Running head: INSIGHT AND STRATEGY IN MULTICUE LEARNING

Insight and Strategy in Multiple Cue Learning

David A. Lagnado

University College London

Ben R. Newell

University of New South Wales

Steven Kahan and David R. Shanks

University College London

Keywords:

Multiple cue learning, self-insight, strategy, rolling regression, implicit vs. explicit learning

Address for correspondence:

David A. Lagnado

Department of Psychology

University College London

Gower Street

London WC1E 6BT, UK

d.lagnado@ucl.ac.uk

Telephone: +44 (0) 20 7679 5389

Fax: +44 (0) 20 7436 4276

Abstract

In multiple-cue learning (also known as probabilistic category learning) people acquire information about cue-outcome relations and combine these into predictions or judgments. Previous studies claim that people can achieve high levels of performance without explicit knowledge of the task structure or insight into their own judgment policies. It has also been argued that people use a variety of suboptimal strategies to solve such tasks. In three experiments we re-examined these conclusions by introducing novel measures of task knowledge and self-insight, and using 'rolling regression' methods to analyze individual learning. Participants successfully learned a four-cue probabilistic environment and showed accurate knowledge of both the task structure and their own judgment processes. Learning analyses suggested that the apparent use of suboptimal strategies emerges from the incremental tracking of statistical contingencies in the environment.

Insight and Strategy in Multiple Cue Learning

A fundamental goal of cognition is to predict the future on the basis of past experience. In an uncertain world this involves learning about the probabilistic relations that hold between the available information and an outcome of interest, and integrating this into a singular judgment. Thus a stock-broker draws on various market indicators to predict the future price of a share, and a race-course pundit uses factors such as form and track condition to pick the likely winner of a race.

The underlying structure of this kind of prediction is captured in the multiple cue learning framework (Brunswik, 1943; Hammond, 1955), which focuses on how people learn from repeated exposure to probabilistic information. This paradigm, also known as probabilistic category learning, has been applied in numerous areas of psychology, including human judgment (Brehmer, 1979; Doherty & Kurz, 1996; Klayman, 1988; for a review see Goldstein, 2004), learning and memory (Gluck & Bower, 1988; Knowlton, Squire & Gluck, 1994; Shanks, 1990), neuroscience (Ashby & Ell, 2001; Poldrack et al., 2001) and social cognition (Gavanski & Hoffman, 1986; Rappoport & Summers, 1973).

A prototypical experimental task is the weather prediction task (Knowlton et al., 1994). In this task people learn to predict a binary outcome (rainy or fine weather) on the basis of four binary cues (four distinct tarot cards, see Figure 1). Each card is associated with the outcome with a different probability, and these combine to determine the probability of the outcome for any particular pattern of cards. The trials in a task are made up of a representative sampling of possible card combinations. On each trial participants see a specific pattern of cards, predict the weather, and receive feedback as to the correct outcome. This enables them to gradually learn the cue-outcome relations, and thus improve the accuracy of their predictions.

There has been extensive research on the conditions under which people are able to master such tasks (Brehmer, 1980; Doherty & Balzer, 1988; Klayman, 1988). The main findings are that people perform well when the cues are few in number and linearly related to the outcome, when the content of the problem is meaningful, and when they receive sufficient trials and appropriate feedback. An important question that has received less attention concerns the relation between people's performance and their knowledge of what they are doing.

The Apparent Dissociation of Learning and Insight

An intriguing finding emerging from recent studies is that even when people perform well in such tasks, they seem to lack insight into how they achieve this (Evans, Clibbens, Cattani, Harris, & Dennis, 2003; Gluck, Shohamy & Myers, 2002; Harries, Evans & Dennis, 2000; Wigton, 1996; York, Doherty & Kamouri, 1987). This is illustrated in Gluck et al.'s (2002) study with the weather prediction task (hereafter WP task). They found that although participants attained high levels of predictive accuracy they demonstrated little explicit knowledge about what they were doing. In particular, in questionnaires administered at the end of the task they gave inaccurate estimates of the cue-outcome probabilities, and there was little correspondence between self-reports about how they were learning the task and their actual task performance.

This apparent dissociation between learning and insight has been taken as evidence for two separate learning systems (Ashby et al., 2003; Knowlton et al., 1996; Reber & Squire, 1999; Squire, 1994): (i) an implicit (or procedural) system that operates in the absence of awareness or conscious control, and is inaccessible to self-report; (ii) an explicit (or declarative) system that requires awareness and involves analytic processing. Tasks like the WP task, which require the gradual learning and integration of probabilistic information, are

generally considered to involve (i). The lack of self-insight on this task is thus explained by the operation of an implicit system to which participants lack access.

It is also argued that these two systems are subserved by distinct brain regions that can be differentially impaired (Ashby et al., 2003; Knowlton et al., 1996; Poldrack et al., 2001). Thus the WP task has been used to reveal a distinctive pattern of dissociations amongst patient populations. For example, Parkinson's disease patients with damage to the basal ganglia show impaired learning on the task, despite maintaining good explicit memory about task features (Knowlton, Mangels & Squire, 1996). In contrast, amnesic patients with damage to the medial temporal lobes appear to show normal learning but poor declarative memory of the task (Knowlton et al., 1996; Reber, Knowlton & Squire, 1996).

If correct, these conclusions have wide repercussions for everyday reasoning, and for the understanding and treatment of patients with neurological damage. However, there are several reasons to be cautious about this dual-process framework. In this paper we will focus on two fundamental issues: the measurement of insight and the analysis of individual learning.

The Measurement of Insight

It is important to distinguish between a learner's insight into the structure of a task (*task knowledge*) and their insight into their own judgmental processes (*self-insight*). In the case of the WP task, this translates into the difference between a learner's knowledge of the objective cue-outcome associations, and their knowledge of how they are using the cues to predict the outcome. And there is no guarantee that the two coincide. Someone might have an incorrect model of the task structure, but an accurate model of their own judgment process. Politicians seem particularly prone to this.

Previous research tends to run the two together. Thus it is not always clear whether claims about the dissociation between insight and learning refer to a dissociation between

self-insight and learning, task knowledge and learning, or both. Further, this conflation can infect the explicit tests given to participants. Questions that are designed to tap someone's insight into their own judgment processes may instead be answered in terms of their knowledge about the task. Such confusion needs to be avoided if firm conclusions are to be drawn about the relation between learning and insight.

