PROCEEDINGS OF THE ROYAL SOCIETY A

rspa.royalsocietypublishing.org

Optimally designing games for behavioural research

Anna N. Rafferty[1], Matei Zaharia[1] and Thomas L. Griffiths[2]

[1]Computer Science Division, and [2]Department of Psychology, University of California, Berkeley, CA 94720 USA

Research

Cite this article: Rafferty AN, Zaharia M, Griffiths TL. 2014 Optimally designing games for behavioural research. *Proc. R. Soc. A* **470**: 20130828.
http://dx.doi.org/10.1098/rspa.2013.0828

Received: 12 December 2013
Accepted: 7 April 2014

Subject Areas:
statistics, Computer modelling and simulation

Keywords:
optimal experiment design, Markov decision process, cognitive science, computer games

Author for correspondence:
Anna N. Rafferty
e-mail: rafferty@cs.berkeley.edu

Computer games can be motivating and engaging experiences that facilitate learning, leading to their increasing use in education and behavioural experiments. For these applications, it is often important to make inferences about the knowledge and cognitive processes of players based on their behaviour. However, designing games that provide useful behavioural data are a difficult task that typically requires significant trial and error. We address this issue by creating a new formal framework that extends optimal experiment design, used in statistics, to apply to game design. In this framework, we use Markov decision processes to model players' actions within a game, and then make inferences about the parameters of a cognitive model from these actions. Using a variety of concept learning games, we show that in practice, this method can predict which games will result in better estimates of the parameters of interest. The best games require only half as many players to attain the same level of precision.

1. Introduction

Computer games have become increasingly popular tools in education and the social sciences (e.g. [1–5]). Within education, games can provide authentic contexts for exploring scientific phenomena [6–8], and engage students through storytelling and immersive, dynamic environments [9]. Games can improve student learning by adaptively providing tasks that are of appropriate difficulty given the students' knowledge and through encouraging motivation and persistence [10–12]. Benefits also hold for behavioural research. For psychology, games provide a way of recruiting large numbers of engaged participants, and offer a powerful method for increasing participant satisfaction and diminishing

Royal Society Publishing

participant disinterest. They may also facilitate longer, more involved behavioural experiments. However, designing games for education and psychological research can be difficult. These games must provide insight into a student's knowledge or cognitive processes, which requires interpreting behaviour in the game. In order to gain as much information about these quantities as possible, a game designer must adjust the settings of many parameters in a game, such as the level design, the incentive structure and the placement of items within the game. This process usually requires significant time and effort, normally based on trial and error.

Designing traditional experiments can also require significant time and effort. While the number of parameters to adjust may be smaller, the experimenter must still set quantities such as what time delays to include in a memory task or what treatment dosages to compare in a medical experiment. The statistical theory of optimal experiment design aims to ease this problem by identifying the design that will give the most information about the dependent variable [13,14]. In chemistry, this technique has been used to discover the value of various parameters relevant to a reaction, making laboratory syntheses more successful (e.g. [15–17]), and the approach was used to develop and validate a new method for synthesizing a compound that has now been used in industry [18]. Optimal experiment design has also been used in pharmacology and clinical applications (e.g. [19–23]), resulting in greater certainty about the effectiveness of new drug therapies while reducing trial costs. Across fields, the idea of setting experiment parameters to optimize the information gained about the phenomena under investigation has made it easier to obtain precise answers while minimizing resource use (e.g. [24,25]).

In this paper, we introduce optimal game design, a new formal framework that extends optimal experiment design to identify game designs that will diagnose people's knowledge more efficiently. We investigate how to identify the game design with maximal utility, where utility is defined as the expected information gain about the parameters of a cognitive model. Like optimal experiment design, our procedure takes an existing design and considers how to modify it to be most informative. For traditional experiments, these modifications might include parameters such as at what time intervals to test recall; for games, these modifications include parameters like the amount of points for different types of accomplishments or the location and frequency of particular objects in the game. The framework leverages the skills of human designers for creating the initial game, and by automating the process of refining that game design, the framework limits the trial and error necessary to find a game that will provide useful data.

Adapting optimal experiment design methods to game design requires predicting people's behaviour within games, which may differ from behaviour in more traditional behavioural experiments. Typically, experiments have relatively simple incentive structures and individual actions in an experiment are not dependent on one another. Games often include a variety of competing incentives and actions in the game are likely to naturally build on one another. To model people's behaviour in games, we use Markov decision processes (MDPs), which are a decision theoretic model for reasoning about sequential actions. This model incorporates the added complexity of games by calculating both the current and future benefit of an action. By combining MDPs with ideas from optimal experiment design, we create a framework for finding the game that will provide the highest expected information gain about a question of interest. This framework provides the potential to investigate psychological questions and estimate student knowledge based on games, without needing to modify our questions to specifically account for the game environment.

We first provide background on optimal experiment design and MDPs. We then combine these ideas to create a framework for optimal game design. The remainder of the paper applies this general framework to the specific case of learning Boolean concepts. We introduce a novel concept learning game and use our approach to optimize the game parameters. Through behavioural experiments, we show that optimized game designs can result in more efficient estimation of the difficulty of learning different kinds of Boolean concepts. Our results demonstrate that this estimation can be complicated by people's own goals, which may not match incentives within the game, but can be accommodated within our framework. We end by summarizing the benefits of optimal game design as well as the limitations of this framework.

2. Background

Our framework relies on ideas from Bayesian experiment design and MDPs, which we will introduce in turn.

(a) Bayesian experiment design

Bayesian experiment design, a subfield of optimal experiment design, seeks to choose the experiment that will maximize the expected information gain about a parameter θ [13,26]. In psychology, this procedure and its variations have been used to design experiments that allow for clearer discrimination between alternative models, where θ corresponds to an indicator function about which of the models under consideration is correct [27,28]. Throughout this paper, let ξ be an experiment (or game) design and y be the data collected in the experiment. The *expected utility* (EU) of a game is defined as the expected information gain about the parameter θ

$$\left.\begin{aligned} \mathrm{EU}(\xi) &= \int p(y|\xi)U(y,\xi)\,\mathrm{d}y,\\ \text{where}\qquad p(y|\xi) &= \int p(y|\xi,\theta)p(\theta)\,\mathrm{d}\theta\\ \text{and}\qquad U(y,\xi) &= \int (H(p(\theta|y,\xi)) - H(p(\theta)))\,\mathrm{d}\theta, \end{aligned}\right\} \tag{2.1}$$

where $H(p)$ is the Shannon entropy of a probability distribution p, defined as $H(p) = \int p(x)\log(p(x))\,\mathrm{d}x$. The Bayesian experimental design procedure seeks to find the experiment ξ that has maximal EU. Intuitively, designs that are likely to result in more certainty about θ will have higher utility.

(b) Markov decision processes

The Bayesian experiment design procedure uses $p(\theta|y,\xi)$ to calculate the information gain from an experiment. This quantity represents the impact that the data y collected from experiment ξ have on the parameter θ. In a game, the data y are a series of actions, and to calculate $p(\theta|y,\xi)$, we must interpret how θ affects those actions. Via Bayes' rule, we know $p(\theta|y,\xi) \propto p(y|\theta,\xi)p(\theta)$. We thus want to calculate $p(y|\theta,\xi)$, the probability of taking actions y given a particular value for θ and a game ξ. To do so, we turn to MDPs, which provide a natural way to model sequential actions. MDPs and reinforcement learning have been used previously in game design for predicting player actions and adapting game difficulty [29–31].

MDPs describe the relationship between an agent's actions and the state of the world and provide a framework for defining the value of taking one action versus another (for an overview, see [32]). Formally, an MDP is a discrete time series model defined by a tuple $\langle S, A, T, R, \gamma\rangle$, where S is the set of possible states and A is the set of actions that the agent may take. At each time t, the model is in a particular state $s \in S$. The transition model T gives the probability $p(s'|s,a)$ that the state will change to s' given that the agent takes action a in state s. The reward model $R(s,a,s')$ describes the probability of receiving a reward $r \in \mathbb{R}$ given that action a is taken in state s and the resulting state is s'. For example, the reward model in a game might correspond to points. Finally, the discount factor γ represents the relative value of immediate versus future rewards. The value of taking action a in state s is defined as the expected sum of discounted rewards and is known as the Q-value

$$Q(s,a) = \sum_{s'} p(s'|s,a)\left(R(s,a,s') + \gamma \sum_{a'\in A} p(a'|s')Q(s',a')\right), \tag{2.2}$$

where $p(a'|s')$ is the probability that an agent will take action a' in state s' and is defined by the agent's policy π. We assume that people are noisily rational actors: they are more likely to take actions that they think have higher value. As in Baker *et al.* [33], this can be formally modelled as a Boltzmann policy $p(a|s) \propto \exp(\beta Q(s,a))$, where β is a parameter determining how close the policy is to optimal. Higher values of β mean the agent is more likely to choose the action with highest Q-value, while $\beta = 0$ results in random actions.

