Think again?
The amount of mental simulation tracks uncertainty in the outcome

Jessica B. Hamrick[1] **(jhamrick@berkeley.edu), Kevin A. Smith**[2] **(k2smith@ucsd.edu),**
Thomas L. Griffiths[1] **(tom_griffiths@berkeley.edu), & Edward Vul**[2] **(evul@ucsd.edu)**
[1]University of California, Berkeley, Department of Psychology, Berkeley CA 94720 USA
[2]University of California, San Diego, Department of Psychology, La Jolla, CA 92093 USA

Abstract

In this paper, we investigate how people use mental simulations: do people vary the number of simulations that they run in order to optimally balance speed and accuracy? We combined a model of noisy physical simulation with a decision making strategy called the *sequential probability ratio test*, or SPRT (Wald, 1947). Our model predicted that people should use more samples when it is harder to make an accurate prediction due to higher simulation uncertainty. We tested this through a task in which people had to judge whether a ball bouncing in a box would go through a hole or not. We varied the uncertainty across trials by changing the size of the holes and the margin by which the ball went through or missed the hole. Both people's judgments and response times were well-predicted by our model, demonstrating that people have a systematic strategy to allocate resources for mental simulation.

Keywords: mental simulation; intuitive physics; SPRT; computational modeling

Introduction

How should the mind allocate its computational resources? Consider the game of Angry Birds, where the goal is to launch birds to knock down a tower. To take a shot, the player can imagine—or *mentally simulate*—the path the bird will take and how it will affect the tower. How long should the player spend thinking before they let each bird fly? If they spend very little time thinking, they are likely to miss the target. But, if they spend too long thinking, it will take much longer to receive the satisfaction of beating the level. More generally, if running simulations will provide a more accurate forecast but incur a sampling cost, how long should an agent spend simulating before acting?

In the domain of physical reasoning, research suggests that people make predictions about physical scenes—such as those found in Angry Birds—by running noisy physical simulations (Sanborn, Mansinghka, & Griffiths, 2013; Smith & Vul, 2013; Battaglia, Hamrick, & Tenenbaum, 2013; Smith, Dechter, Tenenbaum, & Vul, 2013; Smith, Battaglia, & Vul, 2013; Smith & Vul, 2014; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2014; Hamrick, Battaglia, Griffiths, & Tenenbaum, in prep.). However, while this research has investigated the *mechanism* for making these predictions, there has been very little investigation into how people *use* this mechanism. In particular, because the simulations are noisy, it may be beneficial to run multiple simulations in order to obtain more accurate predictions. Is there an optimal number of simulations to run in these situations? If so, do people behave optimally?

To investigate how many simulations people run, we focus on a dichotomous prediction task—will a ball in motion on a computer screen go through a hole, or miss it? To model this task, we combine a mechanism of noisy physical simulation with a decision strategy for sample-based agents. We consider the *sequential probability ratio test*, or SPRT, in which an agent takes samples that point to one hypothesis or another, and continues to do so until the net samples in favor of one hypothesis reaches a threshold, at which point that hypothesis wins (Wald, 1947). Often under the name of the *drift-diffusion* model, SPRT has been used to explain behavior in a number of decision-making tasks (e.g. Gold & Shadlen, 2007; Ratcliff & McKoon, 2008; Bitzer, Park, Blankenburg, & Kiebel, 2014), as it provides an optimal cost-benefit tradeoff between sampling and exploiting information (Wald & Wolfowitz, 1950). Importantly, the SPRT strategy predicts that people need to take more samples—and thus also will take a longer time to respond—when there is roughly equal evidence for both hypotheses.

