What to simulate? Inferring the right direction for mental rotation

Jessica B. Hamrick (jhamrick@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

When people use mental imagery, how do they decide *which* images to generate? To answer this question, we explored how mental simulation should be used in the classic psychological task of determining if two images depict the same object in different orientations (Shepard & Metzler, 1971). Through a rational analysis of mental rotation, we formalized four models and compared them to human performance. We found that three models based on previous hypotheses in the literature were unable to account for several aspects of human behavior. The fourth is based on the idea *active sampling* (e.g., Gureckis & Markant, 2012), which is a strategy of choosing actions that will provide the most information. This last model provides a plausible account of how people use mental rotation, where the other models do not. Based on these results, we suggest that the question of "what to simulate?" is more difficult than has previously been assumed, and that an active learning approach holds promise for uncovering the answer.

Keywords: mental rotation, computational modeling

Introduction

One of the most astonishing cognitive feats is our ability to envision, manipulate, and plan with objects—all without actually perceiving them. This *mental simulation* has been widely studied, including an intense debate about the underlying representation of mental images (e.g., Kosslyn, Thompson, & Ganis, 2009; Pylyshyn, 2002). But this debate hasn't addressed one of the most fundamental questions about mental simulation: how people decide *what* to simulate.

Mental rotation provides a simple example of the decision problem posed by simulation. In the classic experiment by Shepard and Metzler (1971), participants viewed images of three-dimensional objects and had to determine whether the images depicted the same object (which differed by a rotation) or two separate objects (which differed by a reflection and a rotation). They found that people's response times (RTs) had a strong linear correlation with the minimum angle of rotation, a result which led to the conclusion that people solve this task by "mentally rotating" the objects until they are congruent. However, this explanation leaves several questions unanswered. How do people know the axis around which to rotate the objects? If the axis is known, how do people know which direction to rotate the objects? And finally, how do people know how long to rotate?

In this paper, we explore these questions through rational analysis (Marr, 1983; Anderson, 1990; Shepard, 1987) and compare four models of mental rotation. We begin the paper by discussing the previous literature on mental imagery. Next, we outline computational- and algorithmic-level analyses of the problem of mental rotation. We then describe a behavioral experiment based on the classic mental rotation studies (e.g., Cooper, 1975), and compare the results of our experiment with each of the models. We conclude with a discussion of the strengths and weaknesses of each model, and lay out directions for future work.

Modeling mental rotation

Previous models of mental rotation have largely focused on the representation of mental images, rather than how people decide *which* mental images to generate. Kosslyn and Shwartz (1977) proposed a model of the mental imagery buffer, but did not say *how* it should be used. Similarly, Julstrom and Baron (1985) and Glasgow and Papadias (1992) were mostly concerned with modeling the representational format underlying imagery. Although Anderson (1978) emphasized the importance of considering both representation and process, he dismissed the problem of determining the direction of rotation as a "technical difficulty".

The only models (of which the authors are aware) that seriously attempted to address the decision of *what* to simulate are those by Funt (1983) and Just and Carpenter (1985). In both of these models, the axis and direction of rotation are computed prior to performing the rotation. One object is then rotated through the target rotation, and is checked against the other object for congruency. However, this approach assumes that the corresponding points on the two objects can be easily identified, which is not necessarily the case. Indeed, the state-of-the-art in computer vision suggests that there is more to this problem than checking for congruency, particularly when the shapes are complex or not exactly the same (e.g., Belongie, Malik, & Puzicha, 2002; Sebastian, Klein, & Kimia, 2003). Additionally, recent research shows that when performing *physical* rotations, people do not rotate until congruency is reached; they may even rotate *away* from near perfect matches (Gardony, Taylor, & Brunye, 2014).

If people are not computing the rotation beforehand, what might they be doing? To answer this question, we perform a rational analysis of the problem of mental rotation (Marr, 1983; Anderson, 1990; Shepard, 1987). At the computational level, we can say that the *problem* is to determine which spatial transformations an object has undergone based on two images of that object (which do not include information about point correspondences). At the algorithmic level, we are constrained by the notion that mental images must be transformed in an analog manner (or in a way that is approximately analog), and that mental images are time-consuming and effortful to generate. Thus, the *goal* is to make this determination while performing a minimum amount of computation (i.e., as few rotations as possible).