3. Optimal game design

We can now define a procedure for optimal game design, identifying the game with maximum expected information gain about some theoretical quantity of interest θ. The optimal game design framework improves an existing game by adjusting its parameters to be more diagnostic; these parameters may correspond to point values, locations of items, or any other factor that can be varied. To apply Bayesian experiment design to the problem of choosing a game design, we define the expected utility of a game ξ as the expectation of information gain over the true value of θ and the actions chosen by the players

$$\mathrm{EU}(\xi) = E_{p(\theta,\mathbf{a})}[H(p(\theta)) - H(p(\theta|\mathbf{a},\xi))], \tag{3.1}$$

where $\mathbf{a}$ is the set of action vectors and associated state vectors for all players. The expectation is approximated by sampling θ from the prior $p(\theta)$, and then simulating players' actions given θ by calculating the Q-values and sampling from the Boltzmann policy.

To compute $\mathrm{EU}(\xi)$, the distribution $p(\theta|\mathbf{a},\xi)$ must be calculated. Intuitively, this quantity connects actions taken in the game with the parameter of the cognitive model that we seek to infer, θ. For a game to yield useful information, it must be the case that people will take different actions for different values of θ. Concretely, we expect that players' beliefs about the reward model and the transition model may differ based on θ. Then, the process of inferring θ from actions assumes that each θ corresponds to a particular MDP. If this is the case, we can calculate a distribution over values of θ based on the observed sequences of actions $\mathbf{a}$ of all players in the game ξ

$$\begin{aligned} p(\theta|\mathbf{a},\xi) &\propto p(\theta)p(\mathbf{a}|\theta,\xi) \\ &= p(\theta)p(\mathbf{a}|\mathrm{MDP}_\theta,\xi) \\ &= p(\theta)\prod_i p(\mathbf{a}_i|\mathrm{MDP}_\theta,\xi), \end{aligned} \tag{3.2}$$

where $\mathbf{a}_i$ is the vector of actions taken by player i and MDP_θ is the MDP derived for the game based on the parameter θ. Calculating this distribution can be done exactly if there is a fixed set of possible θ or by using Markov chain Monte Carlo (MCMC) methods if the set of θ is large or infinite [34].

Now that we have defined $p(\theta|\mathbf{a},\xi)$, we can use this to find the expected utility of a game. Equation (3.1) shows that this calculation follows simply if we can calculate the entropy of the inferred distribution. In the case of a fixed set of possible θ, $H(p(\theta|\mathbf{a},\xi))$ can be calculated directly. If MCMC is used, one must first infer a known distribution from the samples and then take the entropy of that distribution. For example, if θ is a multinomial and $p(\theta)$ is a Dirichlet distribution, one might infer the most likely Dirichlet distribution from the samples and find the entropy of that distribution.

We have now shown how to (approximately) calculate $\mathrm{EU}(\xi)$. To complete the procedure for optimal game design, any optimization algorithm that can search through the space of games is sufficient. Maximizing over possible games is unlikely to have a closed form solution, but stochastic search methods can be used to find an approximate solution to the game with maximum expected utility. For example, one might use simulated annealing [35]. This method allows optimization of discrete and continuous parameters, where neighbouring states of the current game are formed by perturbations of the parameters to be optimized.

4. Optimal games for Boolean concept learning

We have described a general framework for automatically finding game designs that are likely to be highly informative about model parameters. To test how well this framework identifies

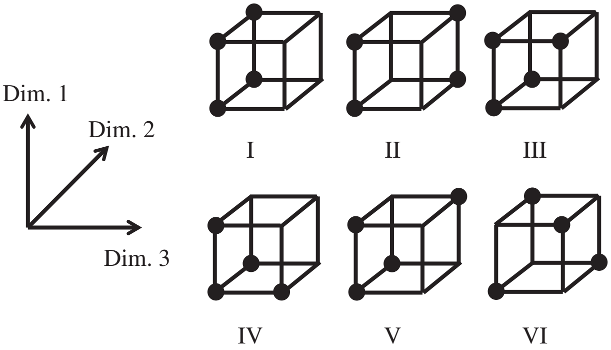

Figure 1. Boolean concept structures. In each structure, eight objects differing in three binary dimensions are grouped into two categories of four elements. Each object is represented as a corner of a cube based on its combination of features, and the objects chosen for one category in each problem type are represented by dots.

informative designs, we applied it to a particular question: what is the relative difficulty of learning various Boolean concept structures? This question has been studied extensively in past work (e.g. [36–39]), so we can compare our results to those produced using more traditional methods. We first describe Boolean concept learning, and then turn to the initial game we created and the application of optimal game design.

(a) Boolean concepts

In Boolean concept learning, one must learn how to categorize objects that differ along several binary dimensions. We focus on the Boolean concepts explored in [36]. In these concepts, there are three feature dimensions, resulting in 2^3 possible objects, and each concept contains four objects. This results in a total of 70 concepts with six distinct structures, as shown in figure 1. Shepard *et al.* found that the six concept structures differed in learning difficulty, with a partial ordering from easiest to most difficult of $\mathrm{I} > \mathrm{II} > \{\mathrm{III}, \mathrm{IV}, \mathrm{V}\} > \mathrm{VI}$ [36]. Similar results were observed in later work [37,38] although the position of type VI in the ordering can vary [39].

Using optimal game design to infer the difficulty of learning Boolean concepts requires a computational cognitive model. We follow previous modelling work that assumes learners' beliefs about the correct concept h can be captured by Bayes' rule [39]

$$\begin{aligned} p(h|\mathbf{d}) &\propto p(h)p(\mathbf{d}|h) \\ &= p(h)\prod_{d\in\mathbf{d}} p(d|h), \end{aligned} \tag{4.1}$$

where each $d \in \mathbf{d}$ is an observed stimulus and its classification, and observations are independent given the category. The likelihood $p(d|h)$ is then a simple indicator function. If the stimulus classification represented by d matches the classification of that stimulus in hypothesis h, denoted $h \vdash d$, then $p(d|h) \propto 1$; otherwise, $p(d|h) = 0$. We seek to infer the prior $p(h)$, which represents the difficulty of learning different concepts and thus gives an implicit ordering on structure difficulty. In our earlier terminology, θ is a prior distribution on concepts $p(h)$. For simplicity, we assume all concepts with the same structure have the same prior probability, so θ is a six-dimensional multinomial. Each θ_i represents the prior probability of a single concept of type i.

(b) Corridor Challenge

To teach people Boolean concepts, we created the game Corridor Challenge, which requires learning a Boolean concept to achieve a high score. Corridor Challenge places the player in a

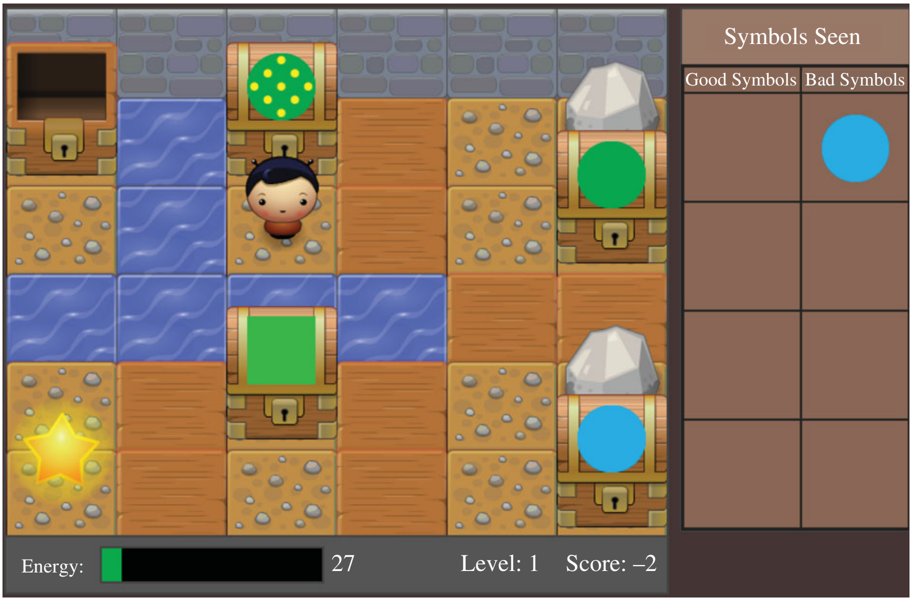

Figure 2. User interface for the Corridor Challenge game (Level 1 of the random game in experiment 1). In this screenshot, the player has opened the first chest and moved to the second island.

corridor of islands, some of which contain a treasure chest, joined by bridges (figure 2).[1] The islands form a linear chain and the bridges can be crossed only once, so players cannot return to previous chests. Some chests contain treasure, while others contain traps; opening a chest with treasure increases the player's score and energy, while opening a chest with a trap decreases these values. Each chest has a symbol indicating whether it is a trap; symbols differ along three binary dimensions and are categorized as a trap based on one of the Boolean concepts. Players are shown a record of the symbols from opened chests and their meanings (see the right-hand side of figure 2). Players are told to earn the highest score possible without running out of energy, which is depleted by moving to a new island or opening a trapped chest. When a player runs out of energy, the level is lost and she cannot explore the rest of the level; surviving a level earns the player 250 points. Corridor Challenge games may consist of several levels. Each level is a new corridor with different chests, but the same symbols are used and they retain the same meaning as on the previous level. At the start of each level, the player's energy is restored, but points are retained from level to level.

