Exploring the space of human exploration

Eric Schulz[1,*] **(ericschulz@fas.harvard.edu),**
Lara Bertram[2,*], **Matthias Hofer**[3], & **Jonathan D. Nelson**[2]

[1]Department of Psychology, Harvard University, Cambridge, Massachusetts, USA
[2]School of Psychology, University of Surrey, Guildford, UK
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[*] Contributed equally to this work.

Abstract

What drives people's exploration in complex scenarios where they have to actively acquire information by making queries? How do people adapt their selection of queries to their environment? We explore these questions using Entropy Mastermind, a novel variant of the Mastermind code-breaking game, in which participants have to guess a secret code by making useful queries. Participants solved games more efficiently and more quickly if the entropy of the game environment was low; moreover, people adapted their initial queries to the scenario they were in. We also investigated whether it would be possible to predict participants' queries within the generalized Sharma-Mittal information-theoretic framework. Although predicting individual queries is difficult, the modeling framework offered important insight on human behavior. Entropy Mastermind offers rich possibilities for modeling and behavioral research.

Keywords: Curiosity; Active Learning; Exploration; Entropy

Introduction

Humans are curious animals. From learning how to speak to launching rockets into space, exploration drives mankind's progress small and large. Human exploration and curiosity have been studied in reinforcement learning, self-directed sampling, and active learning paradigms. Recently there has been a great deal of conceptual work in this area (Coenen, Nelson, & Gureckis, 2018; Gottlieb & Oudeyer, 2018; Gureckis & Markant, 2012; Schulz & Gershman, 2019), which is underpinned by the common assumption that behavior is goal-directed and that people select observations based on a metric of usefulness (Settles, 2009).

Self-directed learning has been investigated empirically in adults and children, in domains including causal learning (Bramley, Dayan, Griffiths, & Lagnado, 2017), categorization (Meder & Nelson, 2012), control tasks (Osman & Speekenbrink, 2012), and 20-questions games (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014). Self-directed learning can lead to improved performance (Gureckis & Markant, 2012; Markant, Ruggeri, Gureckis, & Xu, 2016). For instance, participants actively intervening on a causal system made better inferences about the underlying causal structure than subjects who received identical information in a passive fashion (Lagnado & Sloman, 2004).

What metrics best predict how people evaluate the usefulness of possible queries? Past work has focused on the expected reduction of uncertainty, the extent of predictions' improvement, or the maximization of future rewards (Nelson, 2005). One study optimized experimental materials to maximally distinguish between different measures in an experience-based probabilistic classification task (Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Results showed that participants were better described by probability gain than by information gain or other measures. But different models may better account for human behavior on other tasks. For instance, probability gain does not capture human intuitions well on 20-questions tasks (Nelson et al., 2014).

Gureckis et al. (2012) tested whether participants maximize payoffs or information gain in a game of "battleships", where each query cost money and an attempt to maximize utility would lead to different queries than information-gain based strategies. Surprisingly, participants' sampling behavior was best matched by information gain. The authors argued that using information gain would lead to more knowledge about the underlying structure and therefore can be an effective strategy, no matter what the final task will be. Similar results have been obtained on an active causal learning task (Bramley, Lagnado, & Speekenbrink, 2015).

Studies on reinforcement learning frequently focus on how humans explore promising options, and have tested several exploration algorithms, such as random and directed exploration (Schulz & Gershman, 2019; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). Interestingly, if one defines directed exploration as an "uncertainty bonus" that inflates rewards by their predictive standard deviation, this formalism can again be seen as an approximate measure of information gain (Srinivas, Krause, Kakade, & Seeger, 2012).

A quintessential game of curiosity

In the Mastermind code-breaking game, both learning and exploitation are important. Thus, Mastermind offers a potential platform for bringing together pure information models (like expected information gain) and reinforcement learning models. It was invented in 1970 by Mordecai Meirowitz, and was one of the most successful games of the 20th century. Although Mastermind has been extensively studied in computer science (for references see Berghman, Goossens, & Leus, 2009; Cotta, Guervós, García, & Runarsson, 2010), comparatively little work has been done in cognitive science (but see, e.g., Laughlin, Lange, & Adamopoulos, 1982; Zhao, van de Pol, Raijmakers, & Szymanik, 2018).