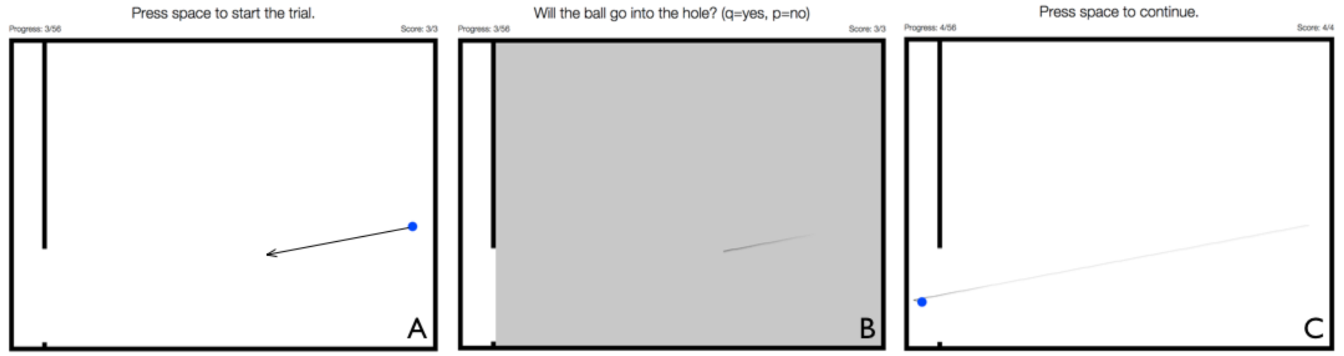
Drawing on the results from both physical simulation and decision-making, we hypothesize that people make predictions by running mental simulations, and that they vary the number of simulations based on their uncertainty. In this paper, we first formalize our model, combining the simulation model from Smith and Vul (2013) with the SPRT decision strategy. Next, we describe an experiment in which we asked participants to respond to the question of, "will the ball go through the hole?", and analyze peoples' judgments and response times. We then demonstrate that our model can explain the empirical pattern of responses and response times we observed. Finally, we discuss the implications of our results on the broader, underlying question: how should people make use of mental simulations?

Making decisions from mental simulations

Consider the task in Figure 1, in which people observe a ball moving inside a box, and must predict whether it will go through the hole. How do people solve this problem? Here, we formalize a model that answers this question by combining noisy physical simulations with a decision-making strategy known as the *sequential probability ratio test*, or SPRT.

Modeling physical simulation

There is a growing body of evidence that people reason about physical scenes like the one in Figure 1 by running noisy simulations. This hypothesis, referred to as the "noisy Newton" hypothesis (Sanborn et al., 2013), states that people have approximate knowledge of physical laws instantiated in a



Press space to start the trial.

Will the ball go into the hole? (q=yes, p=no)

Press space to continue.

Figure 1: **Example experimental trial.** Each panel shows a different part of the trial. *A:* the initial screen presented to the participant. The arrow was not part of the actual stimuli; it has been added to reflect the animation that participants observed after pressing "space". *B:* the screen is occluded after observing the stimulus presentation. The faded gray line shows the path the ball took during the initial presentation. *C:* the final position of the ball, after observing the feedback. As in the middle panel, the faded gray line shows the path of the ball.

runnable model of intuitive physics. Using this model, they can extrapolate the future by running a series of noisy simulations (Smith & Vul, 2013; Battaglia et al., 2013; Smith, Dechter, et al., 2013; Smith, Battaglia, & Vul, 2013; Smith & Vul, 2014; Ullman et al., 2014; Hamrick et al., in prep.).

Smith and Vul (2013) investigated the various sources of uncertainty in these simulations, finding that people's judgments were best captured by a model that took into account both perceptual uncertainty (noise in object locations and their trajectories) and dynamic uncertainty (noise in the object's motion—e.g., a textured floor would cause a ball to deviate from a straight line). Using this model, we hypothesize that people reason about the task in Figure 1 by running noisy physical simulations to estimate where the ball will go.

The SPRT strategy

If people are running simulations to reason about physical scenes, then how many simulations do they run? Because the simulations are noisy, it might be beneficial to run multiple simulations in order to get a better estimate of the outcome. However, each simulation comes with a time cost. To optimize this speed-accuracy tradeoff, we apply the *sequential probability ratio test*, or SPRT (Wald, 1947) to the samples drawn from the physical simulation. By combining these two models, we can make predictions both for people's judgments, and how long they take to make those judgments.

