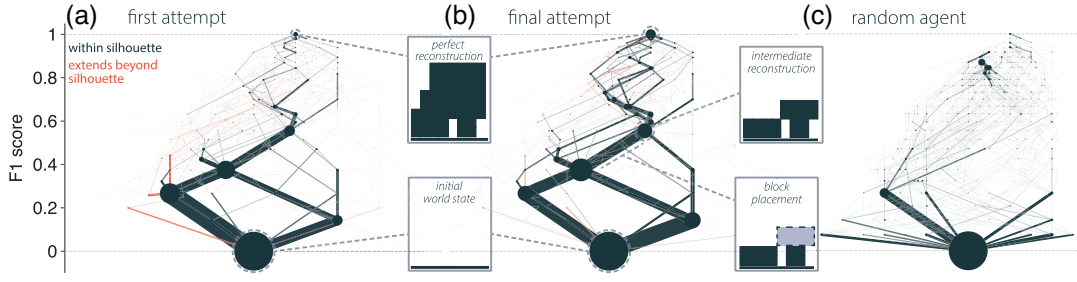

Fig. 6. Distribution of state trajectories for first attempts (A), final attempts (B), and an artificial agent (C) employing a random action-selection policy to reconstruct an example tower. Each trajectory consists of a sequence of states (nodes) connected by actions (edges), beginning from the initial world state ($F_1 = 0$) and directed upward 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.

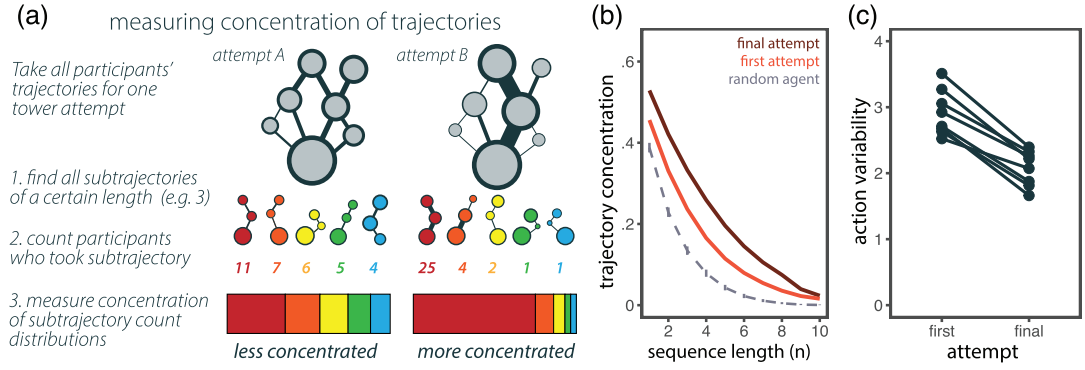

Fig. 7. (A) To estimate the degree of bias toward certain trajectories, we extracted all subsequences of states of a certain length and measured how concentrated construction behavior was on a small number of sequences. (B) Gini index for n-grams of action sequences in first and final attempts, compared to those of a random-policy agent. Higher Gini index reflects a smaller number of frequently appearing action sequences. (C) Variability between sets of actions performed by different participants in first and final attempts. Each line segment represents a different tower.

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 (Table S8). This analysis revealed a positive effect of attempt number ($b = 0.112$, $t = 6.02$, $p < .001$), suggesting that participants converged on a smaller set of procedures across attempts, and this convergence applied to n-grams over action sequences of all lengths (Fig. 7B).

Such convergence toward a smaller number of state sequences is consistent with the possibility that out of all the possible ways of "carving up" each tower into parts, participants had discovered similar ways of doing so. However, an alternative possibility is that this convergence was primarily driven by improvement in participants' ability to build each tower

more accurately, with more accurate reconstructions being inherently more similar to one another than less accurate ones. To distinguish these possibilities, we repeated the previous analysis but only on trials where participants perfectly reconstructed the target tower. We found that Gini values still increased from the first to the final attempt ($b = 0.175$, $t = 5.68$, $p < .001$; Table S9), confirming that convergence in trajectories was not simply a consequence of more accurate reconstructions, but also reflected more consistent ways of reconstructing each tower.

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 place these blocks in a different order. 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 in first attempts, for all target towers ($t(7) = 10.603$, $p < .001$; Fig. 7C). Taken together, these results indicate that despite the relatively high state-space complexity of this task, people share systematic biases toward similar solutions even in their first attempts, and tend to update their strategies across repeated attempts in similar ways, converging on a more similar set of solutions over time.