We introduce a new game, "Entropy Mastermind", for studying exploration-driven problem solving (Fig. 1). A key difference between Entropy Mastermind and the classic game is that in Entropy Mastermind, hidden codes are drawn from known probability distributions. By varying the game environment (the probability distributions from which hidden codes are drawn, depicted visually as an icon array of fruits), Entropy Mastermind allows for research on how the level of entropy affects people's strategies and success in game play.

As a first step toward modeling behavior in a probabilistic framework, we use a model that values both maximizing the probability of a correct query and a curiosity bonus, similar to recent work on reinforcement learning in vast decision spaces (Wu et al., 2018). To model curiosity we use the Sharma-Mittal space of entropy measures, which subsumes several well-known entropy measures as special cases (Crupi, Nelson, Meder, Cevolani, & Tentori, 2018). According to the setting of two parameters, known as the *order* and *degree*, this entropy space can recover Shannon entropy, Bayes's error, theoretically important intermediate models (for instance from the Arimoto family of entropy measures), and the Rényi and Arimoto families of entropy measures. Whereas information gain has traditionally been thought of as reduction in Shannon entropy, any entropy metric could be used.

In what follows, we formally define the Sharma-Mittal space as a unifying framework for information gain measures. We then report an exploratory study assessing and modeling human behavior with Fruit Salad Mastermind, a version of Entropy Mastermind. Participants adapted their queries to the level of entropy in the environment, solving games in less-entropic environments faster and more efficiently. Both the exploration and exploitation parts of the model were important to account for human behavior.

Mapping the space of curiosity

In Mastermind both *learning* the true code and *guessing* the true code are important. (To make this intuitive, suppose that there are two possible codes, given everything that has been learned to date, and that one of these codes has 90% probability of being the correct code. Although the same information will be gleaned from testing either code, clearly it is sensible to test the code that has 90% probability of being correct, thus ending the game more efficiently on average.) We implement this idea via a softmax response rule on a value function which is based on the probability of each query being the correct code in the immediate time step, as well as a curiosity-driven exploration bonus:

$$P(\text{action} = a_i) \propto P(\text{success}|a_i) + \beta \times \text{curiosity bonus}(a_i) \tag{1}$$

How promising a code seems is determined by its current probability of being correct $P(success|a_i)$. This probability is always the same given a specific history of queries and feedback. The curiosity bonus(a_i) is weighted by a free parameter β and can be defined as how much an action promises to reduce uncertainty over the space of possible hypotheses (i.e., how much it reduces uncertainty about possible codes).

The uncertainty in a discrete random variable $K = k_1, k_2, \ldots k_n$ can be measured by its entropy. We use a generalized class of entropy measures that unifies multiple past proposals (Crupi et al., 2018). This class is called the Sharma-Mittal space, and can be defined as shown below:

