How do People Solve the "Weather Prediction" Task?: Individual Variability in Strategies for Probabilistic Category Learning

Mark A. Gluck,[1,3,4] Daphna Shohamy,[1,3] and Catherine Myers[2,3]

[1]Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, New Jersey 07102, USA

[2]Department of Psychology, Rutgers University, Newark, New Jersey 07102, USA

Probabilistic category learning is often assumed to be an incrementally learned cognitive skill, dependent on nondeclarative memory systems. One paradigm in particular, the weather prediction task, has been used in over half a dozen neuropsychological and neuroimaging studies to date. Because of the growing interest in using this task and others like it as behavioral tools for studying the cognitive neuroscience of cognitive skill learning, it becomes especially important to understand how subjects solve this kind of task and whether all subjects learn it in the same way. We present here new experimental and theoretical analyses of the weather prediction task that indicate that there are at least three different strategies that describe how subjects learn this task. (1) An optimal multi-cue strategy, in which they respond to each pattern on the basis of associations of all four cues with each outcome; (2) a one-cue strategy, in which they respond on the basis of presence or absence of a single cue, disregarding all other cues; or (3) a singleton strategy, in which they learn only about the four patterns that have only one cue present and all others absent. This variability in how subjects approach this task may have important implications for interpreting how different brain regions are involved in probabilistic category learning.

Probabilistic category learning has been studied extensively in both animals and humans since the 1950s, and has proven to be a fertile domain for the development and testing of formal models of learning and memory (Medin and Schaeffer 1978, Nosofsky 1984, Estes 1986; Gluck and Bower 1988a,b). In the last several years, this paradigm from cognitive psychology has become popular within cognitive neuroscience as a method for studying the neural substrates for learning incrementally acquired cognitive skills. However, it appears that these tasks may actually be solvable by a range of different strategies. This study presents techniques to deduce what strategies a subject is likely to be using on the basis of post-hoc analyses of behavioral responding. These techniques may be useful for determining whether various groups—such as different clinical populations—are using qualitatively different strategies as compared with control subjects.

Many studies of the cognitive neuroscience of probabilistic category learning have used a paradigm known as the weather prediction task, developed in our lab at Rutgers University in the early 1990s as a variation of an earlier probabilistic category learning design from Gluck and Bower (1988a). As described in Knowlton et al. (1994) (Experiment 1, Task 2), subjects in the weather prediction task are given multidimensional stimuli and asked to classify them into one of two categories. These stimuli are composed from a set of four tarot cards (Fig. 1), each of which contains a unique geometric pattern.

The stimulus presented on each training trial consists of one or more of these cards presented in a random spatial order. Table 1 shows the 14 patterns that were used in the Knowlton et al. (1994) weather-prediction study. Each pattern is represented as a numeric four-digit sequence corresponding to whether each of the four cards is present (1) or absent (0). Thus, pattern A = 0001 has card 4 (squares) present, pattern B = 0010 has card 3 (diamonds) present, pattern C = 0011 has both card 3 and card 4 present, and so on. On each trial, subjects see one of these patterns, and are asked to predict whether there will be good or bad weather (sun or rain). The actual weather outcome is determined by a probabilistic rule based on the individual cards, whereby each card is a partially accurate predictor of the weather.

In the past several years, this weather-prediction task has been used in over half a dozen different studies. Amnesic patients have been reported to learn the weather-prediction task normally during the first 50 trials, although they are impaired relative to controls with further training (Knowlton et al. 1994, 1996a). Patients with basal ganglia dysfunction due to Parkinson's disease or Huntington's disease show impaired learning from the very start of training (Knowlton et al. 1996a,b). These results have been inter-

[3]All authors contributed equally to this work.
[4]Corresponding author.
E-MAIL gluck@pavlov.rutgers.edu; FAX (973) 353-1272.

Article and publication are at http://www.learnmem.org/cgi/doi/10.1101/lm.45202.

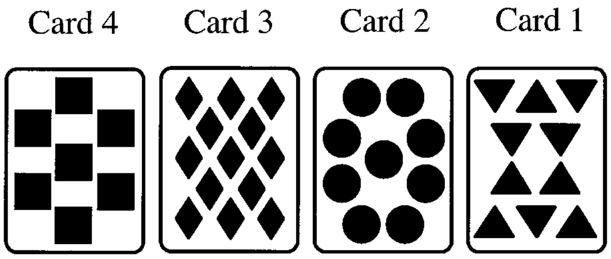

Figure 1 The four cards used as cues in Knowlton et al. (1994). Each card was associated with each possible outcome with a fixed probability.

preted as supporting the claim that there are separate memory systems for declarative memory (in the medial temporal lobes) and nondeclarative learning (in the basal ganglia). Other studies of the weather prediction task have demonstrated severe impairments in patients with schizophrenia (Keri et al. 2000; T. Weickert, T. Goldberg, A. Terrazas, L. Bigelow, J. Malley, M. Egan, and D. Weinberger, in prep.). Functional brain imaging by Poldrack and colleagues has shown that the medial temporal lobes become active early during learning of the weather-prediction task, and gradually become deactivated as the task is learned; conversely, the basal ganglia are inactive early in learning, but gradually become active (Poldrack et al. 1999, 2001). Computational modeling has illustrated how data on amnesic learning of this task could be explained by a cortico-hippocampal model of associative learning (Gluck et al. 1996).

Part of the appeal of the weather-prediction task has been the assumption that it is a cognitive skill that is learned in a procedural or habit-based (nondeclarative) manner. Another appeal is that the behavioral rules for category learning appear to involve some of the same principles as seen in classical conditioning, a canonical example of motor skill learning (Gluck and Bower 1988a,b, 1990; Shanks 1991). In particular, a powerful model of classical conditioning (Rescorla and Wagner 1972), which describes the incremental trial-by-trial changes that occur during classical conditioning of a CS and a US, also describes changes in associative strengths between cues and outcomes in probabilistic category learning tasks (Gluck and Bower 1988a). Although conditioning and category learning undoubtedly involve different neural substrates, it has been intriguing that they share so many similar behavioral properties.

However, despite considerable interest among cognitive neuroscientists in probabilistic category learning (and the weather-prediction task in particular), relatively little is known about how subjects actually approach this task. Prior studies have assumed that people all learn the task the same way. Because the four cue-outcome associations are probabilistic, it has been assumed that subjects learn these associations incrementally, much as if there were four independent conditioning processes going on in parallel, with subjects' choice behavior on each trial reflecting the accumulated associations among all the present cues. This was, in fact, how this learning was modeled in the Gluck and Bower (1988a) neural network model of probabilistic category learning, as well as in the later Gluck et al. (1996) associative network model.

