Explaining intuitive difficulty judgments by modeling physical effort and risk

Ilker Yildirim[*1,2] (ilkery@mit.edu), **Basil Saeed**[*1] (bsaeed@mit.edu),
Grace Bennett-Pierre[3] (gbp@stanford.edu), **Tobias Gerstenberg**[3] (gerstenberg@stanford.edu),
Joshua B. Tenenbaum+[1] (jbt@mit.edu), **Hyowon Gweon+**[3] (hyo@stanford.edu)

[1] Department of Brain and Cognitive Sciences, MIT [2] Department of Psychology, Yale University
[3] Department of Psychology, Stanford University
* indicates co-first authors; + indicates co-senior authors

Abstract

The ability to estimate task difficulty is critical for many real-world decisions such as setting appropriate goals for ourselves or appreciating others' accomplishments. Here we give a computational account of how humans judge the difficulty of a range of physical construction tasks (e.g., moving 10 loose blocks from their initial configuration to their target configuration, such as a vertical tower) by quantifying two key factors that influence construction difficulty: physical effort and physical risk. Physical effort captures the minimal work needed to transport all objects to their final positions, and is computed using a hybrid task-and-motion planner. Physical risk corresponds to stability of the structure, and is computed using noisy physics simulations to capture the costs for precision (e.g., attention, coordination, fine motor movements) required for success. We show that the full effort-risk model captures human estimates of difficulty and construction time better than either component alone. Preprint URL: [http://arxiv.org/xxxxxxx]

Keywords: difficulty estimation; intuitive physics; physical reasoning; hybrid task-and-motion planning

Introduction

Suppose two people are playing with blocks. Sally stacks 10 blocks to build a vertical tower (Figure 1, from 7-H-A to 7-H-B), and Anne puts 10 blocks side by side to form a horizontal line (Figure 1, from 7-E-A to 7-E-B). Which one is harder to make? Even though both Sally and Anne worked with the same number of blocks, we have the intuition that Sally's tower is harder to make than Anne's horizontal line. When prompted to explain why, you might note that Sally's tower is more likely to fall, so she might need to re-stack the blocks, or be more careful than Anne when placing the blocks so they don't fall.

Intuitive judgments of task difficulty are ubiquitous in our actions and decisions (Carroll, 1981). While as adults, we rarely have to think seriously about the difficulty of building block structures, estimating task difficulty is critical for efficiently navigating the physical and social world; for instance, we often think about how difficult it would be to climb a new route at the climbing gym, follow an unfamiliar recipe for dinner, or finish a conference paper before the deadline (Gweon, Asaba, & Bennett-Pierre, 2017).

Sometimes, our prior experience can give us a vague sense for the difficulty of similar tasks (e.g., "making oatmeal is not that hard"). However, judgments of difficulty often have to be made before we actually engage in a task, and estimating task difficulty is particularly important for tasks we've never completed before. For novel tasks, we might want to first evaluate the probability of success to even decide whether to try that task at all, set more reasonable goals, or decide whether additional help is necessary (Gweon & Schulz, 2011). In these cases, one must represent the current state of the world, the desired goal state, and make a plan for how to get from the initial state to the final goal state.

The explanations that we generate can sometimes provide a useful glimpse into the underlying computations. As in the example above, to explain why Sally's tower is harder to build than Anne's line, we may allude to the physics behind the building process (e.g., stability of the structures and the probability of falling over) and the nature of the actions required to complete the task. Yet, it remains unclear whether these explanations reflect the actual inferential processes that give rise to people's judgments. What representations and inferential processes underlie our intuitive judgments of task difficulty?

Estimating difficulty

Difficulty is an abstract, unit-less construct that cannot be measured on a single, absolute scale. Difficulty incorporates both task-specific properties (e.g., Task A can be more difficult than Task B) and agent-specific attributes (e.g., a task can be more difficult for Tom than for Suzy depending on their competence). Difficulty can refer to both physical effort (e.g., force, work, time, fine motor control) as well as mental effort (e.g., attention, cognitive control, inhibition, or even requisite level of skill or intelligence). Even in making seemingly simple inferences such as judging Sally's tower as harder to make than Anne's line, people's explanations often refer to both types of effort: Sally's actions need to be more precise (physical effort) and she needs a higher level of concentration to correctly execute her plan (psychological effort). Unlike the height or weight of the structures that have clear ground-truth values that can be measured in objective units (e.g., 10 inches tall, 5 ounces in weight), the difficulty of building a structure, while certainly related to its height or weight, cannot be easily expressed as a single point estimate.