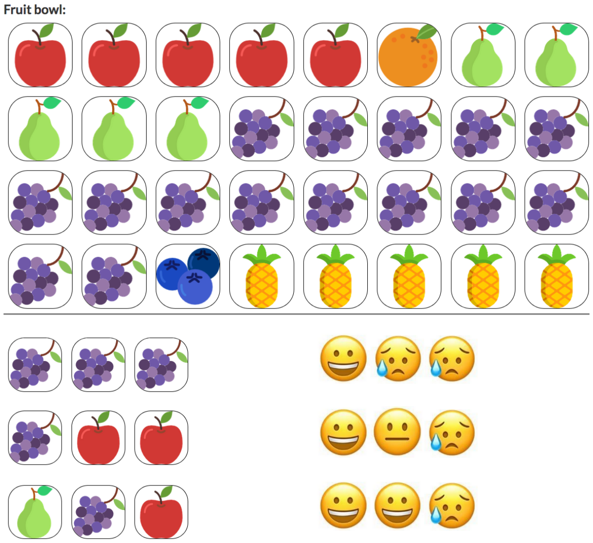

Figure 1: **Fruit Salad Mastermind. Top:** Icon array presenting the fruit bowl that generated the secret code. The fruits are always chosen from the same six fruit types (apples, oranges, blueberries, grapes, pears, and pineapples), by random sampling with replacement. Duplicates are allowed, so it is possible that the same fruit could appear in all positions of the hidden code. Players have to guess which fruit is in which position of three slots of the secret code, by clicking on the position they want to change. Each position is initially blank; clicking cycles through the possible fruits. Once participants are satisfied with the proposed code, they can click on a "Check" button (not shown), and then receive feedback. **Bottom:** History of game play illustrating feedback. In the first guess, the player guessed 3 grape items. The feedback (one smiling face followed by two frowning faces) conveys that exactly one of the items is exactly correct. However, the player does not know which of their guesses is correct: there is no correspondence between the position of the guess and the position of the feedback. In the second guess, the player tests grape in the first position, and apple in each of the other two positions. The feedback (smiling face, neutral face, frowning face) indicates that one of the items is the correct item in the correct location, another item is in the code but needs moved to a new location, and another item is not in the code at all. As before, the guesser has to figure out which feedback smiley face corresponds to which item in the code. The third guess of pear, grape, apple obtains two smiling faces and one frowning face. At this point the guesser can infer that the middle position is grape, and the final position is apple; the guesser must still figure out the first item.

$$\text{entropy}(K) = \frac{1}{t-1}\left[1 - \left(\sum_{i=1}^{n} P(k_i)^r\right)^{\frac{t-1}{r-1}}\right] \tag{2}$$

where r is the order and t the degree of the entropy measure. Note that limits, which exist, are used for points where the above equation is undefined. Although the above equation may not be immediately intuitive, there are a number of ways to build understanding about this space. All of the Sharma-Mittal entropy measures can be thought of as quantifying the average surprise that would be experienced if the value of the random variable K were learned. In the case of Mastermind, this would be the average surprise that would be experienced if one were to learn the true hidden code.

The degree parameter t governs which kind of surprise is averaged. Important values include that if $t = 1$, then $surprise(k_i) = ln(1/P(k_i))$; if $t = 2$, then $surprise(k_i) =$

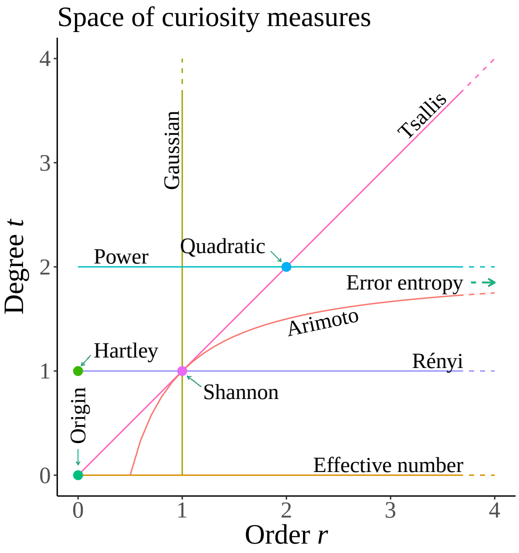

Figure 2: **Sharma-Mittal space.** The Sharma-Mittal family of entropy measures is represented in a Cartesian quadrant with values of the order parameter r and of the degree parameter t. Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function. Several special cases are highlighted.

$1 - P(k_i)$. If $t > 1$, a test is more useful if it is conclusive than if it is not. If $t < 1$, a test is always less useful if it is conclusive than if it is not. The order parameter r determines what kind of averaging function is used. It can be thought of as an index of the imbalance of the entropy function, which indicates how much the entropy measure discounts minor (low probability) hypotheses. For example, when $r = 0$, entropy becomes a (increasing) function of the mere number of the available options. When r goes to infinity, on the other hand, entropy becomes a (decreasing) function of the probability of a single most likely hypothesis.