As such, categorization results are typically analyzed with respect to optimal responding, that is, on each trial, given a particular configuration of cues, did the subject choose the outcome that is most often associated with that pattern over the course of the experiment?

Using such a strategy, subjects would be able to achieve 100% optimal responding. However, in the original weather-prediction study (Knowlton et al. 1994), healthy control subjects averaged only about 70%–75% optimal responding by the end of the experiment. Thus, it is possible that subjects were only imperfectly following such a strategy. However, it is important to realize that in addition, subjects could potentially use several other classes of strategies to approach the weather-prediction task. For example, in the version of the weather-prediction task used by Knowlton et al. (1994) and others, two of the four tarot cards are highly predictive of the weather, with each being associated with one or the other outcome with ~75% probability. The other two cards are less predictive (associated with one or the other outcome with ~57% probability). Thus, a subject who focuses attention on just one of the highly predictive cards, and then responds sun or rain based only on the presence or absence of this one card, could achieve 75% optimal responses, similar to the level of 75% optimal responding that most subjects in the Knowlton et al. (1994) study actually achieved. Thus, such a strategy could conceivably account for the behavior of subjects in the Knowlton et al. (1994) experiment.

In summary, if one only knows the aggregate percent optimal responses from a subject, it is difficult to conclude anything about how that subject learned the task. Thus,

Table 1. *Stimulus Patterns Used in the Knowlton et al. (1994, Experiment 1) Weather Prediction Task*

Pattern	Cards present
A	0001
B	0010
C	0011
D	0100
E	0101
F	0110
G	0111
H	1000
I	1001
J	1010
K	1011
L	1100
M	1101
N	1110

although the amnesic and control groups in the Knowlton et al. (1994) study showed similar percent optimal responding, it is difficult to know whether the two groups were actually using the same strategies or whether qualitatively different strategies might underlie learning in the two groups.

This study presents the development and application of techniques by which we may deduce what strategies a subject is likely to be using on the basis of post-hoc analyses of behavioral responding. These techniques are applied to subjects' performance on two versions of the weather-prediction task, an exact replication of the Knowlton et al. (1994) experiment and a newer version of the task, in which the cue-outcome probabilities are slightly more discriminable.

Experiment 1

The purpose of Experiment 1 was to evaluate how subjects approach learning in the weather-prediction probabilistic category learning task. Our first experiment is essentially a replication of the weather-prediction task published originally in Knowlton et al. (1994) (Experiment 1, Task 2). Here, in addition, we used questionnaires (Experiment 1A) and mathematical models (Experiment 1B) to assess the strategies subjects used to learn the task. As described above, there are four tarot cards that can each be present (1) or absent (0) on a given trial. Each tarot card is associated with one of two outcomes (sun vs. rain) with a fixed probability, as shown in Table 2. The overall probablility of each outcome on a given trial is calculated according to the conditional probabilities of each outcome and card occurring together (Table 3). In addition to recording overall percent optimal responding, we recorded how an individual subject responded to each pattern, and to each cue. Following training, subjects responded to a questionnaire that was designed to provide insight to the kinds of strategies each subject used while performing the task. Mathematical models were developed on the basis of information from the self-reports and the questionnaire, to provide a more objective and accurate strategy analysis.

Experiment 1A: Results

Behavior

Over all 200 training trials, subjects achieved a mean 62.41% optimal responses (SD 7.36). Performance is shown in Figure 2A across the four blocks of training; subjects started near chance (50%) in block 1, and improved to >70% optimal responding by block 4. A repeated-measures analysis of variance (ANOVA) confirmed a significant within-subjects effect of block [F(3,84) = 25.68, $P < .001$], with no effect of subject gender [F(1,28) = 1.25, $P = .273$] and no block-gender interaction [F(3,84) = 1.86, $P = .142$).

Table 2. *Cue Probabilities*

	P (Sun|cue present)	P (Rain|cue present)
Cue 4 (squares)	.756	.244
Cue 3 (diamonds)	.575	.425
Cue 2 (circles)	.425	.575
Cue 1 (triangles)	.244	.756

Table 3. *Probability Structure of the Task*

Pattern	Cue 1	Cue 2	Cue 3	Cue 4	P (pattern)	P (rain| pattern)
A	0	0	0	1	0.14	0.143
B	0	0	1	0	0.08	0.375
C	0	0	1	1	0.09	0.111
D	0	1	0	0	0.08	0.625
E	0	1	0	1	0.06	0.167
F	0	1	1	0	0.06	0.5
G	0	1	1	1	0.04	0.25
H	1	0	0	0	0.14	0.857
I	1	0	0	1	0.06	0.5
J	1	0	1	0	0.06	0.833
K	1	0	1	1	0.03	0.333
L	1	1	0	0	0.09	0.889
M	1	1	0	1	0.03	0.667
N	1	1	1	0	0.04	0.75

For each pattern, each card could be present (1) or absent (0). The all-present (1111) and all-absent (0000) patterns were never used. The overall probability of rain, given by summing P (Pattern) *P (rain|pattern) for all patterns, is 50%.

Figure 2B shows performance across just the first 50 trials. A repeated-measures ANOVA confirmed a significant within-subjects effect of trials [F(4,112) = 3.49, $P = .010$), no effect of subject gender [F(1,28) = 1.23, $P = .278$] and no trials-gender interaction [F(4,112) = 0.94, $P = .442$]. However, as seen in Figure 2B, this effect of trials was due to somewhat better-than-chance responding in the first 10 trials; performance actually fell back to near chance levels for trials 11–50.

Setting an arbitrary criterion of 65% optimal responding throughout training, 10 of 30 subjects reached criterion performance within 200 trials.

Questionnaire

Within the questionnaire, subjects were asked about the strategy they had used, the cue-outcome probabilities, and the general task structure.

Open-Ended Question

When given an open-ended question asking the subject to describe the strategy used to predict the weather, responses varied widely. Several subjects simply stated that they had tried to "associate the cards with sunny or rainy weather" or similar. However, many of the more detailed responses fell into a few basic categories.