Nevertheless, there is reasonable consistency and systematicity in human observers' estimates of task difficulty. Recent work suggests that adults can provide difficulty estimates that are tightly correlated with the time it takes to build a structure, and that even preschool-aged children are quite accurate at judging the relative difficulty of two block structures similar to those shown in Figure 1 (Gweon et al., 2017). Critically, these judgments and estimates were made from simply looking at still photographs of the structures, rather than having actually engaged in the building process. This sug-
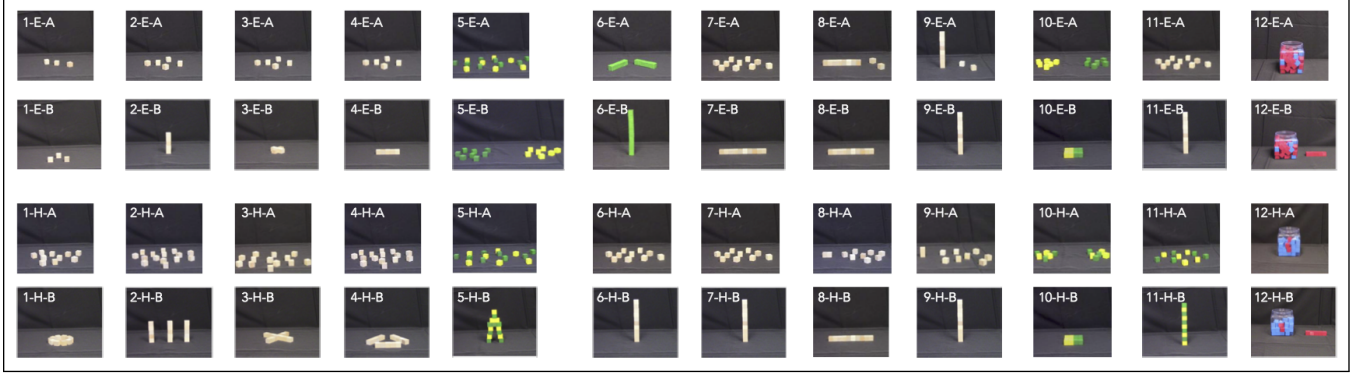
Figure 1: Experimental stimuli comprise 24 trials, each with an initial state (A) and final state (B). Trials are divided into 12 matched pairs, with one task that is intuitively easier and another harder, marked E, H respectively. (Note that E and H are relative designations within a given pair of trials. A similar task may be easy in one pair and hard in another, e.g. pairs 7, 11.)

gests that even though difficulty may not have an objective unit that cannot be directly measured, people use proxy measures like time or amount of work to communicate their intuitions in some measurable format (e.g., this structure might take 15 seconds to complete), and even use these estimates to generate ad-hoc scales that allow them to express difficulty in relative terms (e.g., Task X is more difficult than Task Y). However, the cognitive mechanisms that support these computations remain poorly understood.

Modeling difficulty

Here we give a computational account of these behaviors by drawing on recent work in robotics and cognitive science. We assume that the observer estimates both the *physical effort* required to carry out an efficient plan as well as the *physical risk* arising from the potential instability of the target construction. We model physical effort as the total kinetic energy required to intervene on the world to change the current state to the desired final state; we do this by using a hybrid task-and-motion planner that first finds a symbolic action plan (e.g., pick, place, hold), and then translates this symbolic plan into continuous motion trajectories of the agent's actions. We model physical risk using noisy physics simulations to capture aspects of the construction process that go beyond just transporting blocks from their initial to their final positions. In the face of risk (i.e. potential physical instability), our agent increases its motor precision to create a more stable structure. There are a number of ways in which an agent might increase precision; to minimize noise, an agent might engage in fine motor control and coordination, or deploy cognitive control or attention. All these efforts to ensure appropriate level of precision is reflected in physical risk, a probabilistic estimate of the stability of the final target state. Recent work on intuitive physical reasoning suggests that humans have an "intuitive physics engine" (IPE) that allows them to predict future physical states of the world by engaging in mental simulation (Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Zhou, Smith, & Tenenbaum, 2017). However, the focus of this prior work was primarily on explaining how people predict the future states of the world in the ab-

sence of any agent interventions, such as judging whether a block tower will fall given the weight of its components or in which direction the tower is likely to fall. Yet, an internal physics engine is also critical for action planning and object manipulation. Indeed, most advanced object manipulation systems in robotics combine both agent models and physics (Todorov, 2018; Toussaint, 2015; Toussaint, Allen, Smith, & Tenenbaum, 2018).