Several special cases exist within the Sharma-Mittal space, as Figure 2 illustrates. For example, Shannon entropy is the result of setting $r = t = 1$, and probability gain (also called error entropy) is the result of setting $t = 2$ and letting $r \to \infty$. One of the goals of the present research is to investigate whether people's striving for information (the curiosity goal) can be represented well as a generalized information gain metric, where information is defined as the expected reduction in one of the Sharma-Mittal entropy functions over the probability distribution of the possible codes.

Methods

Participants and Design Forty-seven first-year undergraduate students (38 female, M_{age}=19.04; SD=1.04; range: 18 to 23) participated in our study as part of a cognitive psychology class. Participants gave informed consent in accordance with the university's procedures and the Helsinki Declaration. They spent 10.5 minutes on average on the task.

We explored how the entropy of the distribution generating the secret code affected participants' behavior. Participants played as many rounds as they wanted within the assigned time. In each game, one of the entropy conditions was chosen at random and the six fruits were randomly assigned to one of the six proportions of that condition. The resulting generating "fruit bowl" was presented to participants as an icon array above the current game. A "hidden fruit code" was generated from that distribution.

Materials and Procedure Before starting the experiment, participants were introduced to the rules and interface of the Fruit Salad Mastermind game were and were required to correctly answer four comprehension questions. Participants were told to figure out the secret code using as few guesses as possible. Participants played as many rounds as they wanted within the available time in the lab session.

Entropy conditions The four different entropy conditions specified the distribution from which the underlying secret code was sampled with replacement. In the *very high entropy* (Shannon entropy of code jar 2.58 bits), the secret code was sampled based on the proportions $(5,5,5,5,6,6)$. This means, for example, that there could be 5 pineapples, 5 apples, 5 pears, 5 blueberries, 6 grapes, and 6 oranges, out of a total of 32 items, from which three fruits were sampled with replacement to generate the secret code. In the *high entropy* condition (Shannon entropy 2.08 bits), the secret code was sampled based on the proportions $(1,1,5,5,5,15)$. In the *low entropy* condition (Shannon entropy 1.62 bits), the secret code was sampled based on the proportions $(1,1,1,4,4,21)$. Finally, in the *very low entropy* condition (Shannon entropy 0.99 bits), the secret code was sampled based on the proportions $(1,1,1,1,1,27)$.

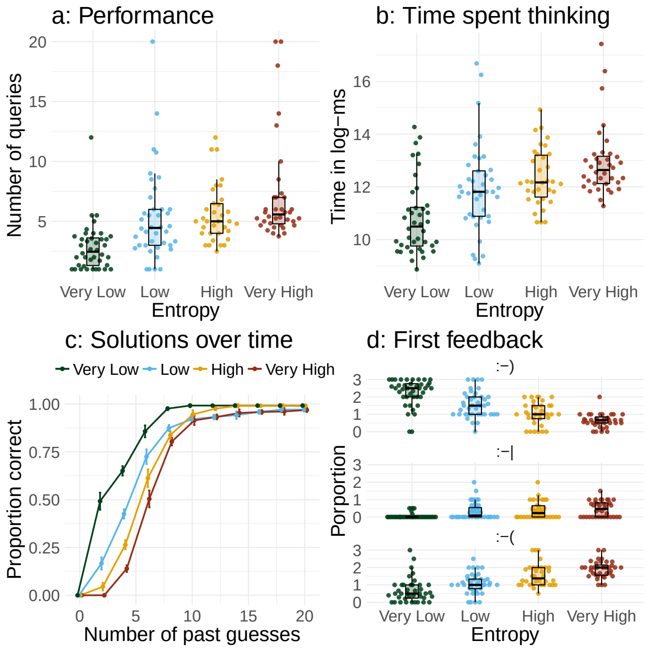

Figure 3: **Behavioral results. a:** Number of queries required to solve a game by entropy condition. **b:** Time spent thinking (measured in log-ms per guess) by entropy condition. **c:** Proportion of correct guesses in dependency of number of past guesses by entropy condition. **d:** Mean proportional feedback after first guess by entropy condition. Points represent mean per participant. Error bars indicate the standard error of the mean.