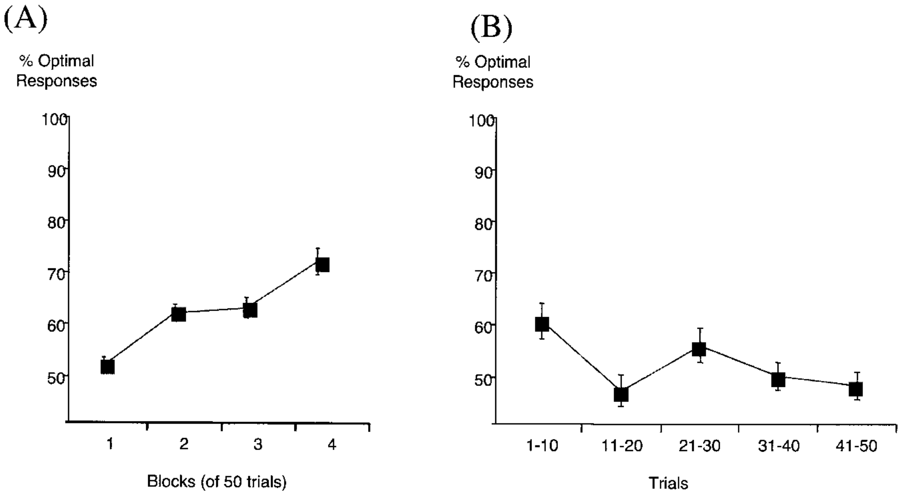

Figure 2 Percent optimal responses (*A*) over all 200 trials, (*B*) over first 50 trials of training.

(1) One-Cue Learning. Basing responses on the presence or absence of a single card (e.g., "I predicted rain whenever I saw the triangle card.").
(2) Multi-Cue Learning. Basing responses on the combinations of cues present on a given trial (e.g., "I noticed that triangles and diamonds usually meant rain, and the circles and squares meant sunny."[5]).
(3) Singleton Learning. Learning the correct response to singleton patterns (A = 0001, B = 0010, C = 0100, D = 1000), in which only a single card appears, and guessing on the remaining trials (e.g., Memorizing the single cards, "The single cards were the easiest, so I concentrated on those.").

Additionally, a few subjects reported that they were learning the correct response to singleton patterns, and then adding evidence from singletons together when more than one card appeared (e.g., "From the result of one-card incidences [sic] I would learn that a pattern predicted a certain outcome. . . . If there were two patterns that had caused sun, I would pick that (and vice versa). If there was a card that had predicted sun and one that had predicted rain, I'd guess").

The remaining subjects reported that they were guessing, memorizing what weather went with each combination of cards, or using a sequence (e.g., respond sun if the last trial's weather was rain and vice versa).

A total of 22 of the 30 subjects reported that they thought the strategy they had used was good or acceptable. There was no obvious relationship between this response and behavioral performance; subjects reporting satisfaction averaged 62.8% optimal responses (SD 7.2), whereas subjects reporting they did not think they had used a good strategy averaged 61.4% optimal responses (SD 8.3). This was not a significant difference (independent-samples *t*-test, $t(11) = 0.40$, $P > .500$).

Multiple-Choice Question

Following the open-ended question, subjects were given a multiple-choice question in which they were asked which best described the strategy they had used to predict the weather as follows: (a) guessing; (b) noticing that certain cards (which?) most often predicted sun and certain cards (which?) most often predicted rain; (c) memorizing which combinations of cards predicted which kinds of weather.

A total of 22 subjects checked (b); 18 subjects checked (c); 10 subjects checked both (b) and (c). Among those subjects checking (b), only 6 subjects listed all 4 cards. Several subjects also checked (a).

There was no obvious correlation between answers on the open-ended question and the multiple-choice question. For example, one of the subjects who correctly reported all four cards and contingencies in the multiple-choice question had claimed previously to be responding on the basis of picture association, reading cards from left to right. Conversely, another subject who claimed to have learned the four singleton patterns was only able to verbalize (on the multiple-choice) that the square card often predicted sun.

Cue-Outcome Contingencies

Next, we investigated subjects' estimates of association strengths for each of the individual cues [similar to a procedure reported in Reber et al. (1996)]. Subjects were asked four questions of the form, "If just the square [or circle, diamond, triangle] card is showing, what percentage of the time would it be sunshine? (Respond with a number from 0–100.)".

For every cue, subject responses varied widely; for example, for the square card (highly predictive of sun), estimates ranged from 20%–100%. As Figure 3 shows, although the group mean of 74.2% for the square card is close to the actual rate of 85.7% sun (see Table 3), the group mean for the other highly predictive triangle card approached chance (51.5%). No individual subject correctly estimated the probability for all four cues within a range of +/−15% of the actual probability. (In fact, only one subject even came close to this criterion, estimating the four probabilities as 85%, 40%, 30%, and 25%).

Regression analyses of the cue estimates, with total

[5] Note that this particular subject's mapping of cues to outcomes is partially incorrect; as per Table 2, squares and diamonds were most often associated with sun, and circles and triangles with rain.

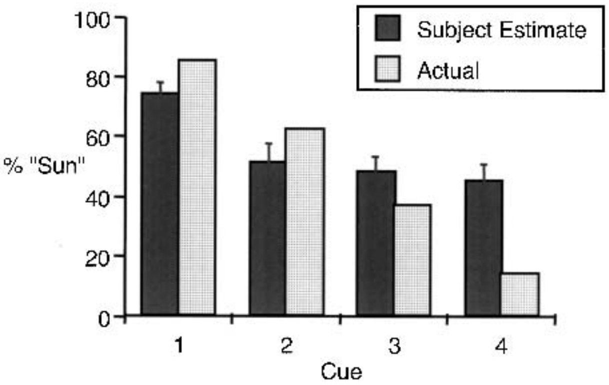

Figure 3 Actual and estimated cue probabilities.

percent optimal responding as the dependent measure, revealed a significant relationship between estimate of P(sun|square card) (ANOVA, F(1,28) = 11.24, P = .002), but not for the other highly predictive triangle card (F(1,28) = 0.66, P = .422), nor either of the less-predictive cards (circle, F(1,28) = 0.33, P > .500; diamond, F(1,28) = 2.71, P = .111).

Subjects were also asked the opposite questions, "What if you knew it was going to be sunny [or rainy] and one card was showing? Which card would be most likely to be showing?" Again, individual responses varied widely, with at least one subject reporting that each possible card was most often present during sun or rain trials.

Overall, 14 subjects (46.7%) reported correctly that the square card was most often present on sun trials, whereas 14 subjects (46.7%) reported correctly that the triangle card was most often present on rain trials. Only six subjects (20%) answered both questions correctly. These six subjects averaged 68.8% optimal responding (SD 7.1), only slightly better than the overall group average of 62.4%.