Imagine an agent in an "initial state" where 10 loose blocks lie scattered on the floor (Fig. 1, 7-H-A). The agent could perform the first intervention by picking up a block, moving it across the space, and placing it at a particular position. The next action could involve picking up another one, and placing it on top of the first one. By combining the IPE with an intuitive understanding of goal-directed actions of an agent, human observers can simulate forward the process of building a block structure. Human observers can also reason backwards from a representation of the final completed task to estimate the number of actions, or the time that it took, to build the tower (Gweon et al., 2017). Recent work suggests that even children engage in physical problem solving in ways that are similar to adults (Cortesa et al., 2018, 2017). One intriguing possibility is that the kinds of proxy measures that human subjects use intuitively (e.g., time to completion, etc.) are generated from such process of physical simulation conditioned on agent action.

Here we evaluate this hypothesis through computational modeling. The process of building such block structures lends itself well to precise formalization. The physical principles involved in these block-building tasks are captured in Newtonian mechanics (e.g., gravity, forces including support relations, motion). Existing physics engines can simulate these principles efficiently and realistically (Coumans, 2010), and are successfully utilized to explain aspects of human physical scene understanding as noted above (Battaglia et al., 2013; Gerstenberg et al., 2017). Here, building on this and other prior work (Yildirim, Gerstenberg, Saeed, Toussaint, & Tenenbaum, 2017), we focus on how IPE might be deployed in our everyday reasoning about actions and their consequences. More specifically, we use IPE model to es-

timate the physical risk (stability) of a given plan; we argue that people's intuitive estimates of task difficulty are rooted in people's abilities to engage in action planning by simulating an agent's goal-directed physical manipulation of objects.

Experiments: Estimated and actual difficulty

We collected data from human adults to characterize the actual difficulty (measured in building time) of the block structures as well as their estimates of expected building time (in seconds) and estimates of difficulty (on an arbitrary scale). **Participants** The data came from three separate tasks. 1) *Difficulty Estimation task*: $n = 59$ (20 females), $M_{age} = 31.62$ (SD $= 8.65$); 2) *Time Estimation task*: $n = 60$ (23 females), $M_{age} = 32.73$ (8.72); and 3) *Build Task*: $n = 20$, $M_{age} = 18.75$ (2.20). The Difficulty Estimation and Time Estimation tasks were conducted on Amazon MTurk. Two participants were excluded for providing identical responses on all trials. The Build task was conducted in the lab (these data were originally presented in Gweon et al., 2017). **Materials** Photographs of 24 sets of structures made of 1-inch wooden blocks were used (12 pairs, 1 easy and 1 hard structure). There were two photographs for each structure: "initial state" (for example, scattered blocks) and a "final state" (for example, a 10 block vertical tower). On the screen they were labeled "Start" and "Finish" respectively. The order of presentation was randomized across 12 pairs and counterbalanced within a each pair. We also used one particularly "hard" and one particularly "easy" structure to help anchor their subsequent judgments it the beginning of the task. **Procedure** Depending on the task, participants used a sliding scale to answer one of these questions for every trial: "How difficult would it be to do this?" (*Difficulty Estimation Task*), "How long would it take to make this?" (*Time Estimation Task*). In order to account for differences in scale use between participants, we z-scored each participant's responses. For the *Build Task* (in-lab), the experimenter laid out blocks in front of the participant in the same configuration as the initial photo on each trial, and asked the participant to create the structure shown in the final state photo. We recorded how long the subject spent building the block structure from start to finish using a key press.

Model

The model consists of two components. The first component evaluates the amount of physical effort that would be required to change an initial spread of blocks on flat surface (Initial State, or State "A") to a particular target configuration of blocks (Final State, or State "B"; see Fig.2). We define physical effort as the total kinetic energy required to transport all blocks to change the current state A to B. We derive this quantity using a hybrid task-and-motion planner. This planner is based on the logic-geometric programming framework and consists of an initial stage of symbolic action sequence planning to go from A to B followed by trajectory optimization to fulfill the symbolic plan with respect to a simulated humanoid robot. The planner optimizes for a geometric cost

term consisting of the sum of squared accelerations.