Behavioral results

We first analyzed the number of required guesses to solve a game as a function of the entropy condition (Fig. 3a). This revealed a positive average rank correlation between how much entropy a condition contained and the number of queries participants required to solve a game (Kendall's $\tau = 0.48$, $t(46) = 12.44$, $d = 1.81$, $BF > 100$). More specifically, participants required fewer queries on average for the very low entropy games as compared to low entropy games ($t(46) = -5.69$, $p < .001$, $d = 0.83$, $BF > 100$). They also required fewer queries for the low entropy games than for the high entropy games ($t(46) = -3.16$, $p = .002$, $d = 0.46$, $BF = 11.8$). Finally, participants needed fewer queries for the high entropy games than for the very high entropy games ($t(46) = -3.96$, $p < .001$, $d = 0.58$, $BF = 97.2$).

Next, we analyzed how much time participants spent thinking to enter a guess by entropy condition (Fig. 3b). Thus, we assessed their mean time to submit a query measured in log-milliseconds. There was a positive average rank-correlation between a game's entropy and participants' average time spent thinking, Kendall's $\tau = 0.39$, $t(46) = 11.50$, $d = 1.68$, $BF > 100$. More specifically, participants spent less time thinking during the very low entropy games than during the low entropy games ($t(46) = -5.05$, $p < .001$, $d = 0.74$, $BF = 97.2$). They also spent less time thinking in the low entropy than in the high entropy games ($t(46) = -3.96$, $p < .001$, $d = 0.58$, $BF = 97.9$). Finally, they spent less time in the low entropy than in the very low entropy games ($t(46) = -3.99$, $p < .001$, $d = 0.58$, $B > 100$).

We also analyzed the proportion of solved games as a function of the number of past guesses, again comparing the different entropy conditions (Fig. 3c). We thus estimated a Bayesian logistic regression of number of past guesses onto the proportion of correct guesses for each of the entropy conditions, using Metropolis-Hastings Markov chain Monte Carlo sampling. The resulting posterior estimate for the effect of number of past guesses onto the probability of guessing correctly was smallest for the very high entropy condition ($\hat{\beta} = 0.15$, 95%HDI=[0.14, 0.16]). The same estimate was higher for the high entropy condition ($\hat{\beta} = 0.19$, 95%HDI=[0.18, 0.20]), which did not differ meaningfully from the low entropy condition ($\hat{\beta} = 0.18$, 95%HDI=[0.17, 0.20]). The very low entropy condition showed the highest estimated effect ($\hat{\beta} = 0.30$, 95%HDI=[0.28, 0.33]). Thus, participants' solution rates differed meaningfully between entropy conditions, with lower entropy leading to faster rates.

In our last behavioral analysis, we looked at the very first query participants submitted as well as the feedback they received for that query (Fig. 3d). The number of smiling faces received on the very first guess was negatively rank-correlated with entropy condition, $\tau = -0.51$, $t(41) = -9.80$, $p < .001$, $d = 1.51$, $BF > 100$, whereas the number of frowning faces showed a positive rank-correlation, $\tau = 0.30$, $t(30) = 6.00$, $p < .001$, $d = 1.06$, $BF > 100$. Interestingly, participants adapted their first queries to the entropy condition, leading to a positive rank correlation between the set size of their first query (the number of unique kinds of fruit contained in the query) and the entropy of the generating distribution, $\tau = 0.40$, $t(46) = 9.00$, $p < .001$, $d = 1.31$, $BF > 100$. Put differently, if the generating distribution were higher entropy, then participants tested a larger number of different fruits as part of their first query.