Questionnaire: Summary

In summary, the results of the questionnaire were highly conflicting. Group averages did not approximate ideal responding, and individual subject responses varied widely, with one subject often reporting different strategies when the question was posed in various ways. There was no obvious correlation between subject strategy and performance, and little good evidence that subjects had learned the cue-outcome contingencies, although they did tend to associate the square card with sun and triangle with rain, when probed to estimate cue-outcome associations. This last finding is generally consistent with the findings of Reber et al. (1996), who found that when probed, subjects, on average, tend to provide relatively accurate estimates of cue-outcome associations.

Experiment 1B: Strategy Analyses

The results from the subject questionnaires suggested that subjects were using more than one strategy to approach the probabilistic categorization task. Some reported using one-cue strategies (responding based on the presence or absence of a single card), or using singleton strategies (learning the optimal response to the four singleton patterns). In fact, if a subject were reliably following one of these strategies, the overall performance rate could approach that obtained by the optimal multi-cue strategy (see below). This suggests that simply knowing a subject's overall performance rate does not necessarily provide any information about which strategy that subject was pursuing.

However, the wide variation within individual subjects among answers to the open-ended question, the multiple-choice question, and the probability estimates suggest that subjects found it difficult to verbalize their strategies. This would be consistent with the assumption that this task taps nondeclarative or implicit memory, and suggests that questionnaire data may be inherently unreliable as an index of how subjects approach the probabilistic category learning task.

We next attempted to derive a more formal method of assessing subject strategies, by quantitatively comparing individual subject data to the ideal subject data that would be expected if a subject were reliably following a particular strategy. On the basis of responses obtained to the questionnaire, we investigated one-cue and singleton strategies, as well as multi-cue strategies, which are often assumed to be how subjects approach this type of task (for example, see Knowlton et al. 1994). Table 4 summarizes these four strategies.

RESULTS

Over all 200 training trials, we found that 27 subjects (90%) were best fit by a singleton strategy, and 3 subjects were best fit by a one-cue strategy (2 circles, 1 triangle).[6]

Our list of potential strategies is clearly not exhaustive, and our strategy analysis clearly cannot prove that an individual subject was following a particular strategy. However, defining a tolerance level of 0.1, all subjects were fit by one of the three classes of strategies we considered. These results do demonstrate that most subjects were behaving in a manner more consistent with a singleton strategy than with a multi-cue strategy.

We also conducted separate strategy analyses for the individual subject data over each block of 50 trials. Figure 4

[6] Within the singleton group, one subject was better fit by assuming that she had summed evidence from multiple singletons on those trials in which two or more cards appeared; whereas there were not enough subjects well fit by this model to justify treating it as a separate group, it is worth noting that this is a potentially more sophisticated strategy than simply learning the singletons and guessing on multi-card patterns. However, on the questionnaire, this subject reported that she had been "memoriz[ing] the pattern of the cards and the sequence of how it appears."

Table 4. *The Rule for Constructing Ideal Subject Data for Each Strategy Investigated*

Strategy	Ideal subject data constructed by:	% Optimal
Multi-Cue	Assume the optimal response (i.e., most frequently-correct outcome) is made on every trial.	100%
Singleton	Assume the subject learns the optimal response for each of the four singleton patterns and guesses on the remaining trials.	75%
One-Cue (highly predictive)	Assume the subject is responding based on the presence or absence of one of the highly-predictive cues (square or triangle) and ignoring all other cards.	87.5%
One-Cue (less predictive)	Assume the subject is responding based on the presence or absence of one of the less-predictive cues (diamond or circle) and ignoring all other cards.	66%

(% Optimal) = potential performance by a subject reliably following this strategy through the entire experiment.

shows the number of subjects best fit by each strategy across the four blocks. There is a shift in best-fit strategies, such that many subjects are best-fit by a singleton strategy early in training, but gradually come to behave in a manner more consistent with a (more effective) multi-cue strategy.

Assuming that block 4 is the best indicator of a subject's ultimate strategy, and most likely to have been reliably followed at least at the end of the experiment, we attempted to correlate best-fit strategy for block 4 with performance. Figure 5 shows that, in general, subjects who were best fit by a multi-cue strategy model tended to generate a relatively high proportion of optimal responses, whereas subjects best fit by a one-cue model tended to generate a low proportion of optimal responses. An ANOVA on block 4 performance with best-fit strategy as the dependent variable revealed a highly significant main effect [F(2,27) = 8.105, $P < .001$). Tukey pairwise tests revealed that subjects best fit by the multi-cue strategy performed significantly better than those subjects best fit by a one-cue strategy ($P < 0.001$). No other pairwise contrasts reached significance (all $P > .100$), that is, there was no significant difference in performance between subjects best fit by the singleton or the one-cue models, or between subjects fit by the singleton or multi-cue models.

Finally, comparing questionnaire responses with best-fit strategy yielded no obvious correspondences. For example, two subjects who had spontaneously verbalized a one-cue strategy were best fit by a multi-cue strategy. Several subjects whose actual behavior was most consistent with a one-cue strategy verbalized that they had been responding on the basis of information from several cues. Thus, again, there was little evidence that questionnaire data accurately represented how subjects had approached the task.

DISCUSSION

In Experiment 1, we used questionnaires and mathematical models to assess strategies used during training on the weather-prediction category learning task. Findings from the questionnaire suggested that there are various strategies with which subjects approach this task. Three main classes of strategies were identified as follows: multi-cue strategies, in which subjects focus on and learn about all four cues; singleton strategies, in which subjects learn primarily about those patterns in which a single card appears; and one-cue strategies, in which subjects respond on the basis of presence or absence of a single cue.

We developed mathematical models on the basis of these strategies, to obtain a more objective and accurate assessment of the kinds of strategies subjects used while learning this task. Strategy analyses with these models confirmed that subjects use a variety of strategies while learning this task. Specifically, we found that 90% of subjects were best fit by a singleton strategy, over all 200 trials. By investigating strategies within 50-trial blocks, we found evidence of a shift; in the first training block, most subjects appeared to start with a singleton strategy; by the last block, subjects appeared to shift away from a singleton strategy and toward a more optimal multi-cue strategy. This shift toward multi-

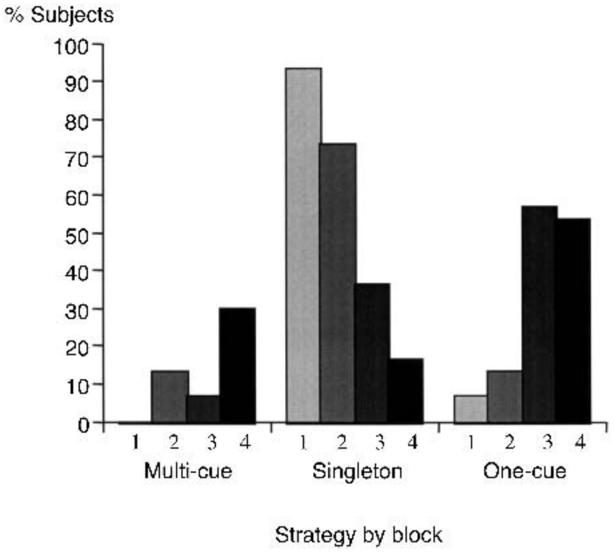

Figure 4 Percentage of subjects best fit by various strategy models across the four training blocks.