The original "congruency" hypothesis (Shepard & Metzler, 1971) is a rational solution to this problem, in the sense

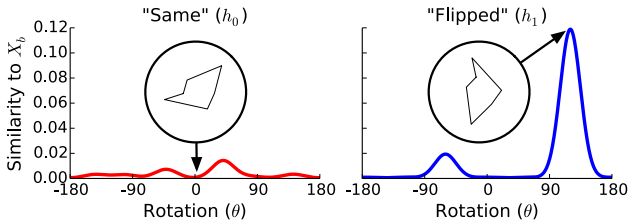

Figure 1: **Example stimuli and similarities.** This figure shows a "flipped" stimulus pair with a rotation of 120°, and the corresponding similarity functions for each hypothesis. Arrows indicate where each shape lies on the curve.

that the smallest amount of computation coincides with rotating through the minimum angle. However, it violates the constraint that we do not know the points of correspondence between the images, which is what necessitates the use of imagery. Noting that a rational solution need not maintain a single trajectory of rotation, we explore an alternative model, which—rather than computing the angle of rotation—engages in an *active sampling* strategy.

Active sampling is the idea that people gather new information in a manner that increases certainty about the problem space. An everyday example of this can be observed in the game of "20 questions", in which one person thinks of a concept, and another has to guess the concept in 20 questions or less. The first question is almost always "person, place, or thing?", because the answer provides the most possible information about the concept of interest. Active sampling has gained support across several areas of cognitive science (e.g., Gureckis & Markant, 2012), including other spatial domains (Juni, Gureckis, & Maloney, 2011). In the case of mental rotation, actively choosing rotations may be the best way to gather evidence about the similarity between the observed shapes when the angle of rotation is unknown.

How should we rotate?

In this section, we formalize our rational analysis and propose four models of mental rotation: one based on existing models; two which are extensions of the first but with relaxed assumptions; and one based on the active sampling approach.

The task we are interested in modeling involves observing two images and determining whether one image depicts the "same" object as the other image (differing by a rotation), or a "flipped" version of the object in the other image (differing by a reflection and then a rotation).

Computational-level analysis

We denote the shapes as X_a and X_b and assume X_b is generated by a transformation of X_a, i.e. $X_b = f(X_a, \theta, h)$, where θ is a rotation, $h = 0$ is the hypothesis that the images depict the same object, and $h = 1$ is the hypothesis that the images depict mirror-image objects. The posterior probability of each hypothesis given the observed shapes is then: $p(h \mid X_a, X_b) \propto \int p(X_b \mid X_a, \theta, h) p(h) p(\theta)\, d\theta$, where $p(X_b \mid X_a, \theta, h)$ is the probability that X_b was generated from X_a. Because we want to determine which hypothesis is more likely, the quantity of interest is a posterior odds ratio $\mathcal{B} := p(h=0 \mid X_a, X_b)/p(h=1 \mid X_a, X_b)$ which (assuming that all rotations are equally likely) is equivalent to:

$$\mathcal{B} = \frac{(\int p(X_b \mid X_a, \theta, h=0)\, d\theta) \cdot p_0}{(\int p(X_b \mid X_a, \theta, h=1)\, d\theta) \cdot p_1}, \qquad (1)$$

where $p_0 = p(h=0)$ and $p_1 = p(h=1)$, for brevity. If $\mathcal{B} > 1$, then we accept the hypothesis that the images depict the same object ($h = 0$); if $\mathcal{B} < 1$, then we accept the hypothesis that the images depict flipped objects ($h = 1$).

Algorithmic constraints

We represent a shape of N vertices with a $N \times 2$ coordinate matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, and denote the rotation and/or reflection transformation as $f(\mathbf{X}, h, \theta) := \mathbf{X}\mathbf{F}_h^T \mathbf{R}_\theta^T$, where $\mathbf{R}_\theta$ is a rotation matrix, and $\mathbf{F}_h$ is either the identity matrix $\mathbb{I}$ (when $h = 0$) or a reflection matrix across the y-axis (when $h = 1$).

We define $p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h)$ to be the similarity between $\mathbf{X}_b$ and a transformation of $\mathbf{X}_a$: $p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h) := S(\mathbf{X}_b, f(\mathbf{X}_a, h, \theta))$. We do not know which vertices of $\mathbf{X}_b$ correspond to which vertices of $\mathbf{X}_a$, so the similarity S must marginalize over the set of possible mappings. For brevity, let $\mathbf{X}_m = \mathbf{M} \cdot f(\mathbf{X}_a, h, \theta)$ where $\mathbf{M}$ is a permutation matrix. Then:

$$S(\mathbf{X}_b, f(\mathbf{X}_a, h, \theta)) := \frac{1}{2N} \sum_{\mathbf{M}} \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_{bn} \mid \mathbf{x}_{mn}, \mathbb{I}\sigma_S^2), \qquad (2)$$

where $2N$ is the total number of possible mappings,[1] and $\sigma_S^2 = 0.15$ is the variance of the similarity. Example similarity curves are shown in Figure 1.