(c) Optimizing Corridor Challenge

Applying optimal game design to Corridor Challenge requires specifying the parameters to optimize in the search for the optimal game, formulating the game as an MDP, and specifying the model for how the player's prior on concepts (θ) relates to the MDP parameters. The structure of Corridor Challenge allows for many variants that may differ in the expected information gain. To maximize expected information gain while keeping playing time relatively constant, we limited the game to two levels, with five islands per level. We then used optimal game design to select the number of points gained for opening a treasure chest, points lost for opening a trap chest, the energy lost when moving, the symbols that appeared on the chests and the Boolean concept used to categorize the chests. For simplicity, we assumed that the number of points gained (or lost) for a particular action is equal to the amount of energy gained (or lost) for that particular action.

[1] Corridor Challenge uses freely available graphics from http://www.lostgarden.com/2007/05/dancs-miraculously-flexible-game.html.

Given particular specifications for these variants of the game, we can define an MDP. Note that we define the MDP based on a player's beliefs, since these govern the player's actions, and these beliefs do not include knowledge of the true concept that governs the classification of the symbols.

States: the state is represented by the player's energy, her current level and position in the level, and the symbols on all chests in the current level.

Actions: the player can open the current chest (if there is one) or move to the next island.

Transition model: the player transitions to a new state based on opening a chest or moving to a new island. In both cases, the symbols on the chests stay the same, with the current symbol removed if the player opens the chest. If a player chooses to move, she knows what state will result: her position will move forward one space and her energy will be depleted by a known constant. If the result is negative energy, then the game transitions to a loss state. However, if a player opens a chest, her beliefs about what state will occur next is dependent on $p(h|\mathbf{d})$, her beliefs about the true concept given the data $\mathbf{d}$ she has observed so far. The player will gain energy if the symbol x on the current chest is in the concept. Taking an expectation over possible concepts h, this probability is $p(x \text{ in concept}) = \sum_h I(h \vdash x)p(h|\mathbf{d})$, where $I(h \vdash x) = 1$ if x is in the concept h and 0 otherwise. The probability of decreased energy is $(1 - p(x \text{ in concept}))$. Based on the Bayesian model mentioned earlier, the player's current beliefs $p(h|\mathbf{d})$ are dependent on the prior probability distribution over concepts. Thus, the transition model assumed by the player is dependent on the parameter θ that we would like to estimate, which is this prior distribution.

Reward model: When the player moves from one island to another, the reward model specifies $R(s, a, s') = 0$, and when the player opens a chest, $R(s, a, s')$ is a fixed positive number of points with probability $p(x \text{ in concept})$ and a fixed negative number of points with probability $(1 - p(x \text{ in concept}))$.

By using the MDP framework and assuming that the player updates her beliefs after seeing information, we ignore the value of information in understanding people's decisions; that is, we assume people make decisions based on their current information and do not consider the effect that information gained now will have on their future decisions. We examine ways to relax this assumption in §8.

5. Experiment 1: inferring difficulty

To test our framework, we first used the optimal game design procedure to find a version of Corridor Challenge with high expected information gain, and then ran an experiment in which players played either the optimized game or a randomly chosen game with lower expected information gain.

(a) Optimization of Corridor Challenge

We used simulated annealing [35] to stochastically search over possible designs of Corridor Challenge. The expected information gain of a game was approximated by sampling 35 possible θ vectors uniformly at random (reflecting a uniform prior on θ), simulating the actions of $n = 25$ players in the game, and using the simulated data to infer $p(\theta|\xi, \mathbf{a})$. We approximated $p(\theta|\xi, \mathbf{a})$ using the Metropolis–Hastings MCMC algorithm [34], with a Dirichlet proposal distribution centred at the current state. The parameter β for the Boltzmann policy was set to 1.

To execute the search, we parallelized simulated annealing by using several independent search threads. Every five iterations, the search threads pooled their current states, and each thread selected one of these states to continue searching from, with new starting states chosen probabilistically such that states with high information gain were more likely to be chosen. Each search state is a game, and the next state was found by selecting a parameter of the current game to perturb. If the selected parameter was real-valued, a new value was chosen by sampling from a Gaussian with small variance and mean equal to the current value; if the selected parameter was discrete, a new value was selected uniformly at random.

The stochastic search found games with significantly higher information gain than the initial games, regardless of starting point. This demonstrates that the evaluation and search procedure may be able to eliminate some trial and error in designing games for experiments. Qualitatively, games with high information gain tended to have a low risk of running out of energy, at least within the first few moves, and a diverse set of stimuli on the chests. These games also generally had positive rewards with larger magnitudes than the negative rewards. The game with the highest information gain used a true concept of type II, although several games with similarly high information gain had true concepts with different structures. While the information gain found for any given game is approximate, since we estimated the expectation over only a sample of possible priors, this was sufficient to separate poor games from relatively good games; we explore this relationship further in experiment 2.

(b) Methods

After optimizing Corridor Challenge, we conducted a behavioural experiment to assess whether an optimized game resulted in greater information gain than a random game.

Participants. Fifty participants were recruited online and received a small amount of money for their time.

Stimuli. Participants played Corridor Challenge with a parameters set based either on an optimized game (expected information gain of 3.4 bits) or on a random game (expected information gain of 0.6 bits).[2] The symbols differed along the dimensions of shape, colour and pattern.

Procedure. Half of the participants were randomly assigned to each game design, and played the game in a web browser. Regardless of condition, the participants were shown text describing the structure of Corridor Challenge, and then played several practice games to familiarize them with interface. The first practice game simply had chests labelled ‘good’ and ‘bad’; the next three games used Boolean concepts of increasing difficulty based on previous work. All practice games used different symbols from one another and from the final game. Practice games used the point and energy values from the game chosen for their condition (i.e. the random game or the game found by the search) in order to make players aware of these values, but the symbols in the practice games were identical across conditions. The fifth and final game was chosen based condition: either the optimized game or the random game. After completing the final game, participants were asked to rate how fun and how difficult the game was, both on 7-point Likert scales. Additionally, they were shown the stimuli and categorization information that they observed during the final game, and asked to classify the remaining stimuli from the game that were not observed.

(c) Results

To assess the information gained from each game, we calculated posterior distributions over the prior probability of each Boolean concept based on the players’ actions. These distributions were calculated using a Metropolis–Hastings algorithm on both the prior and the noise parameter β. Samples were generated from five chains with 100 000 samples each; the first 10% of samples from each chain were removed for burn-in. To infer the actual information gained for each game, we infer the maximum-likelihood Dirichlet distribution based on these samples from the posterior. We then calculate the entropy of the inferred Dirichlet. The difference between this entropy and the entropy of the (uniform) prior distribution is the actual information gain of the game.

Figure 3 shows the inferred distribution over the prior probability of learning a concept of each type (θ_i) based on participants’ actions for the optimized game and the random game; if a concept has higher prior probability, it will be easier to learn. Qualitatively, the distributions inferred from the optimized game appear more tightly concentrated than those from the random

[2] All game specifications and data recording participants’ gameplay can be found at http://cocosci.berkeley.edu/data/optimal-games-data/OptimalGameDesignData.html.