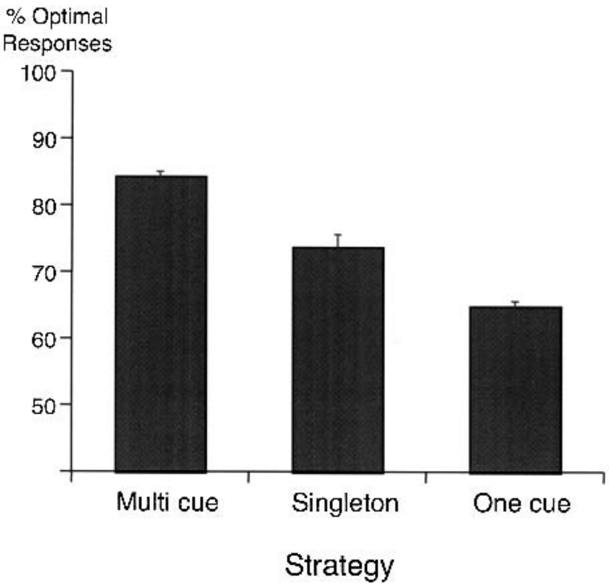

Figure 5 Percent optimal responding as a function of best-fit model.

cue strategy was also reflected in better performance on the final training block (Fig. 5).

We also found little correspondence between subjects' self reports and their actual performance, as assessed by mathematical analyses. This is interesting, given that self reports are often used to gain insight into how subjects learn. This finding is not really surprising, given that the weather prediction task has always been assumed to be learned in an implicit manner by use of procedural, nonverbalizable rules. This finding further suggests that even when subjects are using a strategy that could, potentially, be verbalized (such as "press sun whenever you see a triangle"), they may be learning the strategy in a nondeclarative manner.

One important limitation to these findings is that overall, subjects were not performing so well. Two-thirds of subjects were performing at <65% optimal over the 200 trials. Therefore, it is difficult to determine whether they were following a specific strategy, changing strategies rapidly in an attempt to improve performance, or perhaps simply guessing. In an attempt to address this, in Experiment 2, we altered the weather prediction task slightly by making the specific cue-outcome probabilities slightly more discriminable, while maintaining the same formal structure of the task. We expected that this would have the effect of making the task somewhat easier to learn (and hence, of improving performance rates) while maintaining the general probabilistic nature of the task.

Experiment 2

In Experiment 1, we attempted to assess the kinds of strategies subjects use to learn the weather-prediction task. We presented a new technique for examining learning strategies, and identified three main classes of strategies used. However, one limitation of these findings is that the task is quite difficult. The purpose of Experiment 2 was to develop a task that would be formally identical to the Knowlton et al (1994) task in every way, but that nonetheless would be slightly easier to learn. Obtaining slightly higher levels of performance is particularly important, given that this task is often used to assess learning patients with various kinds of memory dysfunction, and poor performance among healthy controls could make the task less sensitive to impairments in patient populations.

In the present experiment, we slightly modified the cue probabilities to make the task easier to master. Specifically, the four cues were associated with sun with probabilities of 0.8, 0.6, 0.4, and 0.2, which meant that an individual who always responded with the most likely category for each pattern could correctly predict the weather on up to 83% of trials (compared with 76% under the probabilities in Experiment 1). The pattern frequencies and conditional probabilities for Experiment 2 are shown in Table 5.

We expected that subjects would show overall better performance, but that we would find similar patterns of learning as in the original task (i.e., incremental acquisition) and similar patterns of strategy use as revealed by mathematical models.

RESULTS

Overall Performance Measures

Over all 200 trials, subjects averaged 74.5% optimal responses (SD 11.3). Figure 6 shows these data broken into blocks of 50 trials. A repeated-measures ANOVA confirmed significant improvement across trials (within-subjects effect

Table 5. *Total Frequency of Occurrence of Each Pattern, Along With Number of Times Each Pattern Occurred With Sun or Rain Outcome in Experiment 2*

Pattern	Cards present	Sun	Rain	Total
A	0001	17	2	19
B	0010	7	2	9
C	0011	24	2	26
D	0100	2	7	9
E	0101	10	2	12
F	0110	3	3	6
G	0111	17	2	19
H	1000	2	17	19
I	1001	3	3	6
J	1010	2	10	12
K	1011	5	4	9
L	1100	2	24	26
M	1101	4	5	9
N	1110	2	17	19
Total		100	100	200

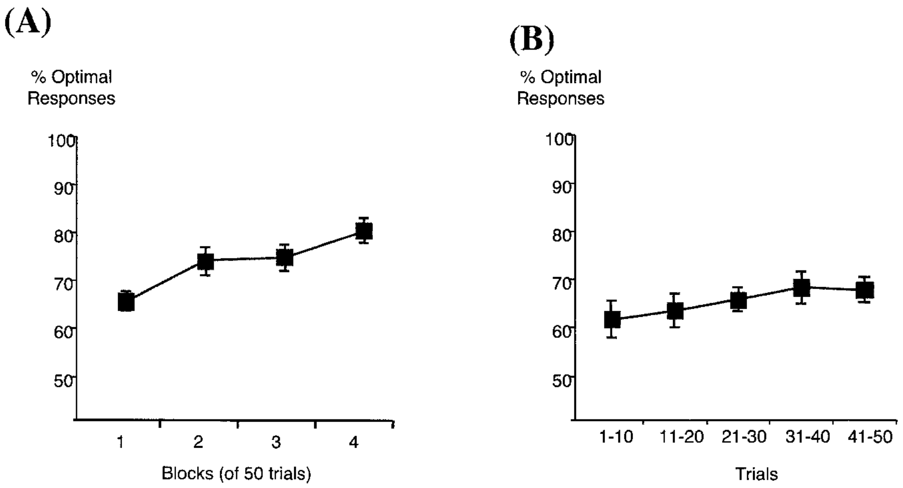

Figure 6 Percent optimal responses (*A*) over all 200 trials, (*B*) over first 50 trials of training.

of block [F(3,84) = 12.84, $P < .001$], with no effect of subject gender [F(1,28) = 0.07, $P > 0.500$], and no gender-block interaction [F(3,84) = 0.96, $P = 0.416$).