We assume that the observed shapes must be transformed by a small amount at a time, and each transformation takes a non-negligible amount of time. If the current mental image is $\mathbf{X}_t$, then:

$$\mathbf{X}_{t+1} = \begin{cases} f(\mathbf{X}_t, 0, \varepsilon) & \text{rotate by } \varepsilon \text{ radians,} \\ f(\mathbf{X}_t, 1, 0) & \text{flip,} \\ f(\mathbf{X}_a, 0, 0) & \text{reset to } 0^\circ, \text{ or} \\ f(\mathbf{X}_a, 1, 0) & \text{reset and flip,} \end{cases} \qquad (3)$$

where $\varepsilon \sim \left|\mathcal{N}(0, \sigma_\varepsilon^2)\right|$ and σ_ε^2 is the variance of the step size.

To summarize, we approximate the likelihood term of Equation 1 using the similarity function defined in Equation 2. Because we assume mental rotations are performed sequentially, this similarity can only be computed for the actions listed in Equation 3.

Specific models of mental rotation

In order to approximate Equation 1 using samples of the similarity function, we must decide *which* places to sample and *when* stop sampling. The models below differ in how they make these decisions.

[1] It is $2N$ and not N^2 because, in polar coordinates, vertices are always connected to their two nearest neighbors in the θ dimension.

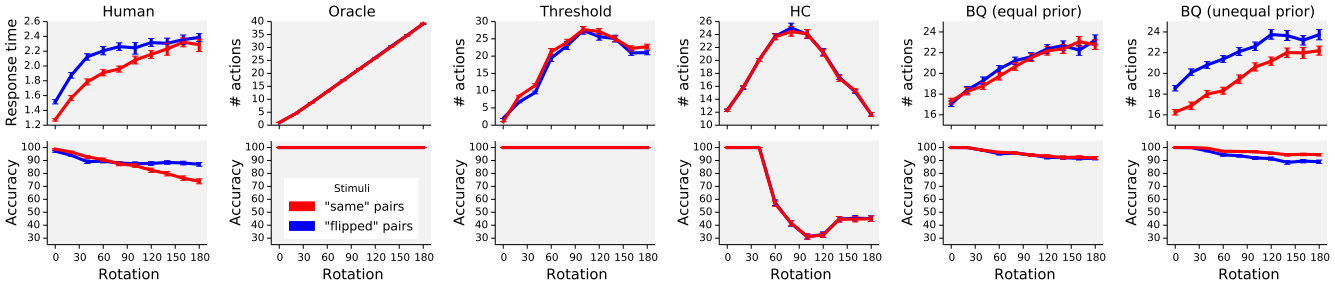

Figure 2: **Response time and accuracy comparison.** Top: RT of correct responses as a function of the minimum angle of rotation. Bottom: accuracy as a function of the minimum angle of rotation. All error bars are 95% confidence intervals.

Oracle model One hypothesis is that people compute the direction and extent of rotation beforehand using *a priori* knowledge of the correspondence between points in the images (Funt, 1983; Just & Carpenter, 1985). To reflect this hypothesis, we created an "oracle" model which is told which points on each shape correspond. From that correspondence, it computes the correct rotation and rotates through it.

To determine the correct rotation, we solve for the rotation matrix by computing $(\mathbf{X}_a\mathbf{F}_h^T)^{-1}_{\text{left}} \cdot \mathbf{X}_b$, where $(\mathbf{X}_a\mathbf{F}_h^T)^{-1}_{\text{left}}$ is the left inverse of $\mathbf{X}_a\mathbf{F}_h^T$. We then check each h to see if the computation produces a valid rotation matrix; the h that does is the correct hypothesis. This gives us the true value of θ, so Equation 1 becomes a generalized likelihood ratio test, where θ is set to the MLE value, rather than being marginalized:

$$\mathcal{B} = \frac{\max_\theta p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h=0) \cdot p_0}{\max_\theta p(\mathbf{X}_b \mid \mathbf{X}_a, \theta, h=1) \cdot p_1}. \tag{4}$$

If we give equal weight to the two hypotheses, then the priors cancel out; if we weigh one hypothesis more heavily, then our decision will be biased towards that hypothesis. However, unless the likelihood ratio is already very close to 1, small biases in the prior will not make much of a difference.