We consider binary (or two-alternative forced choice) decisions, where an agent must choose one of two hypotheses, H_0 or H_1. In the case of the task in Figure 1, H_1 is the hypothesis that the ball goes in, and H_0 is the hypothesis that it does not. The agent may take samples X_i from a Bernoulli distribution with an unknown parameter p (the probability of sampling evidence for H_1), and from these samples estimate the probability that H_1 is correct: $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} X_i$, where N is the total number of samples. Then, the decision rule that minimizes the probability of error is $\hat{H}(X_1, \ldots, X_N) = H_0$ when $\hat{p} < 0.5$ and $\hat{H}(X_1, \ldots, X_N) = H_1$ when $\hat{p} > 0.5$.

In the best possible case, the agent takes infinitely many samples and chooses the *maximum a posteriori* (MAP) hypothesis with probability p. In practice, the agent cannot take infinitely many samples. Thus, to determine when to stop sampling (i.e., what the value of N is), the agent continues to sample until the net evidence Y_N reaches some threshold, either $Y_N = T$ to select in favor of H_1 or $Y_N = -T$ to select in favor of H_0. The net evidence is the sum of samples in favor of H_1 minus those in favor of H_0, or $Y_N = \sum_{i=1}^{N} 2X_i - 1$.

Independent of the actual number of samples taken, the probability of choosing the MAP hypothesis (H_1) is:

$$\Pr[\hat{H}(Y_N) = H_1 \,|\, H_1, T, p] = \frac{p^T}{p^T + (1-p)^T}, \qquad (1)$$

and the expected number of samples taken before reaching either $Y_N = T$ or $Y_N = -T$ is given by:

$$\mathbb{E}[N \,|\, T, p] = \frac{T}{1-2p} - \frac{2T}{1-2p} \cdot \frac{1 - ((1-p)/p)^T}{1 - ((1-p)/p)^{2T}}, \qquad (2)$$

which is derived by Feller (1968, ch. XIV, eq. 3.4).

Combining simulation and SPRT

In order to combine simulation and SPRT, we used the model from Smith and Vul (2013) to sample possible trajectories of the ball, from which we estimated a truncated normal posterior predictive distribution of where the ball will go. From this distribution, we compute the probability p that the ball goes in the hole as the probability mass overlapping the hole. This probability can then be used to compute Equations 1 and 2, which give a formal hypothesis for what decisions people make, and how long it takes them.

Because our experiment (described in the next section) was performed online, we needed to fit the parameters of the model from Smith and Vul (2013) to reflect these different viewing conditions. To do this, we performed an online replication of the experiment from Smith and Vul (2013) in which we asked people to catch a ball like the one in Figure 1 using a paddle that could move up and down along the y-axis (see