The second component of the model captures the physical risk – due to potential instabilities – involved in transitioning from State A to B. For humans, these aspects might include how much attention and motor precision is required to keep the structure from falling during its intermediate stages. We approximate these various facets of physical risk in a probabilistic estimate of the stability of the target state B. We define physical risk as how likely B is to fall, which we estimate by aggregating the outcomes of multiple forward physics simulations each starting from a configuration where the 3-D position of each block in State B is perturbed with a small random noise (cf. Battaglia et al., 2013).

Modeling physical effort

We model physical effort using a hybrid task-and-motion planner – specifically, the logic-geometric programming solver (Toussaint, 2015). Inputs to this planner are symbolic representations of State A and State B. For each object (i.e. a single block), the input representation specifies its dimensions, position, and support relations (the list of other objects supporting this object). Support relations across all objects can be described in a simple directed graph, which we refer to as a "support graph", with nodes indicating objects and edges indicating support relations.

Assignment of identities to objects within each symbolic representation is arbitrary and identities across States A and B need to be matched carefully. To illustrate, consider 11-E-A and 11-E-B in Fig.1. All blocks appear identical and there is no reason subjects or the model should assume a particular assignment based on visual appearances of blocks. Instead, in the model, we choose a one-to-one assignment of blocks in State A to blocks in State B that minimizes the total Euclidean distance between the initial and final positions of all blocks while taking into account the blocks' colors (e.g., a green block in 11-H-A cannot be assigned to a yellow block in 11-H-B). To find the mapping with a minimal total Euclidean distance we use the Hungarian Algorithm (Kuhn, 1955).

Now given the representations of State A and State B and the optimal mapping of objects from one state to the other, the planner proceeds in two stages: 1) a *symbolic planner* and 2) a *geometric planner*.

Symbolic planning In the first stage, a task planner generates a sequence of abstract actions to transform State A to State B. In our setting, this is equivalent to finding an ordering of blocks by which State B will be constructed from blocks in State A. We find this ordering by exploiting the support graph of state B. We find a sequence of actions that builds state B from the ground up, making sure that each block to be placed has its support already in place. Fig.2 shows one such assignment of blocks for an example state pair and the resulting symbolic plan by respecting the ordering arising from the support graph of state B. For every block that needs to be moved, we use two action primitives (implementing GRASP and then PLACE actions with just a single hand of the robot). The length of the plan ranged from 4 to 30 actions across the
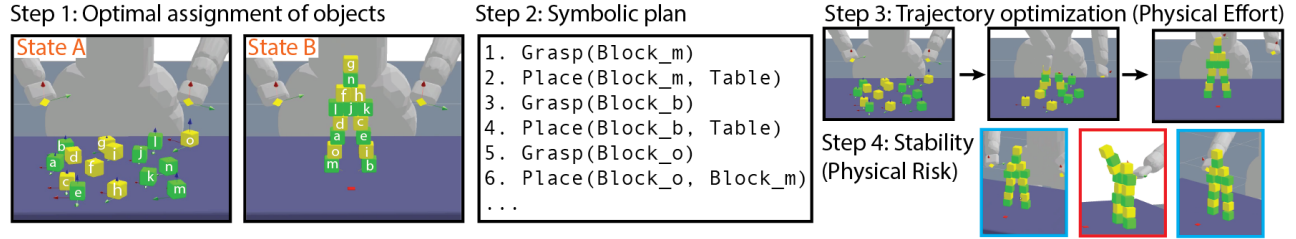
Figure 2: Overview of the modeling approach. In step 1, the model chooses an assignment of objects from state A to B minimizing the total Euclidean distance between initial and final 3D positions. In Step 2, the model finds a symbolic plan that respects the support graph of State B. Two action primitives are sufficient to solve all of our trials, `Grasp(Obj)` and `Place(Obj1, Obj2)`. In Step 3, we perform trajectory optimization to find a full 3D path for each block to be transported from an initial to a final position. Finally, in Step 4, we estimate stability (we define physical risk is as 1 - stability) by introducing random but small amounts of perturbations to State B. Blue frames show stable runs, the red frame is unstable.

trials depending on the number of blocks that were needed to be transported.