Threshold model A model which does not know point correspondences could use the following algorithm: (1) pick a random direction; (2) take a single step; (3) if that step decreased similarity, then begin rotating in the reverse direction, otherwise continue rotating in the original direction; (4) continue rotating in the chosen direction until a "match" is found (defined as finding a value of S that exceeds a threshold); and (5) if no match was found, flip, and start over from step one. We only allow for the "flip" action after no match has been found, because there is no particularly principled way for the Threshold model to choose when to flip. We assume that the locations where S is greater than the threshold correspond to the true θ (or points near the true θ). So, as with the Oracle model, we use Equation 4.

Hill Climbing model In the current formulation of the problem, choosing the threshold is straightforward because we know both the exact geometry of the shapes and that a linear transformation exists which will align them. However, this choice is not always clear *a priori*, as the global optimum depends on many factors (e.g., shape complexity, dimensionality, perceptual uncertainty, and whether the shapes are identical). One way to deal with the problem of choosing a threshold would be use a global optimization strategy; however, this would not result in the linear RT found by Shepard and Metzler (1971). A second alternative is to perform a Hill Climbing (HC) search; i.e., rotate in the direction that increases similarity until no further improvement can be found. In contrast with the Threshold model, this results in arriving in a *local* maximum (which may or may not be the global maximum). Thus, as with the Oracle and Threshold models, we use Equation 4. We only allow for the "flip" action after a local maximum has been reached, because like the Threshold model, there is otherwise no principled way for the HC model to choose when to flip.

Bayesian Quadrature model While the previous few models all focused on *searching* for the global maximum, we need only *approximate* Equation 1. We hypothesize a model based on the idea of *active sampling* (e.g., Gureckis & Markant, 2012): instead of searching for a maximum, we maintain a probability distribution over our *estimate* of Equation 1, and then sample actions which are expected to improve that estimate. This strategy has the benefits that it does not make assumptions about the scale of the similarity function; and, by choosing to sample places which are informative, this method implicitly minimizes the amount of rotation.

We denote Z_h as our estimate of the likelihood for hypothesis h, and write its distribution as: $p(Z_h) = \int [\int S(\mathbf{X}_b, f(\mathbf{X}_a, \theta, h)) p(\theta)\, d\theta]\, p(S)\, dS$, where S is the similarity function, and $p(S)$ is a prior over similarity functions. This method of estimating an integral is known in the machine-learning literature as *Bayesian Quadrature* (Diaconis, 1988; Osborne et al., 2012), or BQ. Denoting $S_h = S(\mathbf{X}_b, f(\mathbf{X}_a, \theta, h))$, we first place a *Gaussian Process* (Rasmussen & Williams, 2006), or GP, prior on the log of S_h in order to enforce positivity after it is exponentiated, i.e. $\mathbb{E}[Z_h] \approx \int \exp(\mu_h(\theta)) p(\theta)\, d\theta$, where $\mu_h := \mu(\log S_h)$ is the mean of the log-GP (Osborne et al., 2012). To approximate this integral, we fit a second GP over points sampled from the log-GP, which we denote as $\bar{S}_h := \exp(\mu_h)$. Then, from Duvenaud (2013), we have $\mathbb{E}[Z_h] \approx \int \bar{\mu}_h(\theta) p(\theta)\, d\theta$ and $\mathbb{V}(Z_h) \approx \iint \text{Cov}_h(\theta, \theta') \bar{\mu}_h(\theta) \bar{\mu}_h(\theta') p(\theta) p(\theta')\, d\theta\, d\theta'$, where $\bar{\mu}_h := \mu(\bar{S}_h)$ is the mean of the second GP, and $\text{Cov}_h :=$

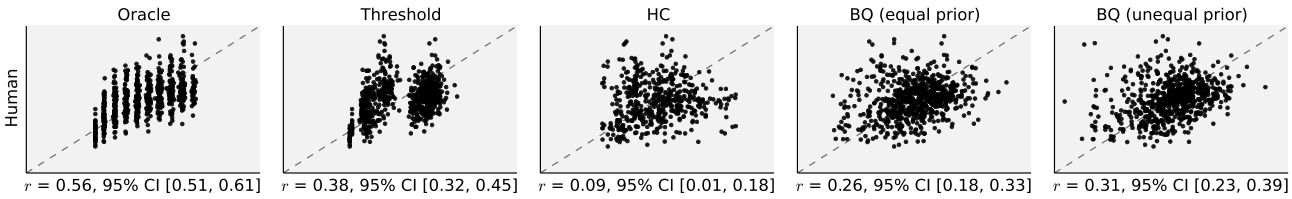

Figure 3: **Model vs. human RTs.** Each subplot shows the z-scored model RTs (x-axis) vs. the z-scored human RTs (y-axis). Pearson correlations are shown beneath each subplot. The dotted lines are $x = y$.