3. Discussion

In this paper, we investigated how people reason about physical assembly problems and update their approach to solving them over time. Specifically, we developed a web-based environment where participants aimed to reconstruct a set of 2D block towers, and measured how accurately and quickly they could do so across successive attempts at building each tower. We found that participants achieved strong performance even on their first attempts and improved substantially with additional practice. Moreover, our findings suggest that low-level changes in motor fluency were insufficient to fully explain this improvement. Instead, improvement was driven by genuine changes in the decisions made by participants about how to build each tower, with participants updating their procedures to a greater degree when their prior attempt had been less successful. In addition, although there were many possible ways of reconstructing each tower, we found that the procedures participants used to initially construct these towers were strikingly consistent across individuals. Moreover, participants converged on increasingly similar procedures across attempts, suggesting shared biases toward similar approaches to solving these assembly problems.

What accounts for the consistency in participants' assembly behavior, especially given that for some towers there were as many as several thousand valid ways to reconstruct them? One possibility is that shared mechanisms for physical understanding lead to similar mental simulations in planning (Proffitt & Gilden, 1989; Spelke & Kinzler, 2007; Smith & Vul, 2013). Alternatively, the consistency we see in people's initial strategies might have been driven more by participants' use of simple rules and heuristics (e.g., to build layer by layer; Shelton et al., 2022). While our random agent baseline simulates the minimum level of consistency expected under the physical constraints of the task, alternative algorithms could be used to evaluate

specific hypotheses concerning the source of homogeneity in participants' solutions. For example, one possibility is that people build "greedily," initially prioritizing larger blocks that cover more of the silhouette, but gradually updating the value of these initial actions in light of whether their reconstruction was ultimately successful (Barto, Bradtke, & Singh, 1995).

Another possibility is that the consistency we observed reflects a tendency for participants to decompose these towers into visual parts in similar ways, and that these parts form the basis for how they then build these towers. Supposing the perceptual organization of complex shapes does constrain how people approach assembly problems, what characterizes the parts that people favor? Identifying the parts that people use to parse visual objects has long been a central target for classical theories of perceptual organization, which have emphasized spatial and shape-based cues to parthood (Hoffman & Richards, 1984; Palmer, 1977; Schyns & Murphy, 1994; Tversky & Hemenway, 1984; Wertheimer, 1923). Building on this tradition, a related notion is that the parts people use to parse a complex visual object are those that are easy to identify and remember (e.g., according to Gestalt or other principles), and can be used to form more compressed representations of other, similar objects (Biederman, 1987). In other words, people confronting an assembly problem may invoke a mental library containing these useful part concepts to imagine a compact motor program that could be executed to generate the target object from those parts (Ellis et al., 2020; Lake, Salakhutdinov, & Tenenbaum, 2015; Tian, Ellis, Kryven, & Tenenbaum, 2020; Wong et al., 2022). On this view, the value placed on parts that appear in different objects suggests a route by which prior experience with specific objects guides the kinds of representational primitives that emerge. Future studies could test these ideas by manipulating the prevalence of different parts in the set of objects people are asked to build, and measuring the impact of exposure to these parts on the assembly procedures they converge upon.

A major focus of the current study was on how practice building an object affects a person's approach to building it later. To what degree does such building experience not only affect how people build it later, but also its underlying mental representation, such that they perceive or remember it differently? This question has been explored in prior work investigating other visual production modalities, such as drawing (Fan, Yamins, & Turk-Browne, 2018; Wammes, Meade, & Fernandes, 2016) and handwriting (James, 2010, 2017). For example, in one recent study, participants who repeatedly produced drawings of similar objects (e.g., beds and chairs) were better able to discriminate them in a subsequent categorization task, relative to control objects that were not repeatedly drawn (Fan et al., 2018). Moreover, this drawing practice was accompanied by changes in patterns of connectivity between visual and parietal cortex, suggesting a potential neural substrate by which experience improves people's ability to transform the contents of a perceptual representation into representational actions (Fan et al., 2020). A promising direction for future work is to test the degree to which practice plays a similar role in the context of physical assembly, thus providing a measure of how strongly these production-driven learning consequences generalize beyond the domain of drawing and handwriting (Schwartenbeck et al., 2021). Insofar as they do, such findings would lend support to the notion that, at least in some contexts, how people internally represent an object is characterized by a fundamental duality—its correspondence to a static entity with certain perceptual properties, but also to a generative process that gives rise to it (Fan et al., 2018;

Fernandes, Wammes, & Meade, 2018; James, 2017; Lake et al., 2015). Regardless, the results of such studies will be invaluable for advancing our mechanistic understanding of how active and constructive behaviors relate to learning more generally (Chi & Wylie, 2014).