We also considered just the first 50 trials, divided into sub-blocks of 10 trials. Figure 6B shows these data. In fact, there was a slight increase in percent optimal responding across the first 50 trials; however, a repeated-measures ANOVA showed that this was not statistically significant (within-subjects effect of trials [F(4,112) = 0.88, $P = .477$), no effect of subject gender [F(1,28) = 0.75, $P = .393$), and no interaction [F(4,112) = 1.02, $P = .400$).

Setting an arbitrary criterion of 65% optimal responding across all 200 trials, 24 subjects (80.0%) reached criterion performance. This is compared with 33% of subjects reaching criterion in Experiment 1.

Strategy Analyses

Over all 200 training trials, we found that a singleton model provided best fit for 24 subjects (80%), whereas a multi-cue model provided a best fit for 4 subjects (13.3%), and one-cue models provided fits for 2 subjects (6.7%). Defining a tolerance level of 0.1, all subjects but one were fit well by one of the three classes of strategies we considered.

We also conducted separate strategy analyses for the individual subject data over each block of 50 trials. Figure 7 shows the percent of subjects best fit by each strategy across the four blocks. As in Experiment 1, we found a shift in best-fit strategies, with many subjects best fit by a singleton strategy early in training, but behaving in a manner more consistent with a multi-cue strategy later in training.

Finally, we evaluated whether best-fit strategies in the last block were related to overall performance. Figure 8 shows that subjects who were best fit by a multi-cue strategy model tended to generate a relatively high proportion of optimal responses, whereas subjects best fit by a one-cue or singleton model tended to generate a relatively low proportion of optimal responses. An ANOVA on block 4 performance with best-fit model as the independent variable confirmed a significant effect of strategy on performance [F(2,27) = 6.04, $P < .010$); post-hoc Tukey tests revealed that the multi-cue group significantly outperformed the singleton group ($P < 0.001$) and the one-cue group ($P < 0.050$); the singleton group did not differ significantly from the one-cue group ($P > .500$).

DISCUSSION

In Experiment 2, we used a new version of the weather-prediction task. This version of the task was formally identical to the task in Experiment 1 in all respects, except that the actual probabilities of cue-outcome associations were slightly more discriminable. Whereas in Experiment 1, cues 1–4 were associated with sun with probabilities of .25, .43, .57, and .75, respectively, in Experiment 2 the corresponding probabilities were .2, .4, .6, and .8. These new probabilities resulted in overall slightly better performance; subjects averaged 74.4% optimal correct performance over all 200 trials in Experiment 2, compared with 62.1% correct performance in Experiment 1. Furthermore, whereas in Experiment 1, relatively few subjects reached a criterion of at least 65% correct responses over all 200 training trials, in Experiment 2, substantially more subjects were able to reach criterion performance (33% of subjects in Experiment 1, compared with 80% in Experiment 2). However, despite slightly better performance, overall patterns of learning were similar—here, too, learning was incremental across all 200 trials, and we found a similar pattern of strategy use. These findings are consistent with fMRI findings of similar patterns of activation with both versions (Poldrack et al. 1999, 2001). This version of the weather-prediction task, therefore, appears not to differ from the initial version in any fundamental way, at least in healthy young controls.

General Discussion

The main point of our results is to demonstrate that a task widely used in the cognitive neuroscience literature may not be solved by subjects in the ways that researchers thought previously. In particular, there appears to be considerable variability in how subjects approach this task.

Using mathematical models of three classes of strategies, the multi-cue, one-cue, and singleton strategies, we found that data from all subjects were consistent with one of these strategies, and, in fact, 59 of 60 subjects across two

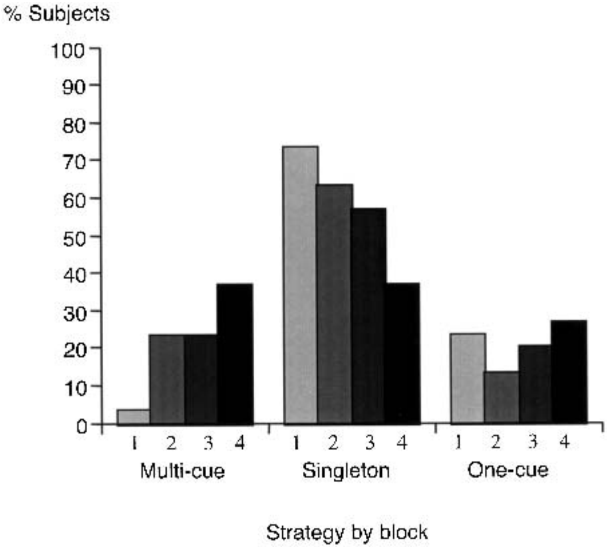

Figure 7 The percentage of subjects for whom multi-cue, one-cue, and singleton strategies provided a best-fit model for each of the four training blocks.

experiments were fit within a tolerance level of 0.1. In both the original and the modified version of the weather-prediction task, the singleton strategy appeared to account for most subjects' data, particularly early in training. There was also evidence of a shift, with many subjects showing behavior more consistent with an optimal multi-cue strategy toward the last block of training.

Because both the one-cue and singleton strategies are simple and easy to verbalize, there is no particular reason to expect that subjects would use nondeclarative memory to mediate learning on the basis of these strategies. However, our finding in Experiment 1 that subjects' actual performance (assessed with models) was poorly correlated with their self reports, supports the idea that even these verbalizable rules are acquired in an unconscious, nonverbalizable way.

Obviously, one cannot definitely assume a given subject was using a particular strategy just because that strategy provided a best-fit model; at least one subject in Experiment 2 was not well described by any of the mathematical models we explored, suggesting that he may have been using a different strategy to approach the task. In future work, it may be interesting to apply clustering techniques, such as multidimensional scaling, to response profile data, to see if we can identify additional or alternate strategy techniques.