Insensitivity of Explicit Tests

There are several other problems with the explicit tests commonly used to measure task knowledge and self-insight. First, these measures tend to be retrospective, asked after participants have completed numerous trials, and this can distort the validity of their assessments. The reliance on memory, possibly averaged across many trials, can make it difficult to recall a unique judgment strategy. This is especially problematic if people's strategies have varied during the course of the task, making it hard if not impossible to summarize in one global response. In general it is better to get multiple subjective assessments as close as possible to the actual moments of judgment (cf. Ericsson & Simon, 1980; Harries & Harvey, 2000; Lovibond & Shanks, 2002).

Second, explicit tests often require verbalization, but this can also underestimate task knowledge because participants may know what they are doing but be unable to put this into words. This is particularly likely in probabilistic tasks, where natural language may not be well adapted to the nuances of probabilistic inference.

A third problem, prevalent in the neuropsychological literature, is that explicit tests are often too vague (Lovibond & Shanks, 2002). Rather than focus on specific task-relevant features they include general questions that are tangential to solving the task (e.g., questions about location of cards on screen). Once again this reduces the sensitivity of the test to measure people's relevant knowledge or insight.[1] This can lead to an over-estimation of

insight (e.g., someone may be able recall features of the task irrelevant to good task performance) or to under-estimation (e.g., the questions fail to ask about critical information).

The studies in this paper seek to improve the sensitivity of explicit tests on all these counts. Separate explicit tests will be used for task knowledge and self-insight, and both will involve specific questions of direct relevance to the task. In the case of task knowledge these will concern probability ratings about cue-outcome relations, in the case of self-insight subjective ratings about cue usage. Such ratings-based tests will avoid any problem of verbalization. To tackle the problem of retrospective assessments we will take multiple judgments during the course of the task (either blocked or trial-by-trial).

Analyses of Judgment and Learning

Claims about the dissociation between insight and learning also depend on an appropriate analysis of learning performance. In tasks with a dichotomous outcome, such as the WP task, learning is measured in terms of the proportion of correct predictions over a sequence of trials. Standard analyses then average across both individuals and trials to produce a mean percentage correct for the whole task. While this approach is useful for broad comparisons across different tasks, it provides no information about the dynamics of individual judgment and learning.

The Lens Model

A richer approach is provided by the lens model framework (for overviews see Cooksey, 1996; Goldstein, 2004), which assumes that people construct internal models to reflect the probabilistic structure of their environment. A central tenet is that judgmental processes should be modeled at the individual level before any conclusions can be drawn by averaging across individuals. An individual's judgment policy is captured by computing a multiple linear regression of their judgments onto the cue values (across all task trials). The resultant coefficients for each cue are then interpreted as their *subjective* weights (or cue

utilization weights). In a parallel fashion the *objective* weights for that individual are computed by a multiple regression from the outcomes experienced onto the cue values. This technique allows for the possibility that different individuals experience different objective weights.

Individual learning is then assessed by comparing subjective and objective weights. Moreover, task knowledge can be assessed by comparing an individual's objective weights with their explicit ratings of the cue-outcome associations, and self-insight assessed by comparing their subjective weights with their explicit ratings of their own cue usage.

The lens model framework thus provides an analysis of individual judgment. However, although it avoids the loss of information incurred by averaging over participants, it still loses information by averaging over trials. It fails to capture the dynamics of a learning task – both in terms of potential changes in the environment, and potential changes in a judge's policy. In particular, the reliance on global weights ignores the fact that both subjective and objective weights can vary across the course of a task.

This problem arises even with stationary environments (as in the WP task), because the cue-outcome patterns experienced early on in the task might not be representative of the global structure. Analyzing individuals just in terms of their averaged performance across all the trials ignores this possibility, and can under-estimate performance. Moreover, it overlooks the possibility that individuals might change their judgment policies over time, and that such changes may track variations in the actual environment.

A related shortcoming is that these global analyses assume that the judge has a perfect memory for all trials, and treat earlier trials in exactly the same way as later ones. But both of these assumptions are questionable – people may base their judgments on a limited window of trials, and may place more emphasis on recent trials (Slovic & Lichtenstein, 1971).

Dynamic Models of Learning

The need for dynamic models of learning is now widely recognized (Dayan, Kakade & Montague, 2000; Friedman, Massaro, Kitzis & Cohen, 1995; Kitzis et al., 1998; Smith et al., 2004). A natural extension to the lens model is the 'rolling regression' technique introduced by Kelley and Friedman (2002) to model individual learning in economic forecasting. In their study participants learned to forecast the value of a continuous criterion (the price of Orange juice futures) on the basis of two continuous-valued cues (local weather hazard and foreign supply). Individual learning curves were constructed by computing a series of regressions (from forecasts to cues) across a moving window of consecutive trials. For example, for a window size of 160 trials, the first regression is computed for trials 1 to 160, the next for trials 2 to 161, and so on. This generates trial-by-trial estimates (from trial 160 onwards) for an individual's cue utilization weights, and thus provides a dynamic profile of the individual's learning (after trial 160).

Each learning profile was then compared with the profile of an 'ideal' learner exposed to the same moving window. Regressions for each ideal learner are also computed repeatedly for a moving window of trials, but in this case the actual criterion values (prices) are regressed onto the cues. The estimates of the ideal learner thus correspond to the best possible estimates of the objective cue weights for each window of trials.

This technique provides dynamic models of both actual and ideal learners, and permits comparisons between them over time. For example, Kelley and Friedman showed that whereas ideal learners converged quickly on the actual weights, the participants learned more slowly, and over-estimated these weights towards the end of the task. In this paper we will use similar techniques to analyze individual learning, but with a smaller window size to simulate a more realistic memory constraint. This should provide a finer-grained analysis of the dynamics of individual learning.

It is important to note that even an ideal learner is not guaranteed perfect performance on a probabilistic task. First, the noisy nature of the environment means that optimal performance is bounded. In the WP task 83% correct predictions is the best that can be expected even if the learner knows the objectives probabilities for each cue pattern. Second, predictive success also depends on the learner's choice rule (Friedman & Massaro, 1998; Nosofsky & Zaki, 1998). Given their estimates for the outcome probabilities (conditional on each cue pattern) they might always select the most probable outcome (maximize) or distribute their choices to match the outcome probabilities (matching). Only the former leads to optimal performance. This issue will be revisited in the General Discussion.

Strategy Analyses

An alternative approach is developed by Gluck et al. (2002). They identified three main strategies employed in the WP task: (1) a multi-cue strategy that used all cards to predict the outcome; (2) a singleton strategy that only used cue patterns with a single card (and guessed on all other patterns); and (3) a one-cue strategy that used just one card (and ignored any other cards present). Ideal profiles were constructed for each of these strategies, and these were fit to the participant data. Overall the best fitting model was the singleton strategy, with a shift from singleton to multicue strategies as the task progressed.