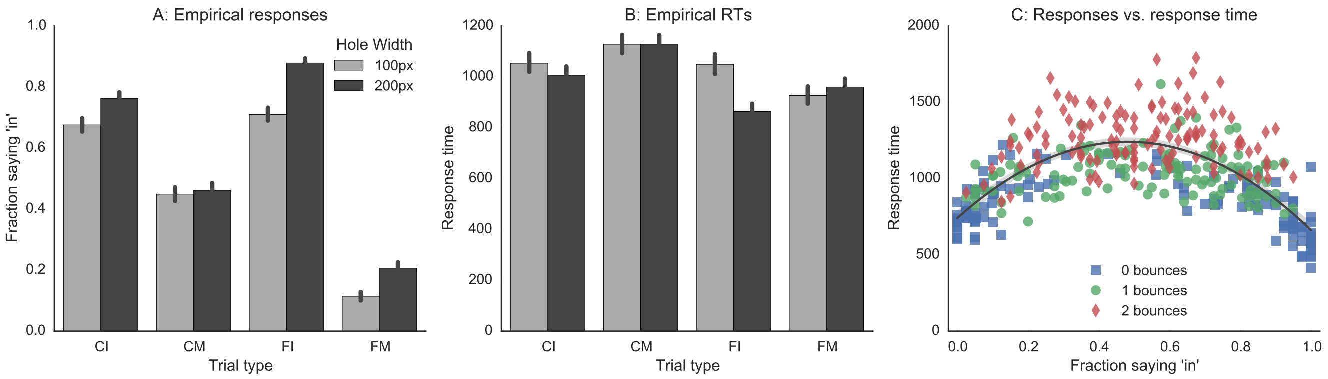


Figure 2: **Response characteristics as a function of trial type.** *A:* each bar shows the proportion of participants saying that the ball will go in the hole for a particular trial type (x-axis) and hole size (color). *B:* like the left subplot, but the y-axis shows bootstrapped logarithmic means of RTs. *C:* each point corresponds to a different stimulus, trial type, and hole size. The x-axis is the proportion of participants saying the ball will go in the hole, and the y-axis is the logarithmic mean RT. The black line indicates a 2nd-order polynomial fit between responses and RTs and the shaded gray region indicates the 95% confidence interval around the fit.

Appendix). If we assume participants in this auxiliary experiment took on average M samples, then the standard deviation of their judgments ($\sigma_{judgments}$) is not equal to the standard deviation of their simulations (σ_{sims}), but is instead related by the equation: $\sigma_{judgments} = \sigma_{sims}/\sqrt{M}$. Therefore, to estimate σ_{sims}, we allowed for a free parameter, $\sigma_{adj} = \sqrt{M}$, which multiplied our original estimate of the standard deviation.

Testing the SPRT model of mental simulation

To determine whether people choose simulations in a way consistent with SPRT, we ran an experiment in which people made a binary judgment about whether a ball traveling across a computer screen would go through a hole (see Figure 1). We designed the trials to elicit a range of responses by varying the margin by which the ball either missed or went through the hole. According to SPRT, when people's simulations are uncertain—i.e., when the probability that the ball goes in the hole is close to $p = 0.5$, such as when the ball just barely goes through the hole—they should be slower to respond. People should be faster to respond when their simulations are more certain, such as when the ball misses the hole by a wide margin.

Participants

We recruited $N = 328$ participants on Amazon's Mechanical Turk using the psiTurk (McDonnell et al., 2014) experimental framework. Participants were treated in accordance with UC Berkeley IRB standards and were paid $0.60 for 6.5 minutes of work. Participants were randomly assigned to one of eight conditions, which determined which stimuli they judged (see Stimuli). We excluded $N = 8$ participants for answering incorrectly on more than one control trial (see Stimuli), leaving a total of $N = 320$ participants.

Procedure

On each trial, participants were shown the scene, including the initial position of the ball and the location of the hole. Participants were instructed to press "space" to begin the trial, after which an animation of the initial stimulus began (see Stimuli). As soon as this animation concluded, a gray box was drawn over the screen, occluding the ball (but not the line depicting the path it had traveled so far; this was left in as a reminder of where the ball had come from). Participants were asked, "will the ball go in the hole?", and were instructed to press 'q' if they thought it would, and 'p' otherwise. After responding, text appeared saying "Correct!" or "Incorrect." The gray occluder was removed, and participants were shown a feedback animation of the path of the ball (see Stimuli). The final frame of this animation remained on the screen until participants pressed "space" to advance to the next trial.

Participants were given seven instruction trials prior to the experiment to familiarize them with the procedure. Then, participants made judgments on 48 experimental trials in a random order. There were also eight control trials, which were shown in a random order after every seven experiment trials.

Stimuli

The stimuli consisted of two animations—the *stimulus presentation* and the *feedback* animations—depicting a blue ball with a radius of 10px moving in a box with dimensions 900px × 650px. In both animations, the ball had a velocity of 400px/s, and as it moved, it traced a gray line (see Figure 1). The stimulus presentation had a duration of 0.775s and depicted the ball moving in a particular direction. The feedback had a duration of 1.5s and picked up where the stimulus presentation left off; it showed the ball either going into the hole or bouncing off the wall that contained the hole for. Across all stimuli, the ball traveled the same distance during both animations, and could bounce on the other walls 0, 1, or 2 times before going into the hole or hitting the wall with the hole.