$\mathrm{Cov}(\log S_h)$ is the covariance of the log-GP.

Assuming independence, we can now write $p(Z_h) \approx \mathcal{N}(Z_h \mid \mathbb{E}[Z_h], \mathbb{V}(Z_h))$, which gives us a distribution over the likelihood ratio in Equation 1: $p(\mathcal{B}) \approx \mathcal{N}(Z_0 \mid \mathbb{E}[Z_0], \mathbb{V}(Z_0)) \cdot p_0 / \mathcal{N}(Z_1 \mid \mathbb{E}[Z_1], \mathbb{V}(Z_1)) \cdot p_1$. This distribution cannot easily be calculated, but we are only interested in whether $Z_0 > Z_1$ or $Z_1 > Z_0$. So, we use $Z_D = p_0 \cdot Z_0 - p_1 \cdot Z_1$ and compute $p(Z_D) \propto \mathcal{N}(p_0 \cdot \mathbb{E}[Z_0] - p_1 \cdot \mathbb{E}[Z_1], p_0^2 \cdot \mathbb{V}(Z_0) + p_1^2 \cdot \mathbb{V}(Z_1))$. We then sample new observations until we are at least 95% confident that $Z_D \neq 0$. In other words, when $p(Z_D < 0) < 0.025$, we accept $h = 0$, and when $p(Z_D < 0) > 0.975$, we accept $h = 1$. Because we compare the hypotheses in order to determine when to stop sampling, biasing the prior should result in requiring less evidence for one hypothesis before stopping, and more evidence for the other hypothesis.

To choose where to sample, we compute the expected variance of Z_h given a new observation at θ_a. From Osborne et al. (2012), we compute $\mathbb{E}[\mathbb{V}(Z_h|\theta_a)] = \mathbb{V}(Z_h) + \mathbb{E}[Z_h] - \int \mathbb{E}[Z_h|\theta_a]^2 \mathcal{N}(\mu_h(\theta_a), \mathrm{Cov}_h(\theta_a, \theta_a))\, \mathrm{d}\log S_h(\theta_a)$ for each of the actions in Eq. 3; we pick the one with the lowest value.

Methods

To evaluate the models described previously, we ran a behavioral experiment based on classic mental rotation studies (e.g. Shepard & Metzler, 1971; Cooper, 1975).

Stimuli We randomly generated 20 shapes of five or six vertices (e.g., Figure 1). For each shape, we computed 20 "same" and 20 "flipped" stimuli pairs, with 18 rotations (θ) spaced at 20° increments between 0° and 360° (with 0° and 180° repeated twice, in order to gather an equal number of responses for each angle between 0° and 180°). "Same" pairs were created by rotating $\mathbf{X}_a$ by θ; "flipped" pairs were first reflected $\mathbf{X}_a$ across the y-axis, then rotated by θ.

We generated five additional shapes to be used in a practice block of 10 trials. Across these trials, there was one "flipped" and one "same" repetition of each shape and each angle (60°, 120°, 180°, 240°, or 300°) such that no shape was presented at the same angle twice. We also generated a sixth shape to include with the instructions. This shape had both a "flipped" and "same" version, each rotated to 320°.

Participants and Design We recruited 247 participants on Amazon's Mechanical Turk using the psiTurk experiment framework (McDonnell et al., 2012). Each participant was paid $1.00 for 15 minutes of work, consisting of one block of 10 practice trials followed by two blocks of 100 randomly ordered experiment trials.

All participants saw the same 10 practice trials as described above. There were 720 unique experimental stimuli (20 shapes × 18 angles × 2 reflections), though because stimuli with rotations of 0° or 180° were repeated twice, there were 800 total experimental stimuli. These stimuli were split across eight conditions in the following manner: first, stimuli were split into four blocks of 200 trials. Within each block, each shape was repeated ten times and each rotation was repeated ten times (five "same", five "flipped"), such that across all blocks, each stimulus appeared once. Each block was then split in half, and participants completed two half-blocks.