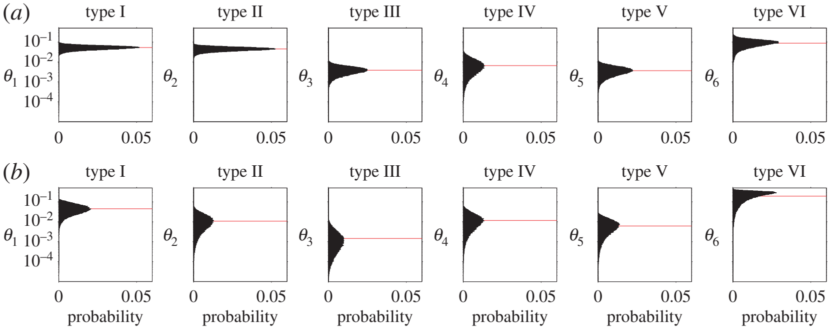

Figure 3. Results of experiment 1, in the form of posterior distributions on concept difficulty from participants' responses in (*a*) the optimized game and (*b*) the random game; red lines indicate the mean of each distribution. Each panel shows the distribution over the inferred difficulty of a concept with the given structure (types I–VI), as reflected by its prior probability in the Bayesian model. Concepts with higher prior probability are easier to learn. Note the logarithmic scale on the prior probability of each θ_i.

game; this is confirmed by the actual information gain, which was 3.30 bits for the optimized game and 1.62 bits for the random game. This implies that we could halve the number of participants by running the optimized game rather than the random game, while achieving the same level of specificity. The amount of information gained by the optimized game was very similar to the predicted information gain, while for the random game, somewhat more information was gained than predicted (1.62 bits versus 0.6 bits). The higher information gain could be the result of error in the prediction of expected information gain: the calculation of this quantity is an approximation, with only a finite number of possible θ vectors sampled to create simulated action vectors. The discrepancy could also result from particulars of the true value of θ. The expected utility is calculated over all possible θ, since its true value is unknown. However, when calculating actual information gain, only the value of θ that reflects participants' cognitive processes is relevant. While the issue of approximation can be lessened by sampling additional θ, the latter issue is inherent to the process of predicting information gain. In experiment 2, we further explore the connection between expected and actual information gain.

For both games, the ordering of the mean prior probabilities of a concept of each type, shown by red lines in figure 3, is the same as that found in previous work, except for type VI. Our inferred distributions for a concept of type VI placed significant probability on a broad range of values, suggesting that we simply did not gain much information about its actual difficulty. We do infer that type VI is easier than types III, IV or V, consistent with some previous findings [39].

6. Experiment 2: estimating information gain

To verify the relationship between actual and expected information gain, we conducted a second experiment in which players played games with a range of information gains. To isolate the impact of the symbols on the chests and the true concept, we fixed the point structures to those found for the optimized game in experiment 1 and conducted new searches over the remaining variables. We then selected games that had varying expected information gains, demonstrating that even without changing the incentive structure a range of information gains was possible.

(a) Methods

Participants. A total of 475 participants were recruited online and received the same payment as in experiment 1.

Stimuli. Participants played one of 19 new games. The 19 new game designs were selected by recording the game design and expected information gain for each iteration of simulated

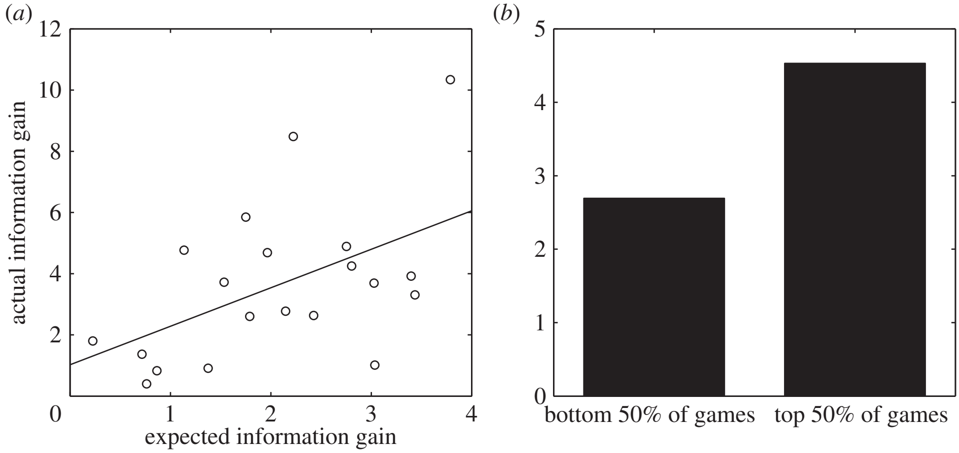

Figure 4. Results of experiment 2. (*a*) Expected versus actual information gains ($r(18) = 0.51, p < 0.025$). Each circle represents a game, and the least-squares regression line is shown. (*b*) Actual information gain for the 10 games with lowest expected information gain versus the highest expected information gain.

annealing. For this experiment, the points were fixed to be the same as in the optimized game in experiment 1, so the search only varied the object on each chest and the true category. We then ran 19 independent search processes, each with a different initial game. From the resulting games, we hand-selected one game from each search thread such that the new collection of games had roughly evenly spaced information gains.

Procedure. Procedure matched experiment 1.

(b) Results

We compared the actual and expected information gains for the 19 new games and the optimized game from experiment 1, all of which used the same point structure. As shown in figure 4*a*, expected and actual information gain were positively correlated ($r(18) = 0.51$, $p < 0.025$). While this correlation might seem modest, it has significant consequences for efficiency: on average, the 10 games with the highest expected utility resulted in a gain of 67% more bits of information than the 10 games with lowest expected utility (figure 4*b*).

One potential objection to the optimal game design framework is that considerable computational power is necessary to predict expected utility. We thus explored whether heuristics based on features of the game might be sufficient to predict information gain. As shown in figure 5*a*, the expected utility of the games showed the highest correlation with actual information gain, although the total number of unique symbols was also positively correlated with information gain ($r(18) = 0.46$, $p < 0.05$). While optimal game design and this heuristic have relatively similar correlations with information gain, we believe that there is still an advantage in using the optimal game design framework, as this approach does not require generating appropriate heuristics for different games and it may not be intuitively obvious which heuristics will be good predictors of information gain. For example, the total number of treasure chests was negatively correlated with information gain, although this correlation was not significant. Additionally, as the number of features to optimize increases, the number of possible heuristics will also increase, making it difficult to choose a heuristic to rely on via intuition; we return to this issue in experiment 3.

7. Experiment 3: sensitivity to rewards

In experiment 2, we showed that expected and actual information gain were correlated for a range of game designs. All of these game designs had the same incentive structure; the only differences