Computational modeling

We now turn to a model-based analysis of participants' exploration strategies. For this, we first need a formal account of intelligent Mastermind play. Logically, all combinations that are still consistent in round i based on the feedback received so far are part of a feasible set $\mathcal{F}_i$. Note that in Entropy Mastermind, not only the feasible codes but also their probabilities (which are not in general equal) are relevant. Code combinations ruled out by prior feedback have zero probability, while the remaining items' probability mass is proportional to the probability of obtaining the item via sampling from the code jar. The effective size of the feasible set is the total number of all non-zero probability codes left in the set. Let the probability that c_i is the hidden code given the current feasible set be denoted $P(c_i)$. The feasible set is guaranteed to shrink after each round unless a guess c_i is repeated. A general playing strategy consists of (i) identifying the set of feasible combinations $\mathcal{F}_i$ (with $\mathcal{F}_0 = \mathcal{E}$), where prior feedback is used to determine which combinations are still viable; and (ii) picking a combination c_i for the next guess. Let us denote the informational usefulness of playing combination c in the current round with $u(c)$. The formula for computing $u(c)$ is

$$u(c) = \text{entropy}(\mathcal{F}_i) - \sum_{r}^{\mathcal{R}} P(f) \cdot \text{entropy}(\hat{\mathcal{F}}_{c,f}), \qquad (3)$$

the difference in entropy (under a particular Sharma Mittal entropy measure with specified order and degree) between the current feasible set and the expected entropy when playing code c. To compute expected entropy, for each possible feedback $f \in \mathcal{R}$, we compute the product of the probability of receiving that feedback $P(f)$ times the entropy of the updated feasible set $\hat{\mathcal{F}}_{c,r}$ when playing combination c and receiving feedback r. To compute $P(f)$ for a given c, we look at all the combinations $c_j \in \mathcal{F}_i$, that lead to feedback f. To this end, we define a feedback function $h(c, c_j) = f$ that returns the feedback f obtained from checking code c against code c_j. The probability of feedback for code c can then be calculated as follows:

$$P(f) = \frac{\sum_{c_j}^{\mathcal{F}} P(c_j) \cdot \mathbb{1}_{h(c,c_j)=f}}{\sum_{c_j}^{\mathcal{F}} \sum_{c_k}^{\mathcal{F}} P(c_k) \cdot \mathbb{1}_{h(c_j,c_k)=f}}.$$

The indicator function $\mathbb{1}_{h(c,c_j)=f}$ ensures that we only sum over codes c_j that generate the required feedback f. The probability of any combination of fruits $c = m_1 m_2 ... m_n$ can be computed as

$$P(c = m_1 m_2 ... m_n) = P(m_1) \cdot P(m_2) \cdot ... \cdot P(m_n) \qquad (4)$$

where each $P(m)$ represents the probability of sampling the corresponding fruit item from the fruit jar. The other term of Equation 3, entropy$(\hat{\mathcal{F}}_{c,f})$, requires us to compute hypothetical feasible sets $\hat{\mathcal{F}}_{c,r}$. Given the current feasible set $\mathcal{F}_i$, a combination c we want to evaluate, and hypothetical feedback f, we need to exclude all combinations $c_j \in \mathcal{F}_i$ for which $h(c, c_j) \neq f$; that is, all combinations c_j that are not consistent with obtaining feedback f.

Lastly, in order to measure how much one would like one such hypothetical set, one has to assign a utility to a feasible set $\mathcal{F}$. For this, we use the Sharma-Mittal entropy framework to compute the entropy of a probability distribution defined over set $\mathcal{F}$, $P_{\mathcal{F}}(c)$. For each combination $c \in \mathcal{F}$

$$P_{\mathcal{F}}(c) = \frac{P(c)}{\sum_{c_j}^{\mathcal{F}} P(c_j)},$$

where the nominator $P(c)$ is computed according to Equation 4 and the denominator is a normalization term.

We assess how well the combination of an entropy-based exploration bonus and the probability of making a correct guess describes players' guesses over time. For this, we analyzed the last five games of the 34 participants who played at least five games in total. Next, we calculated the expected information gain for all of the $6 \times 6 \times 6$ possible fruit combinations that a participant could enter on every trial for every participant, given the participant-specific history of queries in a game. We calculated this information gain for every combination of order $r = [1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, 16, 32, 64]$ and degree $t = [1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, 16, 32, 64]$, i.e. 121 models per participant in total. We then combined the probability of a guess being correct with the information gain assessed by the specific entropy measure following Equation 2 to arrive at a value of an action's usefulness $V(a_t)$, which we put in a softmax function to calculate choice probabilities:

$$P(x) = \frac{\exp(V(a_t(\mathbf{x}))/\tau)}{\sum_{j=1}^{N} \exp(V(a_t(\mathbf{x}))/\tau)} \qquad (5)$$

where τ is a free temperature parameter. For each participant, we calculated a model's AIC$(\mathcal{M}) = -2\log(L(\mathcal{M})) + 2k$ and standardized it using a pseudo-R^2 measure as an indicator for goodness of fit, comparing each model $\mathcal{M}_k$ to a random model: $\mathcal{M}_{\text{rand}}$, $R^2 = 1 - \text{AIC}(\mathcal{M}_k)/\text{AIC}(\mathcal{M}_{\text{rand}})$.

The results of this analysis revealed a mean pseudo-R^2 of 0.041 over all orders and degrees, which was low but significantly better than chance ($t(33) = 20.52$, $p < 0.001$ $d = 1.86$, $BF > 100$). Moreover, the estimated median temperature parameter was $\tau = 1.02$, indicating a relatively wide spread of predictions. There was a significant negative rank-correlation between the degree parameter and model fit, $\tau = -0.37$, $z = -5.84$, $p < .001$, $BF > 100$, whereas this correlation was not significant for the order parameter, $\tau = 0.04$, $z = 0.60$, $p = .54$, $BF = 0.3$. Thus, even though entropies with smaller degree parameters seemed to generally work better at modeling participants' queries, there was no meaningful effect of the different order parameters.

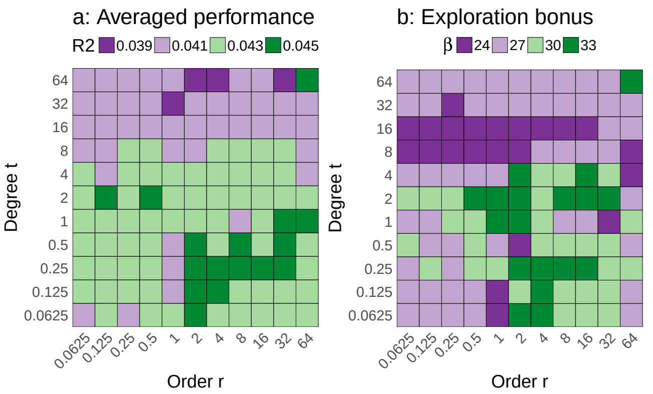

Figure 4: **Modeling results. a:** Averaged r^2 for different Sharma-Mittal parameters. **b:** Clustered model quality for different Sharma-Mittal parameters.

The range of pseudo-R^2 values, $0.038 - 0.045$, also shows that most of the entropy measures led to similar performance. We also assessed the magnitude of the estimated exploration bonus β (Fig. 4b), which had a mean of $\hat{\beta} = 27.81$, and therefore differed significantly from 0, $t(33) = 115.47$, $p < .001$, $d = 10.9$, $BF > 100$. Interestingly, areas of the Sharma-Mittal space with higher r^2 also tended to have higher β estimates.

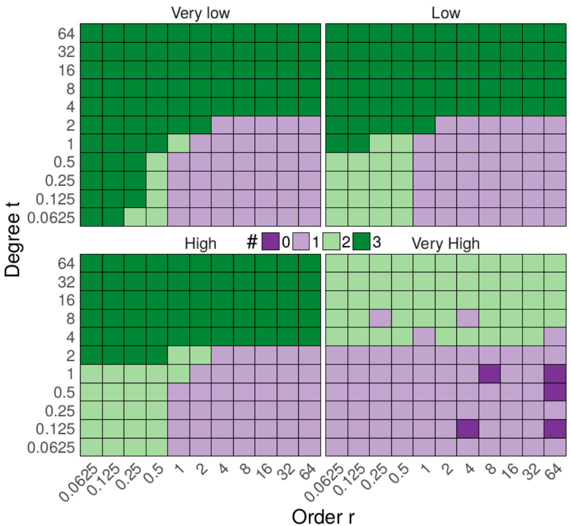

Figure 5: **Number of times most likely fruit was chosen in first query by simulated entropy models across entropy conditions.**