This improves over previous analyses of the WP task, and resonates with recent work on the use of simple heuristics (Gigerenzer et al., 1999). However, it suffers from the use of global cue-outcome associations to construct the ideal profiles for each strategy. This ignores the possibility that participants encounter a non-representative sampling of the environment early on in the task, and therefore may under-estimate the number using a multi-cue strategy. For example, in fitting the multi-cue strategy it is assumed that participants know the correct cue-outcome associations from the outset. This is an unrealistic assumption.

The rolling regression method overcomes this problem by comparing individual learners to an ideal learner exposed to the same information. It also offers a more parsimonious explanation of the apparent strategy shifts in Gluck et al. (2002). People appear to use sub-optimal strategies early on in the task because they are in the process of learning each of the cue weights. The apparent shift to multi-cue strategies emerges as they learn all of these cue weights. The existence of both strong and weak cues will promote this pattern: stronger cues will tend to be learned before weaker ones. A single learning mechanism thus suffices to explain the experimental data.

Strategy and Insight

What is the relation between strategy and insight? Gluck et al. (2002) assume that strategies are implicit, so there need be no connection between the two. In contrast, opponents of the implicit/explicit distinction would maintain that the two are closely related. For example, multi-cue strategies require richer explicit knowledge than single-cue strategies. By using better and more frequent measures of explicit knowledge, and correlating these with implicit measures of learning, this question can be addressed.

In sum, this paper aims to extend our understanding of how people learn in multiple cue tasks, and re-examine the prevalent claim that good performance can be achieved in the absence of insight. In doing so, it will introduce more appropriate measures of both task knowledge and self-insight, and more fine-grained methods for analyzing the dynamics of individual learning.

Experiment 1

Method

Participants and Apparatus

Sixteen students from University College London took part in the study. They were paid a basic turn-up fee of £2, and received additional payment depending on their

performance in the task. The entire experiment was run on a laptop computer using a software program written in Visual Basic 6. Participants were tested individually in a sound-proofed room.

Materials

The stimuli presented to participants were drawn from a set of four cards, each with a different geometric pattern (squares, diamonds, circles, triangles; see Figure 1). Participants saw a total of 200 trials, on each of which they were presented with a pattern of one, two or three cards. Each trial was associated with one of two outcomes (Rain or Fine), and overall these two outcomes occurred equally often. The pattern frequencies are shown in Table 1, along with the probability of the outcome for each of these 14 patterns. The learning set was constructed so that each card was associated with the outcome with a different independent probability. For example, the probability of rain was 0.2 over all the trials on which the squares card (card 1) was present, 0.4 for trials on which the diamonds card (card 2) was present, 0.6 for trials on which the circles card (card 3) was present, and 0.8 for trials on which the triangles card (card 4) was present.

In short, two cards were predictive of rainy weather, one strongly (card 4), one weakly (card 3), and two cards were predictive of fine weather, one strongly (card 1), one weakly (card 2). Overall participants experienced identical pattern frequencies (order randomized for each participant), but the actual outcome for each pattern was determined probabilistically (so experienced outcomes could differ slightly across participants). The position of the cards on the screen were held constant within participants, but counterbalanced across participants.

Procedure

At the start of the experiment participants were presented with the following on-screen instructions:

In the experiment you will be playing a "game" in which you pretend to be a weather forecaster. On each trial you will see between one and three "tarot cards" (cards with squares, diamonds, circles or triangles drawn on them). Your task is to decide if the combination of cards presented predicts RAINY weather or FINE weather. At first you will have to guess, but eventually you will become better at deciding which cards predict RAINY or FINE weather. As an incentive to learn how to predict the weather, you will be paid 1p for every correct forecast you make. A scale on the screen will display your current earnings throughout the experiment. There are 200 trials in total. After each block of 50 trials there will be a short break during which you will be asked some questions about the task and your performance.

After reading the instructions participants moved onto the first block of 50 trials. On each trial a specific pattern of cards (selected from Table 1) was displayed side-by-side on the screen (see Figure 2). Participants were then asked to predict the weather on that trial, by clicking either on RAINY or FINE. Once they had made their prediction, participants received immediate feedback as to the actual weather on that trial, and whether they were correct or incorrect. Correct responses were signalled with a thumbs-up sign, and a 1 pence increment in the on-screen earnings indicator. Incorrect responses were signalled with a thumbs-down, and no change in the earnings indicator.

At the end of each block of fifty trials participants answered two different sets of test questions. In the *probability* test participants were asked to rate the probability of rainy vs. fine weather for each of the four cards: 'On the basis of this card what do you think the weather is going to be like?' They registered their rating using a continuous slider scale ranging from 'Definitely fine' to 'Definitely rainy', with 'As likely fine as rainy' as the midpoint. In the *importance* test participants were asked how much they had relied on each card to make their predictions: 'Please indicate how important this card was for making your predictions'. They registered their rating using a continuous slider scale ranging from 'Not important at all' to 'Very important', with 'Moderately important' as the midpoint.

Results and Discussion

Learning Performance

Across the task participants steadily improved in their ability to predict the outcome. The mean proportions of correct predictions for each block of fifty trials are shown in Figure 3. A linear trend test showed a significant improvement across blocks, $F(1, 15)=10.23$, $MSE = 0.19$, $p < 0.05$, and by the final block mean performance approached the optimal level of 83% correct.

Probability Ratings

After each block of fifty trials participants judged the probability of the weather for each individual card. The mean probability ratings of Rain for each card across the four blocks are shown in Figure 4 (upper panel). An ANOVA with card type (1-4) and block (1-4) as within-subject factors revealed a main effect of card type, $F(3, 45) = 16.79$, $MSE = 2400$, $p < 0.01$, no effect of block, $F(3, 45) = 0.64$, $MSE = 303$, *ns.*, and an interaction between card type and block, $F(9, 135) = 2.65$, $MSE = 338$, $p < 0.01$.

Inspection of Figure 4 shows that participants improved in their ability to discriminate between the probabilities of each card through the course of the task. Recall that the actual probabilities of Rain for each card were 0.2, 0.4, 0.6 and 0.8 for cards 1-4 respectively. By block 4 mean probability estimates were close to the actual values, except for card 4, where the actual value was slightly over-estimated.