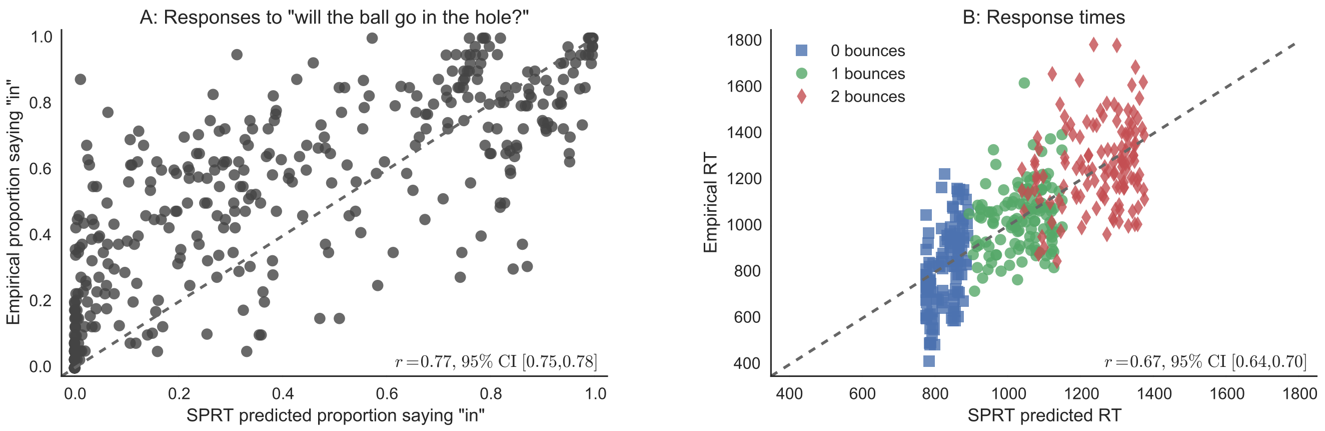


Figure 3: **Model vs. human comparison.** In both plots, each point corresponds to a different stimulus, trial type, and hole size. Dashed lines indicate perfect correspondence between model and people. *A:* The *x*-axis is the probability the model says the ball will go in the hole, and the *y*-axis is the proportion of participants saying the ball will go in the hole. *B:* Color and shape indicate the number of times the ball bounced during feedback. The *x*-axis is the model RTs, and the *y*-axis is the logarithmic mean RTs of participants.

There were 48 different initial animations, equally balanced by number of bounces during feedback (16 each for 0, 1, and 2 bounces). For each of these initial animations, there were four trial types and two hole sizes, for a total of eight versions of each stimulus. The four trial types were: "far in" (FI), where the ball went through the center of the hole; "far miss" (FM), where the ball missed the hole by a wide margin; "close in" (CI), where the ball just barely went through the hole; and " close miss" (CM), where the ball just barely missed the hole. The two hole sizes were 100px and 200px.

In order to ensure that participants never saw the same initial animation twice, we used a Latin square design of Initial Animation × Trial Type × Hole Size. Thus, each participant saw each initial animation once, each trial type 12 times, and each hole size 24 times. This also ensured that the ball would go through the hole half the time, so that participants would not be biased to respond either way. Additionally, there were seven instruction trials and eight control trials, which were the same for all participants. The control trials were designed to be easy and were either of type "straight hit" (with a hole size of either 300px or 350px) or "far miss" (with a hole size of 100px). Thus, participants saw a total of 63 trials.

Results

Responses On average, participants were correct 72.4% of the time and responded that the ball would go in the hole 53.2% of the time ($N = 15216$), excluding catch trials. There was a significant effect of trial type on participants' responses ($\chi^2(3) = 4477.182, p < 0.001$) as well as a significant effect of hole size ($\chi^2(1) = 168.598, p < 0.001$). There was also an interaction between trial type and hole size ($\chi^2(3) = 64.469, p < 0.001$). There was a significant difference between responses for the two hole sizes (for CI, $p < 0.001$; for FI, $p < 0.001$; and for FM, $p < 0.001$) except on the CM trials ($p = 0.45$). Figure 2A shows responses as a function of trial type and hole size.