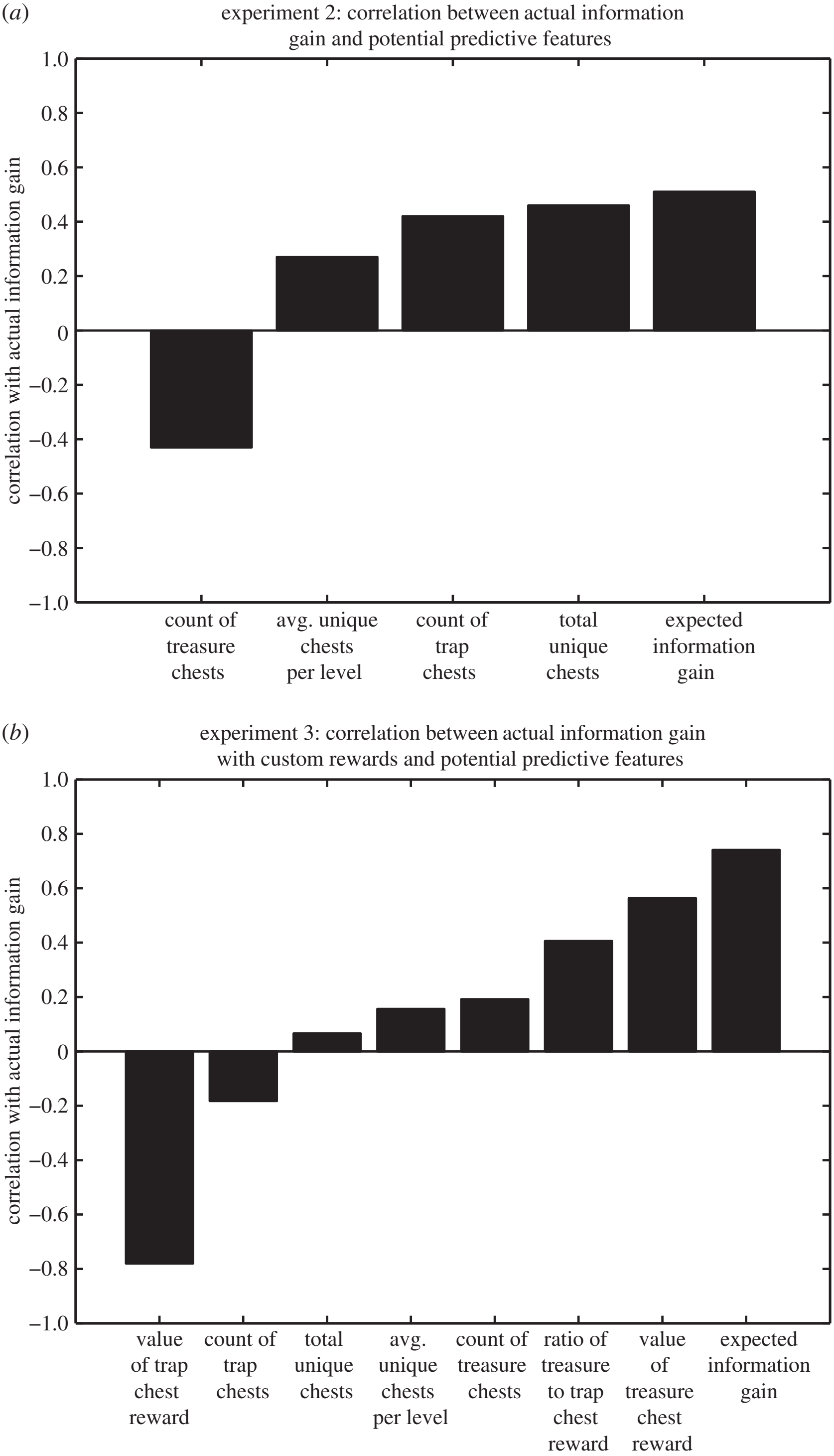

Figure 5. Correlations between proposed heuristics for predicting information gain and the actual information gain. (*a*) Correlations for experiment 2. Expected utility (as calculated using optimal game design) has the highest correlation, and number of unique chests is the only heuristic with a significant correlation to actual information gain. (*b*) Correlations for experiment 3. The correlations to information gain for the value of the trap chest and for expected utility are of similar magnitude, but only expected utility is consistently predictive across the two experiments.

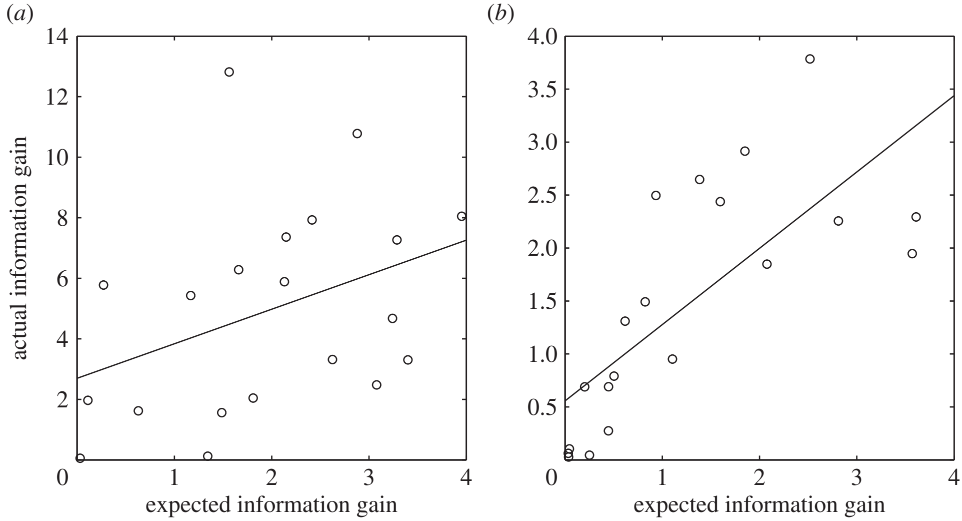

Figure 6. Results of experiment 3. (*a*) Expected versus actual information gains with point-based reward ($r(18) = 0.38$, $p = 0.1$). Each circle represents a game, and the least-squares regression line is shown. (*b*) Expected versus actual information gains with inferred custom rewards ($r(18) = 0.74$, $p < 0.001$).

were the categories being learned and the placement of items within the games. It is also possible to vary the rewards in the game designs, as was done in experiment 1. This raises the question of how much people will internalize the reward structure and behave rationally with respect to it. People may incorporate their own goals into the game, such as wanting to learn about the true concept rather than maximize points, and thus exhibit unexpected behavioural changes based on different reward structures. To investigate this possibility, we generated 18 additional games with a range of expected information gains, allowing the incentive structure as well as the other parameters of the game design to change.

(a) Methods

Participants. A total of 450 participants were recruited online and received the same payment as in experiments 1 and 2.

Stimuli. Participants played one of 18 new games. The search method for this experiment was the same as for experiment 2 except that the point values for opening a treasure or trap chest and the energy lost for movement were allowed to vary. All games came from search threads with independent starting points, and games were hand-selected to span a range of expected information gains.

Procedure. Procedure matched experiment 1.

(b) Results

We analysed the data from these new games combined with the data from the two games in experiment 1. We first calculated the actual information gain about the prior distribution over concept types, assuming the participants' reward functions reflected the point structure. As shown in figure 6*a*, the correlation between expected and actual information gain was not significant ($r(18) = 0.38$, $p = 0.1$). Inspection of participants' actions showed some choices that seemed unlikely to be rational with respect to the model. For instance, a participant might choose to open a chest even when she had low energy and little information about the concept, despite the fact that she could reach the end of the level without this action and earn the large level completion bonus. From the perspective of the model, this action is only predicted if the participant places very high probability on this being a treasure chest.

To test whether participants might be acting based on a different reward function than that given by the incentive structure, we modified the inference procedure to infer a reward function for each game based on the participants' actions. Previously, the inference procedure inferred a posterior distribution over the hypothesis space of six-dimensional multinomials; now, we changed the hypothesis space to be possible reward functions. These functions were specified by the value of opening a treasure chest, the value of opening a trap chest and the value of completing the level. We constrained these values such that the value of opening a treasure chest and of completing the level were non-negative and the value of opening a trap chest was non-positive. We fixed the prior distribution to be equal to the mean of the posterior distribution from the optimized game in experiment 1.[3] MCMC sampling then proceeded as in experiment 1, resulting in a posterior distribution over the values for the parameters of the reward function.

The results showed that participants do seem to be acting based on a different reward function than that given by the point structure. While the reward function varies across games, as expected given that the point structure is likely to have some influence on behaviour, it consistently places relatively low value on completing the level. This could reflect the fact that completing a level is not inherently rewarding to participants. Participant comments are consistent with people being more motivated by understanding the game than achieving maximal points. For instance, one of the most frequent comments by those who did not enjoy the game was that they were 'confused' by the rule or that they did not understand the pattern. Thus, opening chests might be expected to have higher intrinsic reward than completing the level, despite the point structure.

One of the goals of inferring the participants' reward functions was to determine whether using the inferred functions would lead to a correlation between expected and actual information gain. If the confounding factor in the original analysis was the incorrect reward functions, then using these functions to re-calculate both the expected and actual information gains should lead to similar results as in experiment 2. Thus, we fixed the reward function for each game to match the mean of the posterior distribution over reward functions for that game, and then used the original inference algorithm to infer a posterior distribution over the difficulty of learning the six different concept types. As shown in figure 6*b*, expected and actual information gain are in fact correlated when the inferred custom reward functions are used ($r(18) = 0.74$, $p < 0.001$). This demonstrates the importance of knowing participants' goals when interpreting their actions. A participant's actions are only meaningful within the context of her goals and her understanding of how her actions affect her progress towards those goals. While participant actions can be used to make inferences about these factors, this may lead to incorrect conclusions if our assumptions about the relevant factors are wrong.

To determine whether heuristics would also be effective predictors of information gain, we calculated seven heuristics based on the characteristics of the games. Four were the same heuristics as in experiment 2, while three were based on the reward functions, which were the same across all games in experiment 2. We used the inferred custom rewards for these heuristics since the original rewards were inconsistent with participant behaviour. As shown in figure 5*b*, some of these heuristics are quite good at predicting expected utility. The value of a trap chest even has a slightly higher magnitude correlation with information gain than expected utility ($r(18) = -0.78$, $p < 0.001$). Heuristics thus can be effective at predicting information gain. However, their effectiveness seems to be less consistent than expected utility: the best heuristic for experiment 2, the total number of unique symbols, has only a small correlation with information gain for experiment 3 ($r(18) = 0.066$, $p > 0.7$), and the best heuristic for experiment 3, the value of a trap chest, would have no correlation with information gain for the games in experiment 2 since those games all shared the same reward function. By contrast, expected utility as calculated

[3] In principle, one could jointly infer both an arbitrary reward function and the prior distribution, but in practice, this leads to identifiability issues wherein very different parameter configurations all have similar posterior probability. Since our interest here is whether there exists a custom reward configuration that would explain the participants' actions and we have a good estimate of the prior distribution from the previous game, fixing the prior distribution gives the best estimate of the reward functions.

by optimal game design is highly correlated with information gain for both sets of games. This suggests that the computational cost of optimal game design is balanced by its greater consistency.