Procedure Participants were given the following instructions while being shown an example "same" pair and an example "flipped" pair: *"On each trial, you will see two images. Sometimes, they show the **same** object. Other times, the images show **flipped** objects. The task is to determine whether the two images show the **same** object or **flipped** objects."*

On each trial, participants were instructed to press the 'b' key to begin and to focus on the fixation cross that appeared for 750ms afterwards. The two images were then presented side-by-side, each at 300px × 300px, and participants could press 's' to indicate they thought the images depicted the "same" object, or 'd' to indicate they thought the images depicted "flipped" objects. While there was no limit on RT, we urged participants to answer as quickly as possible while maintaining at least 85% accuracy in the experimental blocks.

Results

Of the 247 participants, 200 (81%) were included in our analyses. Of the other 47, we excluded 10 (4%) because of an experimental error, 6 (2.4%) because they had already completed a related experiment, and 31 (12.6%) because they failed a comprehension check, which was defined as correctly answering at least 85% of stimuli with a rotation of either 0°, 20°, or 340°. We also excluded 82 trials for which the RT was either less than 100ms or greater than 20s.

For each model, we ran 50 samples for each of the 800 experimental stimuli. The step size parameter (σ_ε) was fit to human RTs for each of the models, resulting in $\sigma_\varepsilon = 0.6$ for the Threshold and BQ models and $\sigma_\varepsilon = 0.1$ for the Oracle and HC models. We also ran the models under two different priors, $p(h = 0) = 0.5$ (the "equal" prior) and $p(h = 0) = 0.55$ (the "unequal" prior). As expected, this only had a major effect on the stopping criteria for the BQ model.

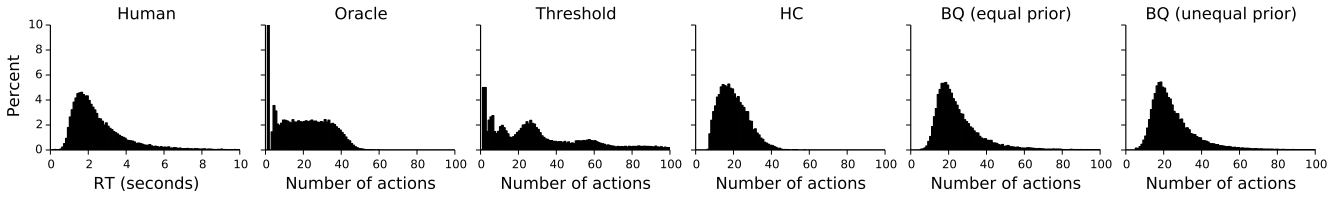

Figure 4: **Response time histograms.** Each subplot shows the distribution of RTs on correct trials for people and the models.

General analysis For analyses of RT, confidence intervals around harmonic means of correct responses were computed using a bootstrap analysis of 10000 bootstrap samples (sampled with replacement). We also used a bootstrap analysis of 10000 bootstrap samples to compute the confidence intervals around both Spearman (ρ) and Pearson (r) correlations. Unless otherwise specified, all correlations were computed over 720 stimuli. For analyses of accuracy, confidence intervals were computed from a binomial proportion with a Jeffrey's beta prior. To test if judgments were above chance on a particular stimulus, we used the same binomial proportion and tested whether $p\left(p(\text{correct}) \leq 0.5\right) \leq \frac{0.05}{720}$, where $\frac{1}{720}$ is a Bonferroni correction for multiple comparisons.

Human The average RT across all correctly-judged stimuli was $M = 1981.1$ msec, 95% CI [1969.4 msec, 1992.1 msec]; the full histogram of RTs can be seen in Figure 4. The minimum angle of rotation was significantly rank-order (Spearman) correlated with average per-stimulus RTs, both for "flipped" ($\rho = 0.49$, 95% CI $[0.40, 0.57]$) and "same" pairs ($\rho = 0.66$, 95% CI $[0.60, 0.72]$). While this replicates the general result of previous experiments (e.g., Shepard & Metzler, 1971; Cooper, 1975), our results are not as linear (Figure 2).

The average accuracy across all stimuli was $M = 88.1\%$, 95% CI $[87.8\%, 88.4\%]$, though there were 64 stimuli (out of 720) for which people were not above chance. The minimum angle was also correlated with participants' average per-stimulus accuracy, though much more so for "same" pairs ($\rho = -0.77$, 95% CI $[-0.81, -0.72]$) than "flipped" pairs ($\rho = -0.36$, 95% CI $[-0.46, -0.26]$). This is the same result found both by Cooper (1975) and Gardony et al. (2014).