Response times For all analyses of response time (RT), we computed averages using bootstrapped logarithmic means (exponential of the mean of the log RTs), using 10000 bootstrap samples. On average, participants responded in $RT = 1009.88$ msec, 95% CI $[998.48, 1022.41]$, excluding catch trials. There were effects of both trial type ($\chi^2(3) = 63.611, p < 0.001$) and hole size ($\chi^2(1) = 8.981, p < 0.01$), as well as an interaction between trial type and hole size ($\chi^2(3) = 27.146, p < 0.001$). As in Figure 2B, hole size only had an effect in the case of the FI trials ($p < 0.001$; for CI, $p = 0.11$; for CM, $p = 0.37$; and for FM, $p = 0.71$).

Participants were fastest to respond on trials with zero bounces ($RT = 799.76$ msec, 95% CI $[783.01, 817.10]$), slower to respond on trials with one bounce ($RT = 1027.48$ msec, 95% CI $[1006.92, 1048.72]$), and slowest to respond on trials with two bounces ($RT = 1251.57$ msec, 95% CI $[1226.62, 1277.71]$).

Relationship of responses and RTs According to SPRT, participants should be slower on trials for which they are less certain (i.e., when their average response is closer to 0.5), and faster when they are more certain (i.e., when their average response is closer to 0 or 1). Figure 2C illustrates that this trend does indeed appear. To demonstrate this more quantitatively, we fit both 1st- and 2nd- order polynomial functions to the relationship between responses and RTs. The 1st-order function had AIC = 5326 and BIC = 5334, while the 2nd-order function had AIC = 5065 and BIC = 5077.

Participants' responses do not fully account for their RTs, however: the number of bounces is also a strong predictor of RT. Although people do make more variable predictions as bounces are added (Smith & Vul, 2013), and more variable trials have longer RTs, there appears to be an additional time cost. We modified the 2nd-order polynomial

Model	Bounces	Correlation
$T = 1$	0	$r = 0.09$, 95% CI $[-0.00, 0.18]$
	1	$r = -0.05$, 95% CI $[-0.16, 0.07]$
	2	$r = 0.05$, 95% CI $[-0.06, 0.16]$
	all	$r = 0.66$, 95% CI $[0.62, 0.69]$
$T = 2$	0	$r = 0.45$, 95% CI $[0.37, 0.53]$
	1	$r = 0.19$, 95% CI $[0.07, 0.29]$
	2	$r = 0.18$, 95% CI $[0.07, 0.28]$
	all	$r = 0.67$, 95% CI $[0.64, 0.70]$

Table 1: **Correlations between human and model RTs.** The SPRT model ($T = 2$) can capture variations in RTs within bounce conditions, while the single sample model ($T = 1$) cannot. To compute the correlations within bounce conditions, we grouped stimuli by the number of bounces shown in the feedback, but within bounce conditions we estimate $\mathbb{E}[B]$ from the simulation model. Thus, the expected number of bounces varies slightly across stimuli, which is why the correlations for the $T = 1$ model are non-zero.

to include the number of bounces as an regressor, and found that the number of bounces is a strong predictor of RT above and beyond the responses (with bounces, AIC = 4915 and BIC = 4931; without bounces, AIC = 5065 and BIC = 5077). From the coefficient, we find that the each bounce adds $RT =$ 150.38 msec, 95% CI $[128.61, 172.14]$. Thus, even if people are using a SPRT-like strategy, there is a time cost per bounce that cannot be explained by simulation variance alone. This suggests that there is a discrepancy between our simulation model and the manner in which people are *actually* running simulations; investigating the details of this discrepancy is an area for future work.

Learning To check for practice effects, we computed Spearman rank correlations (with 95% confidence intervals computed from 10000 bootstrap samples) between trial number and accuracy, as well as between trial number and RT. We found an overall effect of practice on accuracy ($\rho = 0.27$, 95% CI $[0.07, 0.45]$), though in the second half of the experiment, this effect disappeared ($\rho = -0.08$, 95% CI $[-0.38, 0.22]$). There was also an overall effect of practice on RT ($\rho = -0.89$, 95% CI $[-0.93, -0.84]$), which was strong both in the first ($\rho = -0.88$, 95% CI $[-0.94, -0.82]$) and second ($\rho = -0.66$, 95% CI $[-0.83, -0.41]$) halves of the experiment. Future work will need to include a longer training period to alleviate these practice effects.