8. Extensions to the Markov decision process framework

We have demonstrated how MDPs can be used to interpret players' actions within games. Our results show that we can predict which games will result in more information gain and that our inferences about the difficulty of Boolean concepts are consistent with previous work. However, our analyses do not consider players' reasoning about how their actions will affect the information they have to be successful in the game. When a player opens a chest, she gains information about the true concept governing the meaning of the symbols, which might lead her to open a chest primarily to improve her ability to choose actions in the future. While incorporating, this factor is more computationally challenging than the analysis we have provided thus far, we now explore two extensions to the MDP framework related to players' reasoning about changes in their knowledge based on their actions. First, we use partially observable MDPs to analyse the results of experiment 1, determining whether ignoring this factor is leading to incorrect inferences. Then, we explore a computationally tractable approximation to information gain that can be included in the MDP analysis.

(a) Analysing actions via partially observable Markov decision processes

In the MDP formulation, we assume that players' beliefs about the true concept govern the transition and reward models that they believe are operating in the game. These beliefs are updated after they open a chest, but the Q-values for each action do not take into account the fact that when a chest is opened, players' beliefs will change. An alternative analysis that does take this information into account is to assume that the true concept is an unobserved part of the state of the game. The player then does not know the state of the game at each time point, but may gain information about the state by opening chests. Since knowing the state may prove advantageous for choosing actions, this may lead the player to open more or different chests.

Including information in the state that is unobserved corresponds to modelling the game as a partially observable Markov decision process (POMDP). POMDPs are frequently used to model sequential decision-making situations where relevant information cannot be directly observed [40,41]. Formally, a POMDP is defined by the same components as an MDP (the tuple $\langle S, A, T, R, \gamma \rangle$), plus an observation model O and a set of possible observations Z. The observation model defines conditional probability distributions $p(z|s,a)$ for $z \in Z$. Since all or part of the state is unobserved, the observations can be used to gain information about the underlying state. For example, in the case of Corridor Challenge, the observations correspond to seeing that a chest is a trap or a treasure, and based on the symbol that was on the opened chest, this observation will rule out particular concepts. Because parts of the state are unobserved in a POMDP, the Markov property making future actions independent of past actions no longer holds. Agents must consider the history of past actions and observations in their choices, as these can be used to make inferences about the state. Typically, agents are modelled as choosing actions based on their *belief state* $b(s_t)$, which is the distribution over possible states at time t given the actions taken and observations collected. This is a sufficient statistic for representing information from past actions and observations. After each action, the belief state is updated based on the transition and reward model. For a given state i, the updated probability $b(s_{(t+1)} = i)$ that the game is in state i at time $t + 1$ is

$$b(s_{(t+1)} = i) \propto \sum_{j \in S} b(s_t = j) p(s_{(t+1)} = i | s_t = j, a_t) p(z_{t+1} | s_{t+1} = i, a_t). \tag{8.1}$$

Computationally, this dependence on the past means that POMDPs are much more expensive than MDPs. While a number of POMDP solution methods have been developed (e.g. [42–44]), most make use of the assumption that actions are chosen optimally, allowing lower value actions to be ignored. To account for the fact that people do not always choose the optimal action, we can approximate a Boltzmann policy for a POMDP by assuming $p(a_t|b(s_t)) \propto \exp(\beta Q^*(b(s_t), a))$, where $Q^*(b(s_t), a)$ is the optimal Q-function for a given action and belief state; this is consistent with other work using Boltzmann policies for POMDPs [45]. Calculating this policy requires access to the Q-functions, which eliminates many approximate POMDP solution methods that calculate only upper and lower bounds on the Q-function in order to obtain an (approximate) optimal policy more quickly.

To explore the results obtained using a POMDP representation, we modelled the games in experiment 1 using a POMDP. As described earlier, the state is now composed of the unobserved concept governing the meaning of the symbols as well as the observed components present in the MDP representation: the symbols remaining in the level and the energy of the player. The actions and the component of the transition model corresponding to the observed part of the state are the same as in the MDP representation. This definition ignores the fact that in Level 1, the next level of the game could also be treated as an unobserved part of the state. However, there are 9^5 possible configurations of the chests in the second level (five spots for chests, each of which can be empty or have one of eight chests). This results in an extremely large number of states if all possibilities for the second level are considered; assuming that people are not considering all possibilities for the second level and may not even have been paying attention to the directions indicating that there is a second level, we ignore these possibilities in our definition of the state. Since the underlying concept is the same through the game, the transition model for the unobserved concept is the identity. The reward function also does not change for the POMDP representation. The observation model has $p(z = \text{treasure}|s, a = open)$ equal to 1 if the symbol at the current location is in the underlying concept for state s and 0 otherwise; similarly, $p(z = \text{trap}|s, a = open) = 1$ if the symbol at the current location is not in the underlying concept for state s. If the action is *move* rather than *open*, the observation is null: no new information is gained about the underlying concept. Given these definitions for each part of the POMDP, we generated a POMDP specification for each level of both of the games in experiment 1, and then used the pomdp-solve library to solve for a POMDP policy for each level [46]. pomdp-solve uses actual Q-values, rather than bounds, and thus its output can be used to find the approximate Boltzmann policy described earlier.

To use the POMDP solution to infer the difficult of Boolean concepts, we must tie the prior distribution over concepts, θ, to the players' actions in the game. Since the belief state $b(s_t)$ represents the player's beliefs about the probability of each concept after the first t actions and observations, the initial belief state $b(s_0)$ that characterizes the player's beliefs before seeing any information corresponds to the prior distribution over concepts. Thus, we can use MCMC sampling to find a distribution over possible priors, with the only change for different samples being the initial belief state. Unlike in the MDP analysis, the transition and reward models are the same across samples.

Using the same sampling procedure as for the MDP analysis, we found that our inferences about the prior for the two games in experiment 1 were consistent with the results of our initial analysis. For both games, we find that type I concepts are easier than type II concepts, which are easier than concepts of type III, IV or V (figure 7). The difference between a concept in type II and a concept in type IV was less pronounced than in the MDP analysis, but the POMDP analysis still finds $\theta_2 > \theta_4$. As before, the difficulty of type VI concepts is less clear, with little information gained and a some samples in which θ_6 was relatively large. The consistency between these results and those in experiment 1 is encouraging, as it suggests that at least in some cases, the inferences made by the MDP analysis are a good approximation for the less tractable POMDP analysis.

(b) Incorporating information gain into the reward function

While our analysis suggests that including the value of information in the form of a POMDP does not affect the inferences that result from our modelling framework for experiment 1, it remains possible that people are sensitive to this factor. Since POMDP planning is computationally intensive, it is valuable to consider a more tractable strategy for incorporating the value of

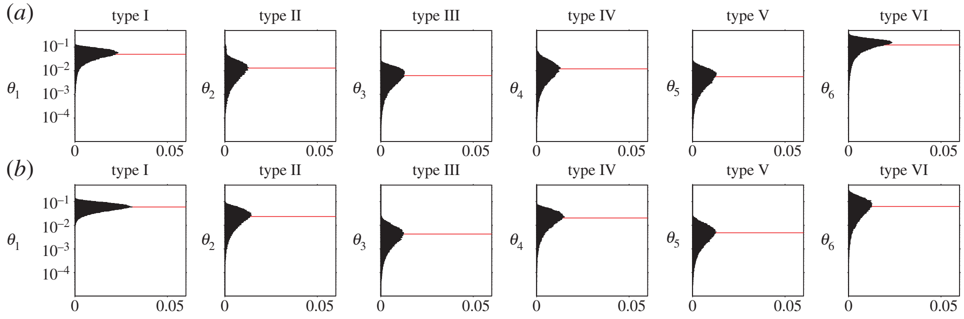

Figure 7. Results of analysing data from experiment 1 using POMDPs, in the form of posterior distributions on concept difficulty from participants' responses in (*a*) the optimized game and (*b*) the random game; red lines indicate the mean of each distribution. Each panel shows the distribution over the inferred difficulty of a concept with the given structure (types I–VI). Note the logarithmic scale on the prior probability of each θ_i.

information into our model of decision-making. One way to do this is to adopt the approach we used in experiment 3, estimating the contribution of value of information as an aspect of the reward function. We augment the reward function in the MDP analysis with a term for the player's expected information gain from the chosen action; that is, how much more certainty will the player be likely to have about the concept after taking this action than before taking the action? While the player's expected information gain in a single step does not directly account for whether the new certainty will enable better choices, it provides a computationally inexpensive way to model the fact that in some cases, people may be motivated by learning more about the underlying concept.