Finally, we compared how often participants put the most likely fruit into their first query with how often simulated models of different order and degree parameters chose the same fruit in their first query, for each entropy condition (see Fig. 5). The higher degree models chose the most likely fruit more often than people did. Specifically, participants put on average 2.14 of the most likely fruit in their first query in the very low entropy condition, 1.60 in the low entropy condition, 1.26 in the high entropy condition and 0.48 in the very high entropy condition. This analysis therefore corroborated our previous finding that the lower degree entropies better matched participants' queries. In relation to previous work modeling behavior with the Sharma-Mittal framework, Entropy Mastermind appears to be more similar to experience-based than to description-based probabilistic classification tasks (see Crupi et al., 2018, Fig. 7).

Discussion and conclusion

We introduced Entropy Mastermind as a game for researching human curiosity and exploration. Participants reported Fruit Salad Mastermind to be engaging and fun. They required fewer queries, spent less time thinking about queries and showed faster learning rates if the distribution generating the secret code had lower entropy. Participants also adapted their queries to the code-generating distribution, and did so in sensible ways. In particular, many of the informational models (Figure 5) use greater proportions of the most-probable fruit in the first guess in lower-entropy conditions; participants also followed this pattern.

Modeling results paralleled earlier findings from other tasks (Crupi et al., 2018) suggesting that it is easier to identify the value of the degree parameter than of the order parameter in the Sharma-Mittal space. Moreover, the general predictive performance of many models was rather similar and relatively low. This might be due to the overall complexity of choices, since there were 216 possible options on every trial, making it difficult to compare among candidate models (also see Parpart, Schulz, Speekenbrink, & Love, 2017).

The difficulty of modelling could also be due to participants using cognitive shortcuts instead of fully entropy-reducing strategies, as has been observed in other domains of active learning (Bramley et al., 2015, 2017). Future studies should therefore investigate both heuristic strategies (Gigerenzer & Gaissmaier, 2011) and boundedly rational approaches (Griffiths, Lieder, & Goodman, 2015). Adaptive experimental designs (Cavagnaro, Myung, Pitt, & Kujala, 2010) could be employed to maximally discriminate among the Sharma-Mittal parameters.

Entropy Mastermind is a promising paradigm to investigate human exploration behavior in complex hypothesis testing scenarios. Although our current modeling framework did not fully map out the space of exploration behavior, we believe that combining the Sharma-Mittal space of entropy measures with an enjoyable game rich in scientific history can further inform our theories of self-directed learning. To fully map out the space of human exploration, we have to keep exploring.

References

Berghman, L., Goossens, D., & Leus, R. (2009). Efficient solutions for mastermind using genetic algorithms. *Computers & Operations Research*, *36*, 1880–1885.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neuraths ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*, 301–338.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 708–731.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*, 887–905.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 1–41.

Cotta, C., Guervós, J. J. M., García, A. M. M., & Runarsson, T. P. (2010). Entropy-driven evolutionary approaches to the mastermind problem. In *International conference on parallel problem solving from nature* (pp. 421–431).

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive Science*, *42*, 1410–1456.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*, 758–770.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*, 464–481.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856–876.

Laughlin, P. R., Lange, R., & Adamopoulos, J. (1982). Selection strategies for "mastermind" problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(5), 475.

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016, sep). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, *10*, 142–152.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*, 119–148.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*, 74–80.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969.

Osman, M., & Speekenbrink, M. (2012). Prediction and control in a dynamic environment. *Frontiers in Psychology*, *3*, 68.

Parpart, P., Schulz, E., Speekenbrink, M., & Love, B. C. (2017). Active learning reveals underlying decision strategies. *bioRxiv*, 239558.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. W. (2012). Information-theoretic regret bounds for Gaussian Process optimization in the bandit setting. *IEEE Transactions on Information Theory*, *58*, 3250–3265.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915–924.

Zhao, B., van de Pol, I., Raijmakers, M., & Szymanik, J. (2018). Predicting cognitive difficulty of the deductive mastermind game with dynamic epistemic logic models. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Cogsci 2018* (pp. 2789–2794).