Model comparison If we assume that every sample takes the same amount of time, RTs as predicted by the model should be directly proportional to $\mathbb{E}[N \mid T, p]$. However, because bounces were a strong predictor of RT, we also incorporated the number of bounces, B, according to $RT = \beta_0 + (\beta_1 + \beta_2 \cdot \mathbb{E}[B]) \cdot \mathbb{E}[N \mid T, p]$. Briefly, β_0 is the time to set up the simulation(s) and to respond, β_1 is the time to simulate with no bounces, and β_2 is the time needed to simulate each bounce. We used the physical simulation model to determine $\mathbb{E}[B]$ as the average number of times the ball bounced across all model simulations. We then fit all parameters (T, σ_{adj}, β_0, β_1, and β_2) to minimize sum squared error between modeled and observed RTs, using 10000 samples from the physical simulation model. The best fitting values were: $T = 2$, $\sigma_{adj} = 0.9$,[1] $\beta_0 = 684.02$ msec, 95% CI $[601.74, 766.31]$, $\beta_1 = 46.00$ msec, 95% CI $[19.67, 72.33]$, and $\beta_2 = 63.59$ msec, 95% CI $[57.18, 70.00]$.

We computed Pearson correlations between the model and people with 95% confidence intervals computed from 10000 bootstrap samples. The fitted model explains participants' judgments of whether the ball would go in the hole very well ($r = 0.77$, 95% CI $[0.75, 0.78]$, see Figure 3A), and is also a good predictor of RT ($r = 0.67$, 95% CI $[0.64, 0.70]$). A model that takes one sample each time (equivalent to SPRT with $T = 1$) is slightly better at explaining people's responses ($r = 0.80$, 95% CI $[0.78, 0.81]$), and is equally good at explaining overall RT in terms of correlation ($r = 0.66$, 95% CI $[0.62, 0.69]$). However, according to both BIC and AIC, the $T = 1$ model is slightly worse (BIC = 5040 and AIC = 5032) than the full model (BIC = 5024 and AIC = 5012) at explaining RTs. Moreover, the $T = 1$ model cannot explain variance in RTs within bounces, whereas the full model with $T = 2$ can (see Table 1). If the number of bounces is excluded, the model with $T = 2$ can predict human RTs to a moderate degree ($r = 0.32$, 95% CI $[0.28, 0.36]$), while the model with $T = 1$ cannot predict RTs at all.

Discussion

In this paper, we asked the question: do people optimally use mental simulations? We hypothesized that people use noisy physical simulations to predict whether a ball would go in a hole, and that they vary the number of simulations in order to exploit the fact that some judgments are easier to make than others. The results of our experiment paint a clear picture that people *do* vary the number of samples they take, as evidenced by the increase in response time on the trials they were most uncertain about. Comparing people's responses and response times to those of the model, we found a strong fit. This provides evidence that people not only rely on approximate physical simulations, but that they vary the number of simulations that they run according to SPRT.

If SPRT is the optimal strategy, then what is the optimal threshold? We found the best fitting SPRT threshold to be $T = 2$, which is consistent with previous research. According to Vul, Goodman, Griffiths, and Tenenbaum (2014), a sample-based agent should only take a small number of samples before making a judgment so long as there is any cost to taking samples. While taking a small number of samples

[1] If $\sigma_{adj} = 0.9$, then $M < 1$. How could people be taking less than one sample? We suspect that $\sigma_{adj} < 1$ because the model overestimates the standard deviation of the ball's trajectory; thus, it is likely that people are taking one sample, and σ_{adj} is adjusting for inflated uncertainty in the model.

provides a worse chance of making a good decision than taking multiple samples, over the long run, this strategy maximizes expected utility across a large number of judgments. There has been some evidence that this story also holds true for mental simulations. For example, Battaglia et al. (2013) analyzed the variability of people's responses in tasks concerning towers of building blocks, and found that participants seemed to use between three and seven samples per judgment. However, this paper is the first to provide evidence not only that people use a small number of simulations, but that they vary the number of simulations in response to task demands.