The observed interaction between card type and block suggests that participants learned about strong cards (and discriminated between them) sooner than they did for weak cards. This is supported by the fact that ratings for card 1 and card 4 differed significantly by block 1, $t(15) = 4.18$, $p < 0.01$, whereas ratings for card 2 and card 3 do not differ on block 1, $t(15) = 0.36$, *ns.*, and only differed significantly by block 3, $t(15) = 2.39$, $p < 0.05$.

In sum, throughout the course of the task participants gradually learned the probabilistic relations between cards and outcomes, and they tended to learn about the strong cards (cards 1 & 4) sooner than the weak cards (cards 2 & 3).

Cue Usage Ratings

After each block participants rated how much they had relied on each card in making their predictions. The main question of interest here is whether participants' usage ratings discriminated between strongly and weakly predictive cards. Thus we combined ratings for the two strongly predictive cards (cards 1 and 4) into one group (*strong*), and ratings for the two weakly predictive cards (cards 2 and 3) into another group (*weak*). The mean cue usage ratings for each group across the four blocks are shown in Figure 4 (lower panel). An ANOVA with card strength (weak vs. strong) and block (1-4) as within-subject factors revealed a main effect of card strength, $F(1, 15) = 7.11$, $MSE = 766$, $p < 0.05$, no effect of block, $F(3, 45) = 0.78$, $MSE = 129$, *ns.*, and a marginal interaction between card strength and block, $F(3, 45) = 2.77$, $MSE = 227$, $p = 0.05$.

Paired comparisons revealed that the difference between strong and weak groups was not significant for block 1, $t(15) = 0.09$, but was marginally significant by block 2, $t(15) = 2.03$, $p = 0.06$, and significant for block 3, $t(15) = 2.54$, $p < 0.05$, and block 4, $t(15) = 3.85$, $p < 0.01$. These tests confirm that as the task progressed participants explicitly reported that they relied more on strong cards than weak ones.

Rolling Regression Analyses

A finer-grained analysis of individual learning is provided by the use of rolling regressions. This technique was introduced by Kelley and Freidman (2002) to model continuous predictions. In order to apply it to the binary predictions made in the WP task logistic rather than linear regression was used. We also used a smaller window size (50 rather

than 160 trials)[2] to provide a more realistic memory constraint and to capture the dynamical fluctuations in the environment.

Two trial-by-trial learning profiles were constructed for each participant, one to capture their own implicit policy, the other to capture the policy of an ideal learner exposed to the same information. Both learning profiles were generated by computing a sequence of logistic regressions across a moving window of 50 trials. Thus each participant's *implicit* profile was composed of four beta-weight curves (from trial 51 to 200), one for each card, tracking the relation between the card values on each trial and the outcome that the participant predicted on that trial. The *ideal* profile for each participant was also made up of four beta-weight curves, in this case tracking the relation between the card values and the actual outcomes. It is termed 'ideal' because each beta-weight corresponds to the best possible (LMS) estimate for the weight given the experienced data (Kelley & Friedman, 2002). This technique allows us to view the dynamics of learning for each individual and to compare this with an ideal learner exposed to the same trials. It also permits comparisons with the explicit ratings that participants make during the task. To illustrate, Figure 5 shows the implicit profiles for two participants (selected as best and worse performers, see below). The card weights have been averaged across blocks of ten trials (the first block corresponding to trials 51-60, the second to 61-70, and so on).

The regression weights on the y-axis show how strongly each card predicts the individual's choice (with positive values predicting rain, and negative values predicting fine). The regression weight for a given card corresponds to the log-likelihood ratio for the outcome given the presence of that card.

$$Weight(card_i) = \ln (P(rain|card_i)/ P(fine|card_i))$$

This can be translated into the odds (or probability) of the outcome given a specific card. For example, a weight of +2.2 for card 4 corresponds to a likelihood ratio = 9.03, and thus to approximate odds of 9:1 in favour of Rain (probability = 0.9).

Informal inspection of Figure 5 shows that the implicit learning curves for the best performer reflect the objective probabilities in the task environment: the cards are appropriately ordered (weight for card 4 > card 3 > card 2 > card 1), associated with the correct outcome (cards 1 & 2 are negative, corresponding to a prediction of fine, cards 3 & 4 are positive, corresponding to a prediction of rain), and steadily increase across trials towards an asymptote. This contrasts with the worst performer, who has the cards in the wrong order, and has an incorrect positive weight for card 2 throughout (which was also reflected by their explicit probability ratings for that card). Formal analyses for each individual learning profile are given below.

Correlational Analyses

Correlational measures were used to assess the relations between a learner's implicit weights, the implicit weights of an ideal learner exposed to the same environment, and the learner's explicit probability ratings. These three measures were computed for each participant, on an individual basis, after 50, 100, 150 and 200 trials (corresponding to the time points at which the explicit probability ratings were made). Thus twelve correlation coefficients were derived for each participant. Because of the small number of cards (n=4) we computed rank correlations (Spearman's rank correlations r_s). For a summary measure at each time point we averaged across individuals to obtain mean correlation coefficients for each of the three kinds of correlation (see Table 2). These were tested for significance using standard *t*-tests.

As can be seen from Table 2, all the mean correlation coefficients were significantly positive. This shows that there were strong positive correlations between: (i) the participants'

implicit weights and their corresponding ideal weights, (ii) the participants' implicit weights and their explicit probability ratings, and (iii) the participants' explicit probability ratings and the ideal weights. In addition, linear trend tests revealed positive trends for all three measures (see final column in Table 2). Thus all three kinds of correlation increased as the task progressed.

Examination of Table 2 also reveals that the mean correlations between implicit and ideal weights were always slightly higher than those between ideal weights and explicit ratings. Rather than reflect a dissociation between implicit and explicit processes, this can be explained by the fact that both implicit and ideal weights were derived from an identical regression procedure (across a window of 50 trials), whereas the explicit ratings were taken at single time-points and used a very different measurement scale.

The individual correlations between implicit and ideal weights were used to select the best and worst performers shown in Figure 5. The best performer had correlations of +0.8, +0.8, +1.0 and +1.0 for the four time points. The worst performer had correlations of +0.2, -0.4, +0.4 and +0.2.