One limitation of our study as it pertains to real-world physical assembly is the focus on building 2D block towers in a virtual environment. While our virtual building environment retained some key aspects of building objects in the physical world, including the relevance of gravity and friction for reasoning about physical stability, there were many other aspects that were not retained in this environment, such as depth information and the biomechanical details governing how a person would actually need to grip a 3D object in order to maneuver it into place. Future work exploring physical assembly could overcome these drawbacks by using recently developed 3D virtual environments to investigate more realistic forms of interaction (Gan et al., 2020, 2021) and could further connect with research in robotics exploring how data from sight and touch might be integrated in order to plan complex actions in the real world (Erdogan, Yildrim, & Jacobs, 2014; Fazeli et al., 2019; Mason, 2018). The generality and scope of our findings might also be extended by using a larger and more diverse set of towers, which would support investigation of the relationship between various properties of these towers (e.g., size, presence of "holes") and how difficult they are to build. Moreover, in order to test the specific hypotheses raised earlier concerning the use of hierarchical representations during physical assembly, it will be advantageous to use more complex objects in future studies that more clearly support hierarchical decomposition (McCarthy, Hawkins, Wang, Holdaway, & Fan, 2021; Wong et al., 2022). Another limitation of the current study is the focus on accurate reconstruction of existing physical structures, rather than reasoning about how to build new ones that satisfy more abstract design criteria, such as the need to provide "shelter" for another object (Bapst et al., 2019). Expanding the suite of physical assembly tasks to include these more open-ended design challenges may provide more direct insight into how humans deploy their general-purpose understanding of how the physical world works to create new things.

In sum, our paper introduces and validates an approach for investigating how people learn how to solve physical assembly problems, providing a window into how physical reasoning and planning interact to achieve specific behavioral goals. Such tools are especially valuable for advancing mechanistic theories of cognition because they support large-scale measurement of complex human behaviors and the evaluation of candidate cognitive models within the same environment. We hope that our findings will inspire further development of mechanistic models that display these and other richly complex behaviors, and direct comparison of these models' behavior to that of humans. In the long run, strong alignment between empirical studies of human and model behavior may lead to both more robust and human-like artificial intelligence, as well as a deeper understanding of human cognition.

Acknowledgments

Thanks to the members of the Cognitive Tools Lab at the University of California, San Diego for helpful discussion. This work was supported by an NSF CAREER Award #2047191

to J.E.F. A subset of these findings were presented as part of the Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.

Data and code availability

All experimental materials, data, and analysis code are publicly available in our GitHub repository: https://github.com/cogtoolslab/block_construction.

Conflict of Interest Statement

The authors report no conflict of interest.

References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310.

Baillargeon, R. (1995). Physical reasoning in infancy. *Cognitive Neurosciences*, 181–204.

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*.

Barto, A. G., Bradtke, S. J., & Singh, S. P. (1995). Learning to act using real-time dynamic programming. *Artificial Intelligence*, *72*(1–2), 81–138.

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.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115.

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130480.

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

Campitelli, G., & Gobet, F. (2004). Adaptive expert decision making: Skilled chess players search more and deeper. *ICGA Journal*, *27*(4), 209–216.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

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*.

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., Shelton, A. L., & Landau, B. (2017). Characterizing spatial construction processes: Toward computational tools to understand cognition. 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.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology*, *9*(12), e1003364.

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning. *arXiv preprint arXiv:2006.08381*.

Éltető, N., & Dayan, P. (2023). Habits of mind: Reusing action sequences for efficient planning. *arXiv preprint arXiv:2306.05298*.

Erdogan, G., Yildrim, I., & Jacobs, R. A. (2014). Transfer of object shape knowledge across visual and haptic modalities. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Fan, J. E., Wammes, J. D., Gunn, J. B., Yamins, D. L., Norman, K. A., & Turk-Browne, N. B. (2020). Relating visual production and recognition of objects in human visual cortex. *Journal of Neuroscience*, *40*(8), 1710–1721.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*(8), 2670–2698.

Fazeli, N., Oller, M., Wu, J., Wu, Z., Tenenbaum, J. B., & Rodriguez, A. (2019). See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*, *4*(26), eaav3123.

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science*, *27*(5), 302–308.

Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., Kim, K., Wang, E., Lingelbach, M., Curtis, A., Bear, D., Gutfreund, D., Cox, D., Torralba, A., DiCarlo, J., Tenenbaum, J., McDermott, J., & Yamins, D. (2020). ThreeDworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

Gan, C., Zhou, S., Schwartz, J., Alter, S., Bhandwaldar, A., Gutfreund, D., Yamins, D. L., DiCarlo, J. J., McDermott, J., Torralba, A., & Tenebaum, J. (2021). The threeDworld transport challenge: A visually guided task-and-motion planning benchmark for physically realistic embodied AI. *arXiv preprint arXiv:2103.14025*.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory*, *6*(3), 225–255.