Mental simulation is a powerful and flexible tool, as it offers a way to make predictions about scenarios that have not yet (or may never) come to pass. In this work, we demonstrated that when people use mental simulation, they are sensitive to their own uncertainty in reasoning about the task and accordingly adjust how many simulations they run. This results joins others (e.g. Hamrick & Griffiths, 2014) in explaining not just *that* people use simulation to reason about the world, but *how* they use it. While there are still many questions left unanswered—e.g., how do people use simulations in non-binary tasks?—this work brings us one step closer to understanding of how mental simulation is used.

Acknowledgments

This work was supported by a NSF GRF awarded to JBH, a UCSD Chancellors Interdisciplinary Collaboratories grant awarded to KAS, NSF CPS grant 1239323, and grant number N00014-13-1-0341 from the Office of Naval Research.

Appendix: Replication of Smith and Vul (2013)

Participants We recruited $N = 60$ participants using psiTurk (McDonnell et al., 2014). Participants were treated in accordance with UC Berkeley IRB standards and were paid $0.60 for 5 minutes of work. We excluded $N = 18$ people for failing to catch the ball on more than one control trial.

Stimuli The stimuli were modified versions of those used in the main experiment, with two differences. First, instead of a wall with a hole in it, there was a paddle of length 100px that could move up and down the y-axis. Second, instead of a full feedback animation, we just displayed the last frame. Because there was no hole that could vary by trial, there were only 48 stimuli, plus seven instruction and eight catch trials.

Procedure Like the main experiment, there were two phases: the training phase and the experimental phase. On each trial, participants were shown the scene, including the initial position of the ball. The paddle begin at the center of the y-axis, and was freely movable at the start of the trial. Participants were instructed to press "space" to begin the trial and display the stimulus animation. After the stimulus presentation, a gray occluder appeared, as well as a timer that began counting down for 2 seconds. During this time, participants had to move the paddle to catch the ball in the position it would be when the timer was up. When the timer finished, the paddle froze, the occluder was removed, and the full path of the ball was revealed. Participants were told whether they caught the ball or not, and then instructed to press "space" to begin the next trial.

Results We fit the model parameters of σ_p, κ_v, κ_m, κ_b, and σ_0 to participant's responses (for details, see Smith & Vul, 2013), finding the best fitting parameters to be $\sigma_p = 31.02$, $\kappa_v = 255.60$, $\kappa_m = 502850.41$, $\kappa_b = 50.42$, and $\sigma_0 = 167.06$. With these parameters, we found very similar results to those from Smith and Vul (2013). In particular, we found a correlation of $r = 0.94$, 95% CI $[0.91, 0.97]$ between the model's predicted means of where the ball would end up and people's average location of the paddle.

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 of the United States of America*, *110*(45), 18327–18332.

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, *8*(102), 1–17.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (3rd ed., Vol. 1). John Wiley & Sons Incorporated.

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*, 535–574.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (in prep.). Inferring mass in complex scenes via probabilistic simulation.

Hamrick, J. B., & Griffiths, T. L. (2014). What to simulate? Inferring the right direction for mental rotation. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., & Gureckis, T. (2014). psiTurk (Version 2.00) [Computer software manual]. New York, NY. Retrieved from https://github.com/NYUCCL/psiTurk

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922.

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

Smith, K. A., Battaglia, P. W., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Smith, K. A., & Vul, E. (2013). Sources of Uncertainty in Intuitive Physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2014). Learning physics from dynamical scenes. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, *38*(4), 599–637.

Wald, A. (1947). *Sequential analysis*. Wiley, New York.

Wald, A., & Wolfowitz, J. (1950). Bayes solutions of sequential decision problems. *The Annals of Mathematical Statistics*, *21*(1), 82–99.