This symbolic plan does not take into account the geometry of the shapes, or their rotations in space, and how they would be transported by an efficient agent. Instead, it treats each object as a point object that is supported by another point object. The only relationships this planner respects are the support relations between objects. The distance between blocks is implicitly taken into account due to the optimal assignment of blocks as pairs prior to symbolic task planning.[1]

Geometric planning The second stage of the planner, the geometric planner computes a full path in 3D for all objects through trajectory optimization. The symbolic plan as well as representations of State A and B are fed to the geometric planner, which uses the geometric information in the initial and final state to create a set of constrained optimization problems. The geometric planner finds the optimal movements that the simulated humanoid robot must take to move the blocks according to the sequence of moves in the symbolic plan. Following Yildirim et al. (2017), the planner minimizes squared acceleration of the joints of the simulated humanoid robot using the KOMO (k-order motion-optimization) package introduced in (Toussaint, 2015).

Measuring physical effort Physical effort is measured by the kinetic energy needed to execute the resulting full plan. Kinetic energy, $\frac{1}{2}mv^2$, requires knowing the velocity by which the object is moved and its mass value for each step of the symbolic plan. Velocity values are calculated from the geometric planner by using the velocity with which the blocks are being moved for each step of the symbolic plan while the solution is being executed. Note that these velocity values result from the optimization procedure that minimizes the sum of squared accelerations. We assume that the blocks are of equal size and equal mass and assigned a unit mass to a unit block. We obtain the total kinetic energy required to execute a plan for a given pair of states A and B by summing up the energy incurred by each step of the symbolic plan.

We used $M = 30$ random samples to represent uncertainty in the initial configurations that arises from the exact positions of scattered blocks (blocks that are not part of a clear structure; or more specifically blocks that are laying on the table without supporting any other block, e.g., all blocks in Fig. 1 5-E-A).[2] This randomness in the initial configurations creates variation from simulation to simulation in how the blocks are assigned between A and B, and what symbolic and geometric solutions are generated. For a given pair A and B, we calculate the total kinetic energy for each randomly drawn A, and average across all 30 samples.

One trial (pair 12E and 12H in which blocks are placed in a bucket) needed additional assumptions to solve. The planner is incapable of searching for specific blocks from a mix of blocks in the box, or calculating paths that avoid passing blocks through the box. Thus, we handled the initial configurations in the following way: We lined up the 20 blocks in a pre-specified order (a 10×2 array) and randomly chose which 5 should be colored blue for trial 12-E and which should be colored red for trial 12-H (cf. Jara-Ettinger, Sun, Schulz, & Tenenbaum, 2018). Like other trials, we solve and obtain the total kinetic energy for each 30 such random initial configurations, and then average across to obtain an estimate of physical effort.

Modeling physical risk

We define the physical risk associated with a trial by how likely its final configuration B is to fall. We estimate this physical risk following Battaglia et al. (2013) by approximating a distribution over how robust the final configuration of blocks B is to small perturbations of the blocks' positions. We ran $N = 100$ simulations and recorded whether or not the final configuration remained stable after noise was applied. For each sample, we modified B by randomly perturbing the position of each block in while ensuring that there is no overlap between blocks (see Fig. 2 for example perturbed configurations). The random perturbations for each block were drawn

---

[1]Note we are separately handling distance and support relation. Finding distance-minimizing pairs with respect to the topological order in the support graph might yield more efficient possible solutions in some cases.