Group Analyses

The implicit profiles for each participant were averaged to yield group learning curves (see Figure 6, upper panel). Inspection of Figure 6 shows that the cards are ordered correctly (weights for card 4 > card 3 > card 2 > card 1) and associated with the appropriate outcome (cards 1 & 2 are negative, cards 3 & 4 are positive). This was confirmed by statistical analyses. An ANOVA with card and block as within-participant factors revealed a main effect of card, $F(3, 45) = 21.20$, $MSE = 245$, $p < 0.01$, no effect of block, $F(14, 210) = .32$, $MSE = 11.1$, *ns.*, and a card by block interaction, $F(42, 630) = 3.71$, $MSE = 17.2$, $p < 0.01$. The interaction between card and block reflects the increasing discrimination between card weights as the task progresses. Separate *t*-tests on the last block showed that card 4 was

weighted higher than card 3, $t(15) = 2.93$, $p < 0.01$, card 3 was weighted higher than card 2, $t(15) = 4.10$, $p < 0.01$, and card 2 was weighted higher than card 1, $t(15) = 2.58$, $p < 0.05$.

Overshooting

The group-level implicit profiles were compared with the group-level ideal profiles. Figure 6 (lower panel) shows both implicit and ideal curves for cards 1 and 4. It suggests that as the task progresses the implicit weights tend to over-shoot the ideal weights. This pattern was observed for all four cards. Statistical tests on the last block showed that this over-shooting was significant for card 4, $t(15) = 2.64$, $p < 0.05$, and card 3, $t(15) = 2.40$, $p < 0.05$, marginally significant for card 1, $t(15) = 1.78$, $p = 0.095$, and not significant for card 2, $t(15) = .51$.

Over-shooting was also present in most of the individual profiles (15 out of 16). This tendency for participants' implicit weights to over-shoot the ideal weights was also found in Kelley and Friedman (2002) and will be considered in the General Discussion.

Strong vs. Weak Cards

To confirm that participants distinguished between strong and weak cards we compared the averaged weights for cards 1 and 4 (*strong*) with the averaged weights for cards 2 and 3 (*weak*). The grouped profiles are shown in Figure 7. An ANOVA with card strength and block as within-participant factors revealed main effects of card strength, $F(1, 15) = 27.02$, $MSE = 20.4$, $p < .01$, and block, $F(14, 210) = 5.25$, $MSE = 17.5$, $p < .01$, and a card strength by block interaction, $F(14, 210) = 2.56$, $MSE = 3.18$, $p < .01$. Individual *t*-tests showed that strong cards were weighted higher than weak cards as early as block 1, $t(15) = 2.20$, $p < 0.05$. These implicit measures correspond well with the explicit measures of cue usage shown in Figure 4.

Strategy Analyses

An alternative way to analyse individual performance is in terms of Gluck et al.'s (2002) strategy analyses. As pointed out in the introduction, one limitation of this approach is that the multi-cue strategy is a very demanding standard. It requires that people know the probabilities associated with each of the 14 patterns, and that they always choose the outcome that is most probable (maximizing). This will be difficult early on in the task when the card weights have not been fully learned. So the failure of this multi-cue strategy to fit many participants, especially early on in the task, does not rule out the possibility that they are adopting some version of a multi-cue strategy.

To address this shortcoming we introduced a variant of the original multi-cue strategy which probability matches rather than maximizes. Like the original multi-cue strategy (*multi-max*) it takes all four cards into consideration, but rather than always responding with the most probable outcome it distributes its responses in proportion to the outcome probabilities. For example, when presented with pattern B (just card 3 present), which is associated with rain 78% of the time, multi-max would always predict rain. In contrast, the multi-match strategy would predict rain 78% of the time, and fine 22% of the time. Evidence from a wide-range of probabilistic tasks suggests that people often adopt such a matching strategy, even when it is not optimal (e.g., Shanks, Tunney & McCarthy, 2002; Tversky & Edwards, 1966; West & Stanovich, 2003).

We used the same method as Gluck et al. (2002) to fit participants' learning profiles to these four strategies. The basic procedure was to calculate the degree to which each model fit the participant's data using a least mean squares measure (see Appendix for details). Across the complete 200 trials the multi-match strategy fit the vast majority of participants (14 out of 16, 88%). The multi-max and singleton strategies fit one participant each (6%) and the one-cue strategy fit none. We also computed model fits over every block of 50 trials. The

results are displayed in Figure 8, and show a clear dominance of the multi-match strategy, with an increase in multi-max as the task progressed.

These analyses complement the findings from our rolling regression analyses, suggesting that people gradually learn the weights of all four cards as the task progresses. They also support our contention that previous strategy analyses by Gluck et al. (2002) under-estimated the number of people adopting multi-cue strategies (by limiting their analysis to a multi-max strategy).

Experiment 2

Experiment 1 showed that people can achieve close to optimal performance in a multi-cue probabilistic learning task, and this is accompanied by accurate task knowledge and good self-insight. This contrasts with previous claims that such learning tasks recruit implicit learning processes (Ashby & Maddox, 2005; Gluck et al., 2002; Knowlton et al., 1994). To further investigate people's explicit knowledge of their own judgment processes, Experiment 2 introduces a trial-by-trial measure of self-insight. This involves asking participants to rate how much they relied on each card immediately after they have made a prediction. A similar technique has been used by Harries and Harvey (2000), and the resultant cue weightings were closely correlated with implicit judgment policies (as revealed by regression techniques). However, this pattern was only found when participants were engaged in an advice integration task, not with a contingency-based task (such as the WP task). The former task involves the integration of several numerical estimates from different advisors to yield a single estimate for a numerical prediction. One difference between Harries and Harvey (2000) and the current study is that they compared participants' self-ratings against regression models computed across the global set of trials, rather than a moving window. This ignores the fact that participants actually experience a stochastic learning environment, and may have a limited memory window. A second difference is that in their study the cues and outcome

were continuous-valued rather than binary. Both of these differences may account for their finding that participants lacked insight in the contingency-based task.

Method

Participants and Apparatus

Sixteen students from University College London took part in the study. Payment and testing conditions were the same as in Experiment 1.

Procedure

The stimuli and instructions were identical to Experiment 1 except for the addition of a cue rating stage after each prediction. On each trial, just after the participant had predicted the weather outcome, a drop-down menu appeared beneath each card present on that trial. Participants were asked "How much did you rely on each card in making your prediction?" and had to select between four options – "Greatly", "Moderately", "Slightly", "Not at all". Once they had registered their judgments they were shown the actual outcome, and then continued to the next trial.

Results and Discussion

Learning Performance

As in Experiment 1 participants improved in their ability to predict the outcome. The mean proportions of correct predictions for each block of fifty trials are shown in Figure 9. In contrast to Experiment 1, however, there is a tail-off in performance towards the end of the task. This is confirmed by a significant quadratic trend, $F(1, 15) = 12.17$, $MSE = 0.005$, $p < 0.05$, but no linear trend, $F(1, 15) = 3.51$, $MSE = 0.007$, *ns*.