To include information gain in the reward function, we set $R_{IG}(s,a,s') = R(s,a,s') + w \cdot \Delta H$, where $R(s,a,s')$ is the reward without including information gain and ΔH is the change in entropy in the player's estimated posterior distribution over concepts based on observing the results of taking action a in state s. The parameter w controls the weighting of information gain within the reward function. We then used the players' actions to infer a posterior distribution over the parameter w as well as the parameters in the original reward function, fixing the transition function to the mean of the inferred transition function from the optimized game in experiment 1. The only difference between this procedure and that in experiment 3 is that w can have non-zero weight.

Using this procedure, we found the weight of information gain in each of the reward functions for the 20 games in experiment 3. We used data from experiment 3 due to the evidence we found in our analysis of the custom reward functions that players may have been motivated by their own goals in these games. Information gain was always inferred to have a positive weight, except in one game where inspection of the samples showed that this parameter was covarying with the point value of opening a treasure chest. This suggests that there are cases where people are sensitive to information gain.

To explore how the new reward functions affected the relationship between expected and actual information gain, we then fixed the reward functions and inferred the prior distribution over Boolean concepts, again as in experiment 3. As shown in figure 8, expected and actual information gain were correlated ($r(18) = 0.82$, $p < 0.001$). This correlation value is similar to that found in experiment 3 ($r(18) = 0.74$). Using the deviance information criterion (DIC; [47]), we compared the fit of the model with information gain and the model without information gain for each game. DIC is related to other information criterion measures and controls for differences between models in the effective number of parameters. This measure is easily computed from MCMC samples; lower DIC reflects a better model fit. The average DIC over the 20 games was 203 for the models with information gain (median: 204), compared to 212 for the models

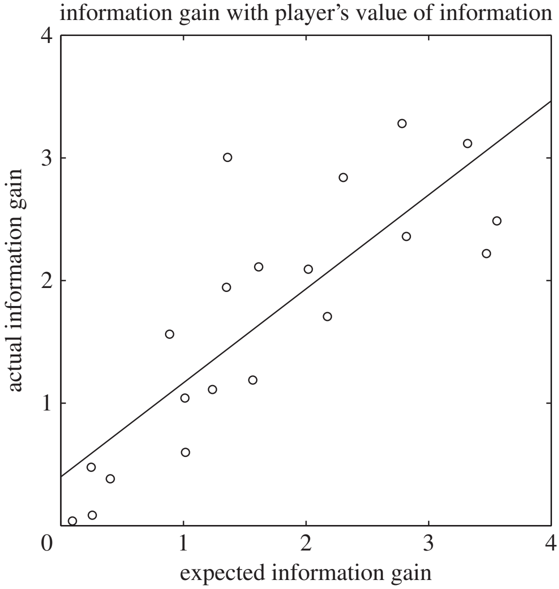

Figure 8. Results of experiment 3 when an inferred reward function that includes information gain is used to estimate the prior distribution over Boolean concepts ($r(18) = 0.82$, $p < 0.001$). Each circle represents a game, and the least-squares regression line is shown.

without information gain (median: 220). This difference suggests that the model which includes information gain is a somewhat better fit to the data than the model without this parameter. Together with the previous analysis, this suggests that while value of information does not always significantly affect players' choices, there are cases where players' behaviour likely reflects some attention to this factor. In general, players may be influenced by a combination of incentives, some extrinsic and provided by the game and some intrinsic and reflecting a desire to learn and understand. Including a factor related to information gain in the reward function is a tractable approximation for the full POMDP analysis, and if this factor was set prior to the optimization, optimal game design could incorporate information gain when searching for the best game design.

9. Discussion

Refining a game to be diagnostic of psychologically or educationally relevant parameters can be a time-consuming process filled with trial and error. While the exact incentive structure or placement of objects in a game may have little impact on how enjoyable or engaging the game is, these factors can significantly impact how useful the game is for diagnosing players' knowledge. We have presented a general framework for deciding how to set these factors by predicting which game designs will have the highest expected information gain. This framework adapts ideas from optimal experiment design and relies on MDPs to link players' actions to cognitively relevant model parameters. We now consider several possible challenges as well as future directions for this framework.

Our framework relies upon the idea that people behave in noisily optimal ways based on their current knowledge and goals. In experiment 3, we saw that invalid inferences can be drawn when one makes incorrect assumptions about one of these factors. Thus, care must be taken to monitor whether the MDP model is a good fit to participants' behaviours, especially since optimization can magnify errors due to model assumptions. One recourse is to modify the task, instructions or model to more closely align the model's assumptions and the participant's beliefs.

For instance, to more closely align model and participant reward functions in experiment 3, one might give points for opening a chest with a previously unopened symbol, making the game's incentive structure closer to that which participants bring to the task. Alternatively, one might give monetary rewards based on the points the participant earned, making it more likely that participants will respond based on the game's reward structure. Results in behavioural economics suggest that aligning monetary incentives with participants' performance results in choices that are more consistent with behaviour outside of an experimental setting, as choices in the experiment have real consequences (e.g. [48]); for games, this is likely to result in behaviour that is more consistent with the game's incentive structure, since this structure will have an effect on participants' monetary rewards.

In §8, we considered extensions to the MDP framework that could allow us to incorporate the value of information into understanding players' actions. We highlighted a tension between incorporating the exact value of information, as in the POMDP analysis, and using an approximation that could be included within the optimization of game design. Incorporating information gain may be especially important in the case of games involving inquiry skills, where part of the challenge is in determining what information to gather to solve a problem. In some cases, the approximation of including the information gain at only a single step (as in the modified reward functions in §8) may be too limited, especially if players must use several actions to uncover information in the game. In that case, it may be necessary to consider other approximations or to attempt to use a POMDP analysis when using the optimal game design framework. Correlations between actual and expected information gain are likely to be much higher when the same representation (MDP or POMDP) is used for prediction and analysis of results. The use of a Boltzmann policy significantly slows computation with the POMDP, since many approximation methods cannot be used. To allow POMDPs to be used within the optimization, one might relax the assumption of a Boltzmann policy for prediction and instead use an optimal policy, reverting to the Boltzmann policy when analysing behavioural results. Exploration of this strategy is necessary to determine its effectiveness in the types of inquiry-oriented games where POMDPs are likely to be most beneficial.

Games are increasingly popular tools for research in a variety of areas, including psychology, cognitive science and education. While games address some issues in traditional experiments, such as flagging motivation or difficulty introducing participants to a complex task, they also create new challenges: it can be difficult to interpret game data to draw conclusions about the research questions, and there may be many possible game designs that a researcher could use, without clear reasons to choose one over the other. We have created an optimal game design framework that provides a way to guide game design and choose designs that are more diagnostic. Beyond providing a principled approach to choosing a game design, more diagnostic games offer key benefits for research. More diagnostic games allow fewer participants to be used to gain the same information about a research question, providing the potential for drawing conclusions more quickly or for asking more complex questions that would otherwise require a prohibitive number of participants. In education, games can be used to assess students' knowledge. By diagnosing knowledge more accurately or over a shorter period of time, instruction can be better targeted to individual learners and more time can be spent on learning rather than assessing. While there are many future directions in which the framework could be extended, this work provides a starting point for more principled approaches to designing games for education and behavioural research.

Funding statement. This work was supported by a Department of Defense NDSEG Fellowship to A.N.R., a Google PhD Fellowship to M.Z. and NSF grant no. IIS-0845410 to T.L.G.

References

1. Michael D, Chen S. 2005 *Serious games: games that educate, train, and inform*. Boston, MA: Thomson Course Technology.