[2]We did not systematically evaluate how the value of M would impact our results. However, we noticed that the variation across our 30 samples were not high to justify a bigger M, and it is possible that even smaller values of M will produce very similar results as reported here.
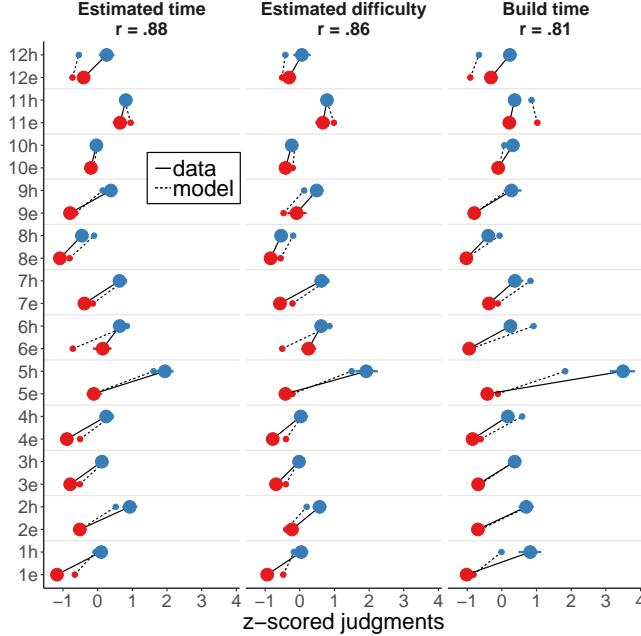
Figure 3: Mean z-scored human judgments (large points connected by solid lines) and fitted model results (small points connected by dotted lines) with sigma = 0.065. The model closely tracks human judgments in each task. It also captures the easy vs. hard distinction in all pairs of easy (red) vs. hard (blue) trials (e.g., 1e vs 1h) except for 11e vs. 11h. Error bars show 95% bootstrapped confidence intervals.

i.i.d. from a Gaussian distribution with 0 mean and standard deviation σ. Each perturbed configuration was simulated forward using a physics engine (we used PhysX physics engine in our simulations) either until all motion stopped or until a given period of time elapsed[3]. We then estimated physical risk of the target configuration B to be the fraction of simulations where the tower fell relative to all N simulations. We fitted σ to the data using a simple grid search exploring values between $0.05d$ to $0.1d$ with increments of $0.005d$ where d is the length of the the block along the axis it is being perturbed (see beginning of the Results section for the details of the fitting procedure).

The full model and lesioned models

We denote the estimated physical effort as E_i and the estimated physical risk as R_i where i indexes the trial number across the 24 trials in our experiments (Fig. 1). Note that units of E_i are in joules and non-negative, whereas R_i is a probability between 0 and 1. Our full model partitions E_i into two factors based on R_i and fits a different coefficient for its risk portion $E_i \times R_i$ and its risk-free portion $E_i \times (1 - R_i)$ based on the average human responses:

$$y_i = \beta_0 + \beta_1 E_i(1 - R_i) + \beta_2 E_i R_i \quad (1)$$

[3] To determine whether a tower fell, we simulated it starting from its initial configuration for 4 seconds. If the difference between the final and initial positions of any of the blocks is above a set threshold, we considered that tower to have fallen.
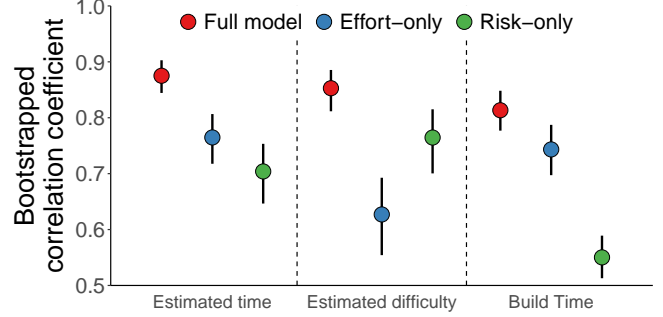

Figure 4: Comparisons of the full model, effort-only model, and risk-only model. Filled circles show median correlation values based on 1000 bootstrapped samples. Error bars show the 2.5 and 97.5 percentile.

where y_i denotes average human response for trial i and $\beta_0, \beta_1,$ and β_2 are coefficients.

We compare the full model with lesioned models that only consider the physical effort, or the physical risk. The *physical effort only* model was fit using $y_i = \beta_0 + \beta_1 E_i$. The *physical risk only* model was fit $y_i = \beta_0 + \beta_1 R_i$. Since we evaluate the *physical risk only* model merely with respect to how well it correlates with the data (see Fig. 4), fitting a slope and intercept does not affect the results.

Results

In the full model, we estimated the noise parameter σ for the physical risk estimation component of the model by fitting the linear regression in Equation 1 for each dataset individually, and then exploring the root mean squared error (RMSE) between model predictions and behavior. We set *sigma* to 0.065 (RMSE was higher for lower noise levels, and didn't change much for a range of levels). The full model which combines effort and physical risk better captures the data than the effort-only model (Figure 4).