This tail-off in learning performance is most likely due to the extra cognitive effort introduced by the trial-by-trial cue usage ratings. Experiment 2 took longer to complete than Experiment 1, and participants might be showing signs of fatigue towards the end of the task.

Probability Ratings

The mean probability ratings of Rain for each card across the four blocks are shown in Figure 10 (upper panel). An ANOVA with card type (1-4) and block (1-4) as within-subject factors revealed a main effect of card type, $F(3, 45) = 30.97$, $MSE = 1601$, $p < 0.01$, no effect of block, $F(3, 45) = 0.88$, $MSE = 219$, *ns.*, and an interaction between card type and block, $F(9, 135) = 2.07$, $MSE = 299$, $p < 0.05$.

Inspection of Figure 10 shows that participants improved in their ability to discriminate between the probabilities of each card through the course of the task. Once again their estimates tended to approximate the actual probability values, with the exception of the over-estimation of one card.

The observed interaction between card type and block suggests that strong cards were learned about (and discriminated between) sooner than weak cards. This is supported by the fact that ratings for card 1 and card 4 differ significantly by block 1, $t(15) = 7.69$, $p < 0.01$, whereas ratings for card 2 and card 3 do not differ on block 1, $t(15) = 0.01$, *ns.*, and only differ significantly by block 3, $t(15) = 3.41$, $p < 0.01$.

In sum, as with Experiment 1, participants gradually learn the probabilistic relations between cards and outcomes, but tend to learn about the strong cards (cards 1 & 4) sooner than the weak cards (cards 2 & 3).

Cue Usage Ratings (Blocked)

The mean cue usage ratings for weakly vs. strongly predictive cards across the four blocks are shown in Figure 10 (lower panel). An ANOVA with card strength (weak vs. strong) and block (1-4) as within-subject factors revealed a main effect of card strength, $F(1, 15) = 13.01$, $MSE = 976$, $p < 0.01$, no effect of block, $F(3, 45) = 0.77$, $MSE = 120$, *ns.*, and an interaction between card strength and block, $F(3, 45) = 4.83$, $MSE = 200$, $p < 0.01$.

Paired comparisons revealed that the difference between strong and weak groups was not significant for block 1, $t(15) = 0.47$, but was significant by block 2, $t(15) = 4.69$, $p < 0.01$,

and for block 3, $t(15) = 4.37$, $p < 0.01$, and block 4, $t(15) = 2.89$, $p < 0.05$. This replicates the finding in Experiment 1 that as the task progressed participants reported that they relied more on strong cards than weak ones.

Cue Usage Ratings (Trial-by-Trial)

In addition to blocked cue usage ratings participants also made similar ratings on a trial-by-trial basis. As before these were grouped into ratings for strong and weak cards. Figure 11 (upper panel) presents the mean ratings for strong and weak cards, where values on the y-axis represent how much participants said they relied on each card, with 4 = "Greatly", 3 = "Moderately", 2 = "Slightly", 1 = "Not at all". Inspection of this figure reveals that from early on in the task participants rated strong cards as more important than weak ones. Indeed by block 2 (trials 10-20) there is a significant difference between strong and weak cards, $t(15) = 2.22$, $p < 0.05$. Moreover, at the individual level, 11 out of 16 participants were already rating the strong cards as more important than the weak ones at block 2. This explicit measure supports the conclusion that participants develop insight into their own cue usage relatively early in the task.

Rolling Regression Analyses

The same analyses were carried out as in Experiment 1. Figure 12 shows the implicit profiles for the best and worst performers, with the best performer achieving correlations between implicit and ideal weights of +1.0, +1.0, +0.8 and +1.0, and the worst performer correlations of +0.6, -0.8, +0.8 and +0.8. The main mistake made by the latter was to get the associations of cards 1 and 4 wrong at block 5 (which was also reflected in their explicit probability ratings).

Correlational Analyses

Correlational measures were again used to assess the relations between a learner's implicit weights, the implicit weights of an ideal learner exposed to the same environment,

and the learner's explicit probability ratings. The results for Experiment 2 are shown in Table 2. As with Experiment 1 all the mean correlations were significant, showing strong positive correlations between: (i) the implicit and ideal weights, (ii) the implicit weights and explicit probability ratings, and (iii) the explicit probability ratings and ideal weights. Further, also in line with Experiment 1, (ii) and (iii) both showed significant linear trends (see final column in Table 2). The only difference from Experiment 1 was the lack of a linear trend for the correlations between implicit and ideal weights. This might be due to the correlation reaching a ceiling by trial 150.

Group Analyses

The implicit profiles for each participant were averaged to yield group learning curves (see Figure 13, upper panel). Inspection of this figure shows that the cards are ordered correctly (weights for card 4 > card 3 > card 2 > card 1) and associated with the appropriate outcome (cards 1 & 2 are negative, cards 3 & 4 are positive). This was confirmed by statistical analyses. An ANOVA with card and block as within-participant factors revealed a main effect of card, $F(3, 15) = 18.43$, $MSE = 270$, $p < 0.01$, no effect of block, $F(14, 210) = .54$, $MSE = 12.5$, *ns.*, and a card by block interaction, $F(42, 630) = 1.63$, $MSE = 18.6$, $p < 0.01$. Separate t-tests on the last block showed that card 4 was weighted higher than card 3, $t(15) = 2.49$, $p < 0.05$, card 3 was weighted higher than card 2, $t(15) = 2.37$, $p < 0.05$, and card 2 was weighted higher than card 1, $t(15) = 2.83$, $p < 0.05$. The interaction between card and block reflects the increasing discrimination between card weights as the task progresses.

Overshooting

The group-level implicit profiles were compared with the group-level ideal profiles. Figure 13 (lower panel) shows both implicit and ideal curves for cards 1 and 4. It replicates the finding in Experiment 1 that implicit weights tended to over-shoot ideal weights as the task progresses. This pattern was observed for all four cards. Statistical tests on the last block

showed that this over-shooting was significant for card 4, $t(15) = 2.61$, $p < 0.05$, and card 1, $t(15) = 2.36$, $p < 0.05$, but not for card 2, $t(15) = 1.72$, or card 3, $t(15) = 1.69$. This over-shooting was present in all of the individual profiles.