Because the weather prediction task is probabilistic and incrementally acquired, it has been assumed in the past that it therefore must depend on nondeclarative (extra-hippocampal) memory systems. In reviewing our results, we feel that it is important to note that just because a strategy can be learned incrementally, it does not necessarily follow that it must be nondeclarative, nor is it the case that just because a strategy can be easily memorized, it must be declarative.

For example, the one-cue strategy can obviously be learned in a nondeclarative, incremental fashion analogous to conditioning to a single cue. However, one can also easily verbalize this strategy using declarative memory: 'Respond "rain" whenever the triangle card is present, and respond "sun" whenever it is not present.' Conversely, just because the singleton strategy is easy to verbalize, does not mean that a person with impaired declarative memory, such as an amnesic patient, should be unable to learn the correct response to each of four stimuli that are presented with very high frequency compared with other stimuli in the task. Thus, we argue that one should be cautious about strong claims regarding which strategies are declarative versus nondeclarative. In fact, it seems plausible to suggest that a normal healthy human may use a variety of strategies and brain substrates—possibly in parallel—to approach a difficult categorization task. Patients with damage to one or more brain systems may be restricted in the strategies they can use to approach this task—but to the extent that they can still use at least one effective strategy, they may be able to perform quite well in terms of overall percent optimal responding. In such a case, a gross measure of overall optimal responding may not differentiate a patient group from healthy controls, but a strategy analysis might. Future studies with various patient groups are obviously indicated to determine the extent to which strategy analysis on probabilistic category learning tasks may help elucidate not just how well patients learn, but what they are learning.

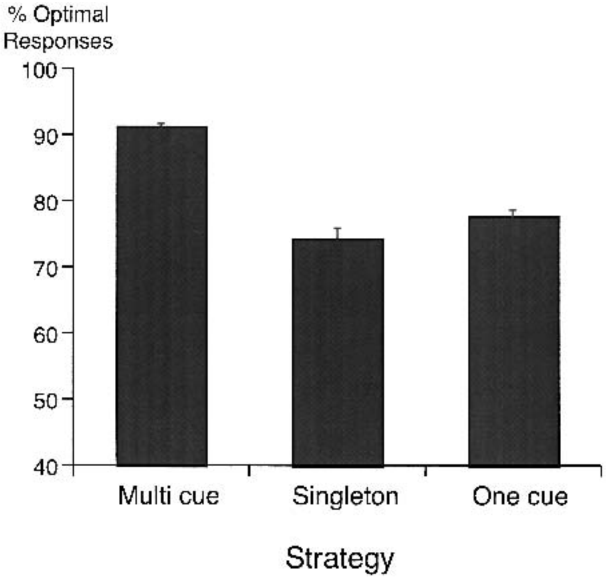

Figure 8 Percent optimal responding as a function of best-fit strategy.

METHODS

Experiment 1A

Subjects

A total of 30 Rutgers University undergraduate students participated, and received class credit in exchange for their participation. The group included 13 males and 17 females, with a mean age of 19.8 years (SD 2.9).

Apparatus

Testing took place in a quiet room with the subject seated at a comfortable viewing distance from a Macintosh iBook laptop computer. The keyboard was masked except for two keys, labeled sun and rain, which the subject could use to enter responses.

Stimuli and Procedure

The general procedure was as described previously in Knowlton et al. (1994) (Experiment 1, Task 2), using the same software, cues, and cue probabilities as in the earlier work.

In brief, the subject was required to learn which of two outcomes (rain or sun) was predicted by combinations of tarot cards that appeared on the screen. There were four cards total, as shown in Figure 1, each associated with each outcome according to a fixed probability (see Table 2). One to three cards could appear on each trial. The actual outcome on each trial was calculated according to the conditional probabilities of each outcome and card occurring together (see Table 3). There were 200 trials total, and the sun and rain outcomes occurred with equal frequency.

On each trial, the cards appeared and the subject was asked to respond with a prediction of sun or rain. Once the subject responded, the correct answer was shown. If the response was correct, a smiley face appeared, a high-pitched tone sounded, and a score bar on the right of the screen increased; if the response was incorrect, a frowning face appeared, a low tone was sounded, and the score bar decreased. Visual feedback and cards remained on the screen for 2 sec, followed by a 1-sec intertrial interval during which the screen was blank.

If the subject did not respond within 2 sec, a prompt appeared, Answer Now!. If the subject did not respond within the next 3 sec, the trial was terminated and the correct answer was shown.

Following the 200 training trials, subjects were given a questionnaire, including both open-ended and multiple-choice questions about the strategies they had used to approach the task, the individual cue-outcome probabilities, and the overall structure of the task.

For the purposes of data analysis, a subject was considered to have made an optimal response if the subject selected the outcome that was most often associated with the current cue pattern, regardless of the actual (probabilistically determined) weather on that trial. Thus, subjects could be scored as making an optimal response on a given trial even if their actual prediction of the weather was wrong. Two patterns (F = 0110 and I = 1001; see Table 3) appeared equally often with each outcome, and so the optimal response for trials on which pattern F or I appeared was undefined. Percent optimal scores were analyzed in blocks of 50 trials. Additionally, following previous studies, we analyzed data from the first 50 trials in sub-blocks of 10 trials.

Experiment 1B

Ideal response profiles were constructed for each strategy, defined as the expected response to each trial in the experiment if a subject were perfectly following that strategy (see Table 4). This ideal profile was compared with individual subject data by means of a least-mean-squared-error measure (see Appendix). The strategy generating the lowest error was defined to be the best-fit model for that subject's data. Note that although the scoring for the multi-cue strategy allows for 100% optimal responding, a subject reliably following one of the other strategies could obtain up to 88% optimal responding.

We conducted this strategy analysis for all 30 subjects over all 200 training trials. The actual probabilities of the experiment only hold true across all 200 training trials. However, on the basis of subject questionnaire responses, it appeared likely that at least some subjects had switched strategies during the course of the experiment. Accordingly, we also conducted separate strategy analyses for the individual subject data over each block of 50 trials.

Experiment 2

Subjects

A total of 30 subjects were recruited from the Rutgers University community, including 17 female and 13 male subjects, with a mean age of 20.73 years (SD 3.64); participants either volunteered or received class credit for an introductory psychology course. All subjects signed statements of informed consent before initiation of behavioral testing.