Figure 3 shows the best fitting full model results against the data at the trial-level granularity for each task. The full model's (combined physical effort and risk) predictions closely tracked the data and are correlated highly with each of the three tasks. Not only did the full model capture the overall distribution of responses, it also recapitulated the easy versus hard distinction in most trial pairs. The model produces the correct rank order (assigning a smaller cost to the easy pair than the hard pair) in a11 of the pairs. The only exception is the pair 11E and 11H, where humans did not show any difference across all three datasets.

To what extent can we explain the behavior just using physical effort or just using physical risk? We correlated the predictions of the full model with participants' judgments of physical effort and physical risk and physical build times. We obtained confidence intervals on the model's correlation with the data by bootstrapping over participants in each dataset. We found that the fitted full model explained participants' judgments behavior better than either of effort or risk in each dataset. Interestingly, whereas in both the Estimated Time

and Build Time dataset the effort-only model better explained behavior than the risk-only model ($p < 0.05$ direct bootstrap hypothesis testing), in the Estimated Difficulty dataset the risk-only model performed significantly better than the effort-only model ($p < 0.05$ direct bootstrap hypothesis testing; Figure 4). To focus on the two judgment datasets, at the surface level, the responses of the Estimate Time and Estimated Difficulty participants are quite similar, and the full model indeed correlates with each of them to a similar degree. However, their responses appear to differentially correlate with the two separate components of our model: the risk-only model correlates better with the Estimated Difficulty dataset whereas the effort-only model correlates better with the Estimated Time dataset (in both comparisons $p < 0.05$, direct bootstrap hypothesis testing; Figure 4). This suggests that participants in these two groups might have attended to different features or components of the reconfiguration problems. Participants in the Estimated Difficulty experiment might have attended more to how difficult it would be to keep the structure stable. Participants in the Estimated Time experiment, on the other hand, seemed to have weighted the overall distance that each block would have to be moved to get from A to B in each pair, which is captured by physical effort in our model.

Discussion

Intuitions about the difficulty of tasks are pervasive in our everyday decisions and actions. Yet, the representations and inferential processes that underlie these intuitions remain poorly understood. The current work brings together ideas and tools from cognitive science and robotics to formalize human adults' judgments and estimates of difficulty in precise, quantitative terms. Here we focused on people's intuitions about the difficulty of simple block-building tasks. We developed a novel computational model that quantifies the degree of difficulty to transform an initial state of blocks into a final target configuration. In doing so, our model takes into account not only the kinetic energy required to move blocks but also the uncertainty in the stability of the structure introduced by noisy placements of blocks. This model captures human judgments well by considering physical effort and the level of care required to achieve one's goals in the physical world.

Even though both cost components were important to account for the data, the extent to which participants considered each cost varied between tasks. We find that physical risk (i.e. the level of care required) was particularly important for people's estimates of task *difficulty*, whereas considerations of physical effort strongly affected people's estimates about how much *time* it would take to go from A to B.

The present study builds on Yildirim et al. (2017). The models presented in both studies use the same solver for path-planning (Toussaint, 2015) to accomplish trajectory optimization in multi-step complex action sequences. However, the present study differs from the previous work in two important ways. First, we account for uncertainty arising from the specific way in which the blocks might be scattered in the initial

configuration. We also handle the mapping problem between states A and B (i.e., a particular wooden block in state A can go to one of many wooden blocks in state B) by choosing the assignment that minimizes the total Euclidean distance between the assigned pairs. Second, our model treats the contributions of physical risk (stability) and physical effort independently from each other whereas Yildirim et al. (2017) integrated stability tests within the intermediate stages of the geometric solver to filter out physically impossible plans. Our choice of isolating risk from effort was not only computationally more efficient but also allowed us to better understand human judgments across different tasks.

The way in which our model evaluates and uses physical risks to plan suggests a novel approach to functionally characterize aspects of mental effort (Kool & Botvinick, 2018). For example, when stacking blocks into a precarious configuration, we need to continuously attend to the task at hand, and finely plan our motor actions. More generally, our model emphasizes the role of internal object representations (e.g., shapes, positions, support relations, mass) as well internal body models that can support thinking about complex object manipulations. We see this intersection of physical reasoning, action planning, and difficulty estimation as a promising platform to formalize and reverse-engineer what computations are involved in the exertion of mental effort.