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*. Citeseer.

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

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, *18*(1–3), 65–96.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*(3), e1002410.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(10), 3098–3103.

James, K. H. (2010). Sensori-motor experience leads to changes in visual processing in the developing brain. *Developmental Science*, *13*(2), 279–288.

James, K. H. (2017). The importance of handwriting experience on the development of the literate brain. *Current Directions in Psychological Science*, *26*(6), 502–508.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, *73*(1–2), 31–68.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759.

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., Liu, Y., & Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*(4), 454–469.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Maglio, P. P., & Kirsh, D. (1996). Epistemic action increases with skill. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 391–396).

Mason, M. T. (2018). Toward robotic manipulation. *Annual Review of Control, Robotics, and Autonomous Systems*, *1*, 1–28.

McCarthy, W. P., Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. *arXiv preprint arXiv:2107.00077*.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–131.
Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*(4), 441–474.
Proffitt, D. R., & Gilden, D. L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 384.
Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.
Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2021). Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. *bioRxiv*.
Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116.
Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. *Psychology of Learning and Motivation*, *31*, 305–349.
Shelton, A. L., Davis, E. E., Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., & Landau, B. (2022). Characterizing the details of spatial construction: Cognitive constraints and variability. *Cognitive Science*, *46*(1), e13081.
Sheridan, H., & Reingold, E. M. (2017). Chess players' eye movements reveal rapid recognition of complex visual patterns: Evidence from a chess-related visual search task. *Journal of Vision*, *17*(3), 4.
Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different physical intuitions exist between tasks, not domains. *Computational Brain & Behavior*, *1*(2), 101–118.
Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.
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.
Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.
Tian, L., Ellis, K., Kryven, M., & Tenenbaum, J. (2020). Learning abstract structure for drawing by efficient motor program induction. *Advances in Neural Information Processing Systems*, *33*, 2686–2697.
Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169.
Van Harreveld, F., Wagenmakers, E.-J., & Van Der Maas, H. L. (2007). The effects of time pressure on chess skill: An investigation into fast and slow processes underlying expert performance. *Psychological Research*, *71*, 591–597.
Van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. In *CogSci*.
van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*, 1–6.
van Opheusden, B., & Ma, W. J. (2019). Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences*, *29*, 127–133.
Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, *69*(9), 1752–1776.
Wertheimer, M. (1923). Laws of organization in perceptual forms. *Psycologische Forschung*, *4*.
Wolfgang, C. H., Stannard, L. L., & Jones, I. (2001). Block play performance among preschoolers as a predictor of later school achievement in mathematics. *Journal of Research in Childhood Education*, *15*(2), 173–180.
Wong, C., McCarthy, W. P., Grand, G., Friedman, Y., Tenenbaum, J. B., Andreas, J., Hawkins, R. D., & Fan, J. E. (2022). Identifying concept libraries from language about object structure. *arXiv preprint arXiv:2205.05666*.
Xia, L., & Collins, A. G. E. (2020). Temporal and state abstractions for efficient learning, transfer and composition in humans. *bioRxiv*.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 Parameter estimates for linear mixed effects model used to predict F1 score from attempt (first and final) and condition.

Table S2 Parameter estimates for linear mixed effects model used to predict number of blocks from attempt (first and final) and condition.

Table S3 Parameter estimates for linear mixed effects model used to predict number of blocks from attempt (first and final) and condition, excluding trials in which the trial ended early due to a block falling.

Table S4 Parameter estimates for linear mixed effects model used to predict the mean time (seconds) between block placements from attempt (first and final) and condition.

Table S5 Parameter estimates for linear mixed effects model used to predict preparation time from attempt (first and final) and condition.

Table S6 Parameter estimates for linear mixed effects model used to predict total build time (in seconds) from attempt (first and final), condition, and variable indicating whether the reconstruction was perfect.

Table S7 Parameter estimates for linear mixed effects model used to predict action dissimilarity from attempt pair, dissimilarity measure, and F1 score of the previous attempt.

Table S8 Parameter estimates for linear model used to predict difference in Gini coefficients from attempt (first and final) and length of action sequence considered (linear and quadratic) (all trials). df = 155.

Table S9 Parameter estimates for linear model used to predict difference in Gini coefficients from attempt (first and final) and length of action sequence considered (linear and quadratic) (perfect reconstructions only). df = 156.