Strong vs. Weak Cards

The implicit profiles for strong and weak cards were also compared (see lower panel of Figure 11). An ANOVA with card strength and block as within-participant factors revealed a main effect of card strength, $F(1,15) = 24.86$, $MSE = 32.7$, $p < .01$, no effect of block, $F(14, 210) = 1.24$, $MSE = 20.9$, *ns.*, and a card strength by block interaction, $F(14,210) = 4.91$, $MSE = 2.16$, $p < .01$. Individual *t*-tests showed that by block 2 strong cards were weighted higher than weak cards, $t(15) = 2.62$, $p < 0.05$. These implicit measures correspond well with the explicit measures of cue usage shown in Figure 10.

Note the similarity between participants' implicit trial-by-trial cue usage for strong vs. weak cards in Figure 11 (upper panel) and their explicit trial-by-trial judgments about their cue usage (lower panel). This suggests a nice fit between what participants say they are doing and what they are actually doing.

Strategy Analyses

Across the complete 200 trials the multi-match strategy fit the vast majority of participants (14 out of 16, 88%). The one-cue strategy fit just two participants (12%) and the multi-max and singleton strategies fit none. We also computed model fits over every block of 50 trials. The results are displayed in Figure 14 (upper panel), and show a clear dominance for the multi-match strategy.

Experiment 3

One interesting question about the use of explicit measures during the learning task (blocked or trial-by-trial) is whether these lead participants to solve the task in a qualitatively different manner. It may be that the high concordance we have observed between explicit and

implicit weights is only characteristic of situations in which conscious attention is directed to the task structure and cue usage by explicit questions about these. To address this issue we conducted a replication of Experiments 1 and 2 with explicit measures taken only at the end of the task.

Method

Twelve students from University College London took part in the study under identical conditions to Experiments 1 and 2 except that there were no probability ratings or cue usage judgments during the task. Explicit measures were only taken once, at the end of the task (after trial 200).

Results & Discussion

The learning performance in this new task was very similar to both Experiment 1 and 2. Mean correct predictions on the final block was 73%, which is close to that achieved in Experiment 2 (72%), but slightly less than in Experiment 1 (78%). A more telling comparison is after 100 trials, when participants in Experiment 1 had received explicit measures at trial 50, and those in Experiment 2 had also had explicit measures on every trial. For Experiment 3 mean correct predictions were 73%, effectively no different to Experiment 1 (72%) or Experiment 2 (74%).

There were also no changes in participants' explicit probability ratings or cue usage ratings at trial 200. The mean probability ratings were 22.6 for card 1, 40.8 for card 2, 59.6 for card 3 and 78.3 for card 4. These were very close to the objective probabilities. An ANOVA showed a significant effect of card, $F(3,33) = 11.94$, $MSE = 580$, $p < 0.01$, and a linear trend, $F(1,11) = 9.03$, $MSE = 359$, $p < 0.05$. The mean cue usage rating for strong cards (75.0) was significantly higher than that for weak cards (59.5), $t(11) = 2.27$, $p < 0.05$. These results are very similar to the explicit measures in the final block of the previous two experiments.

We also computed rank correlations between the implicit and ideal weights after trials 50, 150 and 200, and correlations at trial 200 between implicit weights and explicit probability ratings, and ideal weights and explicit ratings. These are shown in Table 2. In line with the previous experiments the mean correlations between participants' implicit weights and ideal weights were all significant and strongly positive. This suggests that the lack of explicit ratings did not affect participants' ability to learn the cue weights. Indeed they reached a high mean correlation (=.8) by trial 100.

Although still strongly positive, the correlation between participants' explicit ratings and their implicit weights was slightly lower than in Experiment 2. This could be interpreted as showing that the fact that participants in Experiment 2 made continual assessments of the individual cards helped them in expressing their probability ratings for those cards. It does not suggest, however, that they solved the task in a qualitatively different way.

Finally, we ran strategy analyses on the new experiment. Across the complete 200 trials the multi-match strategy fit the majority of participants (83%). This is in close agreement with the strategy fits in Experiments 1 and 2. We also computed strategy fits over blocks of 50 trials (see Figure 14, lower panel). This yields a similar picture to the previous experiments except for a greater number of participants fit by multi-max in Experiment 1 (blocks 2-4).

One possible explanation for this is that the explicit measures at trial 50 encouraged some participants to maximize rather than match. However, on this account it is puzzling that the participants in Experiment 2, who gave explicit measures on every trial, did not show a similar shift to maximizing. Alternatively, these slight cross-experiment differences might reflect some variability in the participant groups, and/or idiosyncrasies of the strategy fit method.

Future research could examine these questions in more detail. However, the strategy analyses clearly show that with or without explicit measures, the vast majority of participants use multi-cue strategies throughout the task. Taken together with the other analyses it suggests that the use of explicit measures is unlikely to change the general way in which people approach the task.

General Discussion

In three experiments we investigated how people solved a multi-cue probability learning task. In contrast to previous studies (Evans et al., 2003; Gluck et al., 2002; Wigton, 1996; York et al., 1987), we found that near optimal performance was accompanied with accurate task knowledge and self-insight. Individual learning during the course of the task was analysed using a rolling regression technique (Kelley & Friedman, 2002). This generated cue usage curves for both actual and ideal learners. Correlational analyses on the individual data showed strong positive correlations between: (i) implicit and ideal weights, (ii) implicit weights and explicit probability ratings, and (iii) explicit probability ratings and ideal weights.

The close fit between implicit and explicit measures is well illustrated by comparing the regression profiles for participants' implicit cue usage (Figure 11, upper panel) with their explicit trial-by-trial ratings (Figure 11, lower panel). Both show a clear discrimination between the strong and weak cues that increases as the task progresses. This demonstrates that from early on in the task participants were accurate in their reports about their own cue usage. This flies in the face of received wisdom about the frailties of people's self-reports.

These findings raise serious questions for the received view that probabilistic category learning is a purely implicit task, and that people master it in the absence of awareness and insight. In particular, the standard assumption that such learning involves an implicit procedural system inaccessible to conscious awareness is challenged by our finding that

participants' explicit task knowledge corresponds closely with the actual structure of the learning environment, and that their knowledge of what cues they use corresponds well with how they actually weight the cues. More generally, our findings are problematic for the standard distinction between explicit and implicit learning systems. According to the classifications introduced and defended by proponents of this distinction, we have shown that a paradigm implicit task is being solved by participants in an explicit manner. This has strong repercussions for the various neuropsychological studies that have built upon this alleged distinction (see below).