Interestingly, the model predictions showed relatively worse fits with how long it actually took participants to solve the different tasks compared to the estimated difficulty and time judgments. It's possible that people care about physical risk to different degrees at different points in the building process – start more liberally but become more constrained and focused at the end (to avoid an almost completed tower from crashing). We can model a change in the level of care depending on the current task at hand by having the planner reason about the physical risk at each time step and then actively controlling the level of noise *sigma* as required.

Our current model assesses stability only once at the end, without evaluating it in the intermediate stages throughout the execution of a plan. Arguably this is different from how a human planner would perform the task. While one might consider this as a limitation of the model, for this particular study, constraining the frequency of physical simulation was not only computationally convenient but also informative: it allowed us to straightforwardly determine the contributions of effort and risk to explaining human judgments. Future work should consider adaptive allocation of computation to estimating stability throughout the intermediate stages of a plan as described above.

In current work, the level of noise in physics simulations was fit to the data. However, it would be interesting to try and infer the level of noise directly from people's actions. Different people will differ in how carefully they execute their actions, and in how skillful they are in interacting with the physical world. Recent work in cognitive development suggests that humans have representations of their own, as

well as others' competence even early in life (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). These representations of competence can inform decisions about how to efficiently allocate tasks across agents (Magid, DePascale, & Schulz, 2018), choose whom to help (Bennett-Pierre, Asaba, & Gweon, 2018), learn from others' effort (Leonard, Lee, & Schulz, 2017), and even teach what would be too difficult for others (Bridgers, Jara-Ettinger, & Gweon, 2016).

This study considered simple blocks-world type scenarios as a "model organism" to characterize the computations that support humans' intuitively inferences about task difficulty. One question is to what extent our current model can capture human judgments across different tasks or scenarios outside this domain. Although much further work is needed to capture people's sense of difficulty in non-physics domains (e.g., navigating an unfamiliar environment, solving a math problem, learning a new skill), we believe that a range of tasks that involve manipulation of physical objects could be modeled using our approach by adjusting a few aspects of the symbolic and geometric planner. Future work should explore how our approach can be extended to different manipulation tasks and to scenarios with more complex physics including objects with different shapes, density, and substance properties (e.g., soft bodies, liquids).

We look forward to these lines of future work that will help us better understand the inferential processes that underlie these behaviors, and build better artificial agents that can make these judgments in ways that humans do.

Acknowledgments

References

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.

Bennett-Pierre, G., Asaba, M., & Gweon, H. (2018). Preschoolers consider expected task difficulty to decide what to do and whom to help. In *Proceedings of the 40th annual conference of the cognitive science society (austin, tx: Cognitive science society)*.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2016). Children consider others expected costs and rewards when deciding what to teach. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 559–564).

Carroll, J. B. (1981). Ability and task difficulty in cognitive psychology. *Educational Researcher*, *10*(1), 11–21.

Cortesa, C., Jones, J., Hager, G., Khudanpur, S., Landau, B., & Shelton, A. (2018). Constraints and development in childrens block construction. In *Cogsci*.

Cortesa, C., 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*.

Coumans, E. (2010). Bullet physics engine. *Open Source Software: http://bulletphysics. org*.

Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 409–414).

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults and preschoolers ability to infer the difficulty of novel tasks. In *Proceedings of the 39th annual conference of the cognitive science society*.

Gweon, H., & Schulz, L. E. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 785.

Jara-Ettinger, J., Sun, F., Schulz, L., & Tenenbaum, J. B. (2018, Feb). Sensitivity to the sampling process emerges from the principle of efficiency. *Cognitive Science*, *42*, 270–286.

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, 1.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, *2*(1-2), 83–97.

Leonard, J. A., Lee, Y., & Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290–1294.

Magid, R. W., DePascale, M., & Schulz, L. E. (2018). Four-and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. *Open Mind*, *1*(4), 194–207.

Todorov, E. (2018). Goal directed dynamics. In *2018 ieee international conference on robotics and automation (icra)* (pp. 2994–3000).

Toussaint, M. (2015). Logic-geometric programming: An optimization-based approach to combined task and motion planning. In *IJCAI* (pp. 1930–1936).

Toussaint, M., Allen, K., Smith, K., & Tenenbaum, J. B. (2018). Differentiable physics and stable modes for tool-use and manipulation planning. *Proc. of Robotics Science&Sysem*.

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenenbaum, J. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. *arXiv preprint arXiv:1707.08212*.