Stimuli and Procedure

Stimuli were the same as in Experiment 1 and in prior studies (e.g., Knowlton et al. 1994; Experiment 1, Task 2). The two outcomes (sun and rain) were equally probable, but each of the four cards was independently associated with each outcome with a fixed probability: P(sun|card 4) = 0.8; P(sun|card 3) = 0.6; P(sun|card 2) = 0.4; P(sun|card 1) = 0.2. The associations between cards 1–4 and rain were, accordingly, P(rain|card) = 1-P(sun|card) or 0.2, 0.4, 0.6, and 0.8, respectively. Trials were then constructed to adhere to these independent probabilities. Table 5 shows the number of times each combination of cards (i.e., each pattern) occurred with each outcome. The 200 trials defined in Table 5 were presented in a random, but fixed order for all subjects.

The apparatus, procedure, and data collection were identical to Experiment 1.

ACKNOWLEDGMENTS

We thank Nessrein Abu-Shaba, Suditi Gupta, Natalie Montealegro, and Souty Shafik for assistance with data collection. We also thank the J.S. McDonnell Foundation for their generous support of this work.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Estes, W.K. 1986. Array models for category learning. *Cog. Psychol.* **18:** 500–549.

Gluck, M. and Bower, G. 1990. Component and pattern information in adaptive networks. *J. Exper. Psychol.: Gen.* **119:** 105–109.

Gluck, M.A. and Bower, G.H. 1988a. From conditioning to category learning: An adaptive network model. *J. Exper. Psychol.: Gen.* **117:** 225–244.

———. 1988b. Evaluating an adaptive network model of human learning. *J. Mem. Lang.* **27:** 166–195.

Gluck, M.A., Oliver, L.M., and Myers, C.E. 1996. Late-training amnesic deficits in probabilistic category learning: A neurocomputational analysis. *Learn. Mem.* **3:** 326–240.

Keri, S., Kelemen, O., Szekeres, G., Bagoczky, N., Erdelyi, R., Antal, A., Benedek, G., and Janka, Z. 2000. Schizophrenics know more than they can tell: Probabilistic classification learning in schizophrenia. *Psychol. Med.* **30:** 149–155.

Knowlton, B., Squire, L., and Gluck, M. 1994. Probabilistic classification learning in amnesia. *Learn. Mem.* **1:** 106–120.

Knowlton, B., Mangels, J., and Squire, L. 1996a. A neostriatal habit learning system in humans. *Science* **273:** 1399–1402.

Knowlton, B., Squire, L., Paulsen, J., Swerdlow, N., Swenson, M., and Butters, N. 1996b. Dissociations within nondeclarative memory in Huntington's disease. *Neuropsychology* **10:** 538–548.

Medin, D.L. and Schaeffer, M.M. 1978. A context theory of classification learning. *Psychol. Rev.* **85:** 207–238.

Nosofsky, R. 1984. Choice, similarity, and the context theory of classification. *J. Exper. Psychol.: Learn. Mem. Cog.* **10:** 104–114.

Poldrack, R., Prabakharan, V., Seger, C., and Gabrieli, J. 1999. Striatal activation during cognitive skill learning. *Neuropsychology* **13:** 564-574.

Poldrack, R.A., Clark, J., Pare-Blagoev, J., Shohamy, D., Creso-Moyano, J., Myers, C.E., and Gluck, M.A. 2001. Interactive memory systems in the human brain. *Nature* **414:** 546–550.

Reber, P.J., Knowlton, B.J., and Squire, L.A. 1996. Dissociable properties of memory systems: Differences in the flexibility of declarative and nondeclarative knowledge. *Behav. Neurosci.* **110:** 861–871.

Rescorla, R.A. and Wagner, A.R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In *Classical conditioning: II. Current research and theory.* (ed. A.H. Black and W.F. Prokasy), Appleton-Century-Crofts, New York.

Shanks, D.R. 1991. Categorization by a connectionist network. *J. Exper. Psychol.: Learn. Mem. Cog.* **17:** 1–11.

Received November 7, 2001; accepted in revised form September 24, 2002.

APPENDIX

Strategy Analysis

We investigated three basic classes of strategy that subjects might be using as follows: multi-cue strategies, in which the subject attends to all four cues, one-cue strategies, in which the subject attends selectively to one particular cue (e.g., the presence or absence of a particular card), and singleton strategies, in which the subject learns how patterns containing individual cues (singletons) predict the outcome.

For each strategy, we constructed ideal data, defined as the pattern of responses expected across the 200 trials if a subject were reliably following that strategy. For example, ideal data for a one-cue model based on the square card would assume that the subject responded sun to all patterns in which the square card was present (A,C,E,G,I,K,M), and rain to all patterns in which the square card was absent (B,D,F,H,J,L,N). Ideal data for a singleton model would involve responding sun to singletons A and B, rain to singletons D and H, and randomly to the remaining patterns. (We also initially considered a rule in which subjects responded to multi-card singletons by summing evidence across singletons—i.e., a majority rule, if two of the three cards are associated with sun, then respond sun—but as only one subject was best fit by this model, we do not consider it further.) Ideal data for a multi-cue rule would involve responding to each pattern on the basis of the most frequent outcome, for example, pattern A appears 19 times, 17 sun and 2 rain, so ideal data would involve responding sun each time A appears. These ideal data thus provided various models of subject performance that could be compared against actual subject response patterns.

In practice, no subject's response profile was perfectly identical to any of these ideal profiles. In the early trials, before any learning had taken place, a subject's responses would be expected to be random (or nearly so); later in the experiment a subject might switch strategies—or even make occasional errors in key pressing. Nonetheless, averaged over all 200 trials of the experiment and over the 4 blocks of 50 trials, many subjects' response profiles were fit quite well by at least one of the ideal data models.

To quantify this fit, we took the squared difference between the number of sun responses generated by a subject and the number predicted by a model, summed across all patterns; this score was normalized by dividing between the sum of squares of total presentations of each pattern,

$$\mathit{Score\ for\ Model\ M} = \frac{\sum_{P}(\#\mathit{sun\_expected}_{P,M} - \#\mathit{sun\_actual}_{P})^2}{\sum_{P}(\#\mathit{presentations}_{P})^2}$$

in which P = pattern A…N; $\#\mathit{presentations}_P$ is the number of times pattern P appears in the 200 trials of the experiment; $\#\mathit{sun\_expected}_{P,M}$ is the number of sun responses expected to pattern P under model M, and $\#\mathit{sun\_actual}_P$ is the actual number of sun responses the subject made to pattern P during the experiment.

The result was a number between 0 and 1 for each strategy, with 0 indicating a perfect fit between the model M and a subject's response profile. Comparing across all strategies examined, the model generating the lowest score was defined as the best-fit model for that subject. If the best fit was less than 0.1, we concluded that there was evidence that the subject's data could be accounted for by assuming that the subject was following the corresponding strategy.