There was a significant effect of trial number both on RT ($\rho = -0.76$, 95% CI $[-0.82, -0.68]$) and on accuracy ($\rho = 0.66$, 95% CI $[0.58, 0.74]$), though the effect on accuracy was not significant during the second half of the experiment ($\rho = 0.50$, 95% CI $[0.33, 0.65]$ for the first half vs. $\rho = 0.16$, 95% CI $[-0.03, 0.34]$ for the second half). These effects may have contributed to the not-quite-linearity of the human RTs; future work should collect more data per participant.

Oracle model The number of actions taken by the Oracle model was perfectly correlated with the minimum angle of rotation (Figure 2). The Oracle model was the best fit to human RTs, with a correlation of $r = 0.56$, 95% CI $[0.51, 0.61]$ (Figure 3), although the distribution of response times did not match that of people (Figure 4). Moreover, the Oracle model was 100% accurate, and therefore could not explain the effect of rotation on people's accuracy.

Threshold model There was an overall monotonic relationship between the minimum angle of rotation and the number of actions taken by the Threshold model (Figure 2), though this relationship did not hold for *individual* shapes (e.g., Figure 5). The Threshold model was able to explain a moderate amount of the variance in human RTs, with a correlation of $r = 0.38$, 95% CI $[0.32, 0.45]$ (Figure 3). Like the Oracle model, the overall distribution of its RTs did not match that of people (Figure 4). The Threshold model had 100% accuracy, and thus did not exhibit a relationship between minimum angle and accuracy. As noted, we fit $\sigma_\varepsilon = 0.6$ for the Threshold model. This had the interesting effect of causing the Threshold model to *over*rotate, because the step size was large enough that it sometimes missed the global maximum, and had to do another full rotation to find it.

HC model The HC was the only model for which there was no monotonic relationship between rotation and RT (Figure 2). Moreover, the HC model was barely above chance ($M = 59.7\%$, 95% CI $[59.2\%, 60.2\%]$) and there were 312 stimuli for which it was not above chance. The HC model was not a good predictor of human RTs ($r = 0.09$, 95% CI $[0.01, 0.18]$), as shown in Figure 3. It was a moderate predictor of human accuracy ($r = 0.24$, 95% CI $[0.17, 0.31]$).

BQ model Like the Oracle and Threshold models, there was an overall monotonic relationship between rotation and the number of steps taken by the BQ model (Figure 2). Unlike the Threshold model, this relationship existed for individual shapes as well (e.g., Figure 5). The BQ model explained variance in human RTs about as well as the Threshold model (Figure 3), with a correlation of $r = 0.26$, 95% CI $[0.18, 0.33]$ for the equal prior and $r = 0.31$, 95% CI $[0.23, 0.39]$ for the unequal prior, and the RT distribution from the BQ model had the same overall shape as that of people (Figure 4).

The BQ model was quite accurate overall (equal prior: ($M = 95.3\%$, 95% CI $[95.1\%, 95.5\%]$; unequal prior: $M = 95.3\%$, 95% CI $[95.1\%, 95.5\%]$). With the equal prior, there were 12 stimuli for which it was not above chance; with the unequal prior, there were 14. The correlation with people's accuracy was $r = 0.23$, 95% CI $[0.16, 0.30]$ (equal prior) and $r = 0.15$, 95% CI $[0.08, 0.21]$ (unequal prior).

Because the BQ model relies on Equation 1 for its stopping criteria (as opposed to just finding a maximum), the prior $p(h)$ had an observable effect (Figure 2). As expected, with just a small bias of $p(h = 0) = 0.55$, there was a clear separation in RTs for "same" versus "flipped" stimuli: because of this bias, the model needed less evidence before accepting

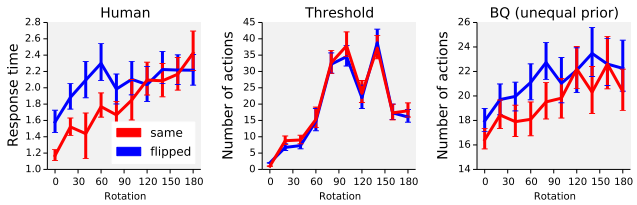

Figure 5: **Typical RT curves for a single object.** These plots correspond to the object shown in Figure 1. Left: human curves are either linear (as with the "same" pairs), or linear and then flat (as with the "flipped" pairs). Middle: the Threshold model does not have a monotonic relationship with rotation. Right: the BQ model is roughly linear.