2. Von Ahn L. 2006 Games with a purpose. *Computer* **39**, 92–94. (doi:10.1109/MC.2006.196)
3. Siorpaes K, Hepp M. 2008 Games with a purpose for the semantic web. *Intell. Syst.* **23**, 50–60. (doi:10.1109/MIS.2008.45)
4. Klopfer E. 2008 *Augmented learning: research and design of mobile educational games*. Cambridge, MA: MIT Press.
5. Puglisi A, Baronchelli A, Loreto V. 2008 Cultural route to the emergence of linguistic categories. *Proc. Natl Acad. Sci. USA* **105**, 7936–7940. (doi:10.1073/pnas.0802485105)
6. Barab SA, Dede C. 2007 Games and immersive participatory simulations for science education: an emerging type of curricula. *J. Sci. Educ. Technol.* **16**, 1–3. (doi:10.1007/s10956-007-9043-9)
7. Barab SA, Scott B, Siyahhan S, Goldstone R, Ingram-Goble A, Zuiker SJ, Warren S. 2009 Transformational play as a curricular scaffold: using videogames to support science education. *J. Sci. Educ. Technol.* **18**, 305–320. (doi:10.1007/s10956-009-9171-5)
8. Ketelhut DJ. 2007 The impact of student self-efficacy on scientific inquiry skills: an exploratory investigation in River City, a multi-user virtual environment. *J. Sci. Educ. Technol.* **16**, 99–111. (doi:10.1007/s10956-006-9038-y)
9. Jacobson MJ, Kozma RB. 2000 *Innovations in science and mathematics education: advanced designs for technologies of learning*. London, UK: Lawrence Erlbaum.
10. Gee J. 2007 *What video games have to teach us about learning and literacy*. New York, NY: Palgrave Macmillan.
11. Marino MT, Beecher CC. 2010 Conceptualizing RTI in 21st-century secondary science classrooms: video games' potential to provide tiered support and progress monitoring for students with learning disabilities. *Learn. Disabil. Q.* **33**, 299–311.
12. Papastergiou M. 2009 Digital game-based learning in high school computer science education: impact on educational effectiveness and student motivation. *Comput. Educ.* **52**, 1–12. (doi:10.1016/j.compedu.2008.06.004)
13. Atkinson AC, Donev AN, Tobias RD. 2007 *Optimum experimental designs, with SAS*. New York, NY: Oxford University Press.
14. Pukelsheim F. 2006 *Optimal design of experiments*, vol. 50. Philadelphia, PA: Society for Industrial and Applied Mathematics.
15. Dantas L, Orlande H, Cotta R. 2002 Estimation of dimensionless parameters of Luikov's system for heat and mass transfer in capillary porous media. *Int. J. Therm. Sci.* **41**, 217–227. (doi:10.1016/S1290-0729(01)01310-2)
16. Emery AF, Nenarokomov AV, Fadale TD. 2000 Uncertainties in parameter estimation: the optimal experiment design. *Int. J. Heat Mass Transf.* **43**, 3331–3339. (doi:10.1016/S0017-9310(99)00378-6)
17. Fujiwara M, Nagy ZK, Chew JW, Braatz RD. 2005 First-principles and direct design approaches for the control of pharmaceutical crystallization. *J. Process Control* **15**, 493–504. (doi:10.1016/j.jprocont.2004.08.003)
18. Shin S, Han S, Lee W, Im Y, Chae J, Lee D-i, Lee W, Urban Z. 2007 Optimize terephthaldehyde reactor operations. *Hydrocarbon Process.* **86**, 83–90.
19. Bruno R *et al.* 1998 Population pharmacokinetics/pharmacodynamics of docetaxel in phase II studies in patients with cancer. *J. Clin. Oncol.* **16**, 187–196.
20. D'Argenio DZ. 1981 Optimal sampling times for pharmacokinetic experiments. *J. Pharmacokinet. Pharmacodyn.* **9**, 739–756. (doi:10.1007/BF01070904)
21. Haines LM, Perevozskaya I, Rosenberger WF. 2003 Bayesian optimal designs for phase I clinical trials. *Biometrics* **59**, 591–600. (doi:10.1111/1541-0420.00069)
22. Simon R. 1989 Optimal two-stage designs for phase II clinical trials. *Control. Clin. Trials* **10**, 1–10. (doi:10.1016/0197-2456(89)90015-9)
23. Derlinden EV, Bernaerts K, Impe JV. 2010 Simultaneous versus sequential optimal experiment design for the identification of multi-parameter microbial growth kinetics as a function of temperature. *J. Theor. Biol.* **264**, 347–355. (doi:10.1016/j.jtbi.2010.01.003)
24. Elvind D, Asmund H, Rolf V. 1992 Maximum information at minimum cost: a north sea field development study with an experimental design. *J. Pet. Technol.* **44**, 1350–1356. (doi:10.2118/23139-PA)
25. Ajo-Franklin JB. 2009 Optimal experiment design for time-lapse traveltime tomography. *Geophysics* **74**, Q27–Q40. (doi:10.1190/1.3141738)
26. Chaloner K, Verdinelli I. 1995 Bayesian experimental design: a review. *Stat. Sci.* **10**, 273–304. (doi:10.1214/ss/1177009939)

27. Cavagnaro DR, Myung JI, Pitt MA, Kujala JV. 2010 Adaptive design optimization: a mutual information-based approach to model discrimination in cognitive science. *Neural Comput.* **22**, 887–905. (doi:10.1162/neco.2009.02-09-959)
28. Myung J, Pitt M. 2009 Optimal experimental design for model discrimination. *Psychol. Rev.* **116**, 499–518. (doi:10.1037/a0016104)
29. Erev I, Roth A. 1998 Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* **88**, 848–881.
30. Andrade G, Ramalho G, Santana H, Corruble V. 2005 Extending reinforcement learning to provide dynamic game balancing. In *Proc. of the Workshop on Reasoning, Representation, and Learning in Computer Games, 19th IJCAI*, pp. 7–12. Published online as Technical Report AIC-05-127. Washington, DC: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence.
31. Tan C, Cheng H. 2009 IMPLANT: an integrated MDP and POMDP learning ageNT for adaptive games. In *Proc. of the Artificial Intelligence and Interactive Digital Entertainment Conf., Stanford, CA, October 2009*, pp. 94–99. Menlo Park, CA: AAAI Press.
32. Sutton RS, Barto AG. 1998 *Reinforcement learning*. Cambridge, MA: MIT Press.
33. Baker CL, Saxe RR, Tenenbaum JB. 2009 Action understanding as inverse planning. *Cognition* **113**, 329–349. (doi:10.1016/j.cognition.2009.07.005)
34. Gilks W, Richardson S, Spiegelhalter, DJ (eds). 1996 *Markov Chain Monte Carlo in practice*. Suffolk, UK: Chapman and Hall.
35. Kirkpatrick S, Gelatt C, Vecchi M. 1983 Optimization by simulated annealing. *Science* **220**, 671–680. (doi:10.1126/science.220.4598.671)
36. Shepard RN, Hovland CI, Jenkins HM. 1961 *Learning and memorization of classifications*. Psychological Monographs: General and Applied, vol. 75(13). Washington DC: American Psychological Association.
37. Nosofsky RM, Gluck M, Palmeri TJ, McKinley SC, Glauthier P. 1994 Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Mem. Cogn.* **22**, 352–369. (doi:10.3758/BF03200862)
38. Feldman J. 2000 Minimization of Boolean complexity in human concept learning. *Nature* **407**, 630–633. (doi:10.1038/35036586)
39. Griffiths TL, Christian BR, Kalish ML. 2008 Using category structures to test iterated learning as a method for identifying inductive biases. *Cogn. Sci.* **32**, 68–107. (doi:10.1080/03640210701801974)
40. Monahan G. 1982 A survey of partially observable Markov decision processes: theory, models, and algorithms. *Manag. Sci.* **28**, 1–16. (doi:10.1287/mnsc.28.1.1)
41. Kaelbling L, Littman M, Cassandra A. 1998 Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134. (doi:10.1016/S0004-3702(98)00023-X)
42. Pineau J, Gordon GJ, Thrun S. 2006 Anytime point-based approximations for large POMDPs. *J. Artif. Intell. Res.* (*JAIR*) **27**, 335–380.
43. Kurniawati H, Hsu D, Lee WS. 2008 Sarsop: efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Proc. Robotics: Science and Systems, Zurich, Switzerland, June 2008* (eds O Brock, J Trinkle, F Ramos), pp. 65–72. Cambridge, MA: MIT Press.
44. Spaan M, Vlassis N. 2005 Perseus: randomized point-based value iteration for POMDPs. *J. Artif. Intell. Res.* **24**, 195–220.
45. Ramírez M, Geffner H. 2011 Goal recognition over POMDPs: inferring the intention of a POMDP agent. In *Proc. 22nd Int. Joint Conf. on Artificial Intelligence, Barcelona, Spain, July 2011*, pp. 2009–2014. Menlo Park, CA: AAAI Press.
46. Cassandra AR. 2005 pomdp-solve, version 5.3. See http://www.pomdp.org/.
47. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. 2002 Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* (*Stat. Methodol.*) **64**, 583–639. (doi:10.1111/1467-9868.00353)
48. Oxoby RJ. 2006 Experiments and behavioral economics. In *Handbook of contemporary behavioral economics: foundations and developments* (ed. M Altman), pp. 441–454. Armonk, NY: M. E. Sharpe, Inc.