$h = 0$ (thus taking less time). This separation is similar to the trend also observed in human RTs. The prior also had an effect on accuracy (though this did not reflect human behavior): the bias towards $h = 0$ meant that the model was more likely to judge a pair as "same", thus, accuracy increased for "same" pairs, but decreased for "flipped" pairs.

Discussion

We set out to answer the question of how people decide *what* to simulate when using mental imagery. Focusing on the specific case of determining the direction and extent of mental rotation, we formalized four models and compared their performance with the results of a behavioral experiment.

The Oracle and Threshold models were the best predictors of human RTs. However, both are somewhat unsatisfying explanations because they rely on *a priori* knowledge that people are unlikely to have. Moreover, they offer no explanation of several aspects of human behavior. First, their overall RT distributions look nothing like people's (Figure 4). Second, they both are 100% accurate, and so cannot explain the systematic relationship between rotation and human accuracy (Figure 2). Third, neither model can explain the difference in people's behavior on "same" and "flipped" stimuli.

In contrast, the BQ model was nearly as good as the Threshold model, yet it makes no assumptions about people's *a priori* knowledge. Furthermore, the BQ model matches people's behavior better than the Oracle or Threshold models in several ways. Its overall RT histogram has the same general shape as people's (Figure 4). Moreover, a closer look shows that the BQ model maintains the monotonic relationship between angle and RT even on individual stimuli, while the Threshold model does not (Figure 5). Finally, the BQ model's adaptive stopping rule is sensitive to the prior, and thus provides a possible explanation for why people are slower to respond on "flipped" stimulus pairs.

Thus, we suggest that the BQ model offers the most promising explanation of people's behavior on the mental rotation task to date. While it is not a perfect account, there are several ways in which it could be improved. For example, while we used holistic rotations in this paper, there is evidence that people compare individual features of shapes (Just & Carpenter, 1976; Yuille & Steiger, 1982). Additionally, a different active sampling approach could maintain a distribution over the location and value of the global maximum, rather than over the integral. We intend to explore these possibilities in future work, building upon the foundation established in this paper and working towards a better understanding of *what* people choose to simulate.

Acknowledgments This research was supported by ONR MURI grant number N00014-13-1-0341, and a Berkeley Fellowship awarded to JBH.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249–277.
Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis*, *24*(24), 509–522.
Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive Psychology*, *7*, 20–43.
Diaconis, P. (1988). Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, *1*, 163–175.
Duvenaud, D. (2013). *Log-gaussian processes for Bayesian quadrature.* Personal communication.
Funt, B. V. (1983). A parallel-process model of mental rotation. *Cognitive Science*, *7*(1), 67–93.
Gardony, A. L., Taylor, H. A., & Brunye, T. T. (2014). What does physical rotation reveal about mental rotation? *Psychological Science*, *25*(2), 605–612.
Glasgow, J. I., & Papadias, D. (1992). Computational imagery. *Cognitive Science*, *16*(3), 355–394.
Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.
Julstrom, B. A., & Baron, R. J. (1985). A model of mental imagery. *International Journal of Man-Machine Studies*, *23*, 313–334.
Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2011). Don't stop 'til you get enough: adaptive information sampling in a visuomotor estimation task. In *Proceedings of the 33rd Annual Meeting of the Cogntive Science Society* (pp. 2854–2859).
Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480.
Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, *92*, 137–172.
Kosslyn, S. M., & Shwartz, S. P. (1977). A simulation of visual imagery. *Cognitive Science*, *1*, 265–295.
Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2009). *The Case for Mental Imagery*. New York, NY: Oxford University Press.
Marr, D. (1983). *Vision*. New York, NY: Henry Holt and Company.
McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., & Gureckis, T. (2012). psiTurk (Version 1.02) [Computer software manual]. New York, NY. Retrieved from `https://github.com/NYUCCL/psiTurk`
Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S. J., & Ghahramani, Z. (2012). Active learning of model evidence using Bayesian quadrature. In *Advances in Neural Information Processing Systems 25* (pp. 46–54).
Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Science*, *25*, 157–238.
Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.
Sebastian, T. B., Klein, P. N., & Kimia, B. B. (2003). On aligning curves. *IEEE Transactions on Pattern Analysis*, *25*(1), 1–9.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.
Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.
Yuille, J. C., & Steiger, J. H. (1982). Nonholistic processing in mental rotation: some suggestive evidence. *Perception & Psychophysics*, *31*(3), 201–209.