Models of Self-Insight

Given that most previous research claims that people lack self-insight, there are few positive proposals about its nature. Most theorizing has focused on dual-route models that posit separate mechanisms for people's verbal reports and their non-verbal behaviour. An influential view is provided by Nisbett and Wilson (1977; see also Wilson, 2002). They argue that when people report how particular cues have influenced their responses they do not access internal states, but base their reports on a priori causal models derived from their beliefs about what people typically do (or ought to do) in such situations. This view meshes well with the general implicit-explicit distinction – the way people actually use the cues is implicit (and inaccessible to awareness) whereas their post-experiment reports are explicit but incorrect.

Given the abstract nature of the stimuli in the WP task it is unlikely that participants have an a priori causal model about what people would (or should) do. Consistent with Nisbett and Wilson's account, this can explain why previous studies using post-experiment questionnaires found poor explicit knowledge. However, it does not explain how participants give accurate self-reports when probed with more sensitive measures.

One answer is to take seriously the possibility that people have access to the internal states that influence their behaviour, and that this drives both their online predictions and their explicit verbal reports. This would explain the strong correlation between participants' implicit cue weightings, their explicit cue usage ratings (both online and blocked), and their blocked probability judgments. It would also account for the tendency to conflate task knowledge and self-insight, because on this model both share the same common cause. Thus, although conceptually distinct, participants' explicit judgments about the structure of the task will usually be highly correlated with their expressed cue usage. This is particularly true when they perform well, because high performance requires that their subjective cue weightings correspond to the objective cue weights in the task environment. In short, we conjecture that people gradually acquire cue-outcome weightings that veridically represent the environment to which they are exposed, and that these weightings drive their online predictions and their explicit reports about both the task structure and their own cue usage.

We believe this to be the most parsimonious explanation of the current data. It also fits with evidence from studies in social cognition (Gavanski & Hoffman, 1987)[3]. There are, however, several alternative models proposed in the literature. Following Lovibond and Shanks (2002) we distinguish two main classes: single process models that posit just one declarative learning mechanism, and dual process models that posit two independent learning mechanisms, one declarative and the other procedural. As noted in the introduction, a declarative system is supposed to involve propositional knowledge that is accessible to conscious awareness, whereas a procedural system involves non-propositional knowledge that is inaccessible (Squire, 1994; Cohen, & Eichenbaum, 1993). The distinction is often cast in terms of 'knowing that' versus 'knowing how', with the intuition that there are certain skills that we learn how to perform without any explicit knowledge of what we are doing.

The two classes of model are depicted in Figure 15. Model A asserts a single declarative learning process that drives people's behavioural responses (e.g., their online predictions), their explicit judgments about the task structure (e.g., their probability ratings), and their explicit judgments about their own cue usage (both blocked and online). In contrast, model B asserts that people's online predictions derive from a procedural learning process, and their explicit judgments derive from a separate declarative learning system.

In a broad sense both models are consistent with the data from our current studies. However, the dual-process model is challenged by the close concordance between the explicit measures and online predictions. This model predicts that explicit measures and online predictions sometimes dissociate, but there is no evidence of this in our studies (and little support in the wide range of studies surveyed by Lovibond and Shanks). Thus, not only is the single process model favoured on grounds of simplicity, insofar as it posits one rather than two learning systems, but it is also supported by the close match between behavioural and explicit measures.

A more direct way to discriminate between these models would be to examine whether disruption to the declarative learning system inhibits behavioural responses. If the declarative system can be impaired or disrupted, without compromising online predictions, this would be strong evidence against a single process model. Thus far there is little evidence in support of such a possibility. For example, although Knowlton et al. (1996) and Reber et al. (1996)[4] claimed that amnesics with disrupted declarative systems could still master the WP task, more recent research has shown that amnesics are impaired on the task (Hopkins, Myers, Shohamy, Grossman & Gluck, 2004).

Strategy Analyses

Gluck et al. (2002) identify three strategies that people use to solve the WP task: a multi-cue strategy that uses all cue-outcome relations, a single-cue strategy that uses just one

cue, and a singleton strategy that only uses a cue when it is presented on its own. Despite the suboptimality of the latter two strategies, they still lead to above chance performance. On the basis of model fits to their empirical data Gluck et al. (2002) make two claims: that the majority of participants adopt a singleton strategy, and that there is a gradual shift in strategy from singleton to multi-cue. The empirical data and analyses from our experiments, however, cast doubt on both of these claims.

First, in Gluck et al.'s strategy analysis only one multi-cue strategy was considered (multi-max). This is a very exacting model because it requires that people know the probabilities associated with all 14 patterns, and always choose the most probable outcome given the pattern. We introduced an alternative strategy -- multi-match – which supposes that people distribute their predictions in line with the learned probabilities. We found that across all 3 experiments this multi-match strategy fit over 80% of participants. So, in terms of a strategy-based analysis, our findings support multi-cue rather than single cue strategies.[4]

Second, there is an alternative account of learning that makes no appeal to strategies. Apparent strategy use and strategy shifts might both arise from the operation of a general learning mechanism that incrementally learns all cue-outcome associations. The key idea is that as people gradually learned these relations they will look as if they initially used sub-optimal strategies, and only later shifted to multi-cue strategies. But this pattern could equally result from a tendency to learn about strong cues first, and only learn about weaker cues later in the task. This is compounded by the fact that early in the task the experienced cue-outcome contingencies may be unrepresentative of the actual contingencies, especially for the weaker cues.

This alternative explanation is supported by the increasing positive correlations between implicit and ideal weights as the task progresses (see Table 2), as well as by the general increase in task knowledge, and their correlations with the implicit measures. Across

the task participants learned to distinguish between all four cues, and knowingly relied more on the strong cues than the weak ones.

In addition, the claim that the majority of people use a singleton strategy is independently implausible. It is unlikely that people can learn each individual cue-outcome relation, but cannot transfer any of this information to trials where more than one cue is present. For example, once a participant has learned that card 1 predicts fine, and that card 2 predicts fine, it is unlikely that when faced with a pattern consisting of both card 1 and card 2 they would guess the outcome randomly rather than predict fine weather.

Our alternative account, in terms of a cue-based learning model, looks very similar to the claim that people use a multi-cue strategy (which is pattern-based). In fact the current WP task is unable to discriminate between the two models, because both yield the same behavior in a linear task environment. However, in a non-linear environment (with configural patterns of cues) the two can be distinguished: multi-cue strategies would succeed whereas simple cue-based models would fail. One possibility here is that learners adapt their approach according to the environment. As a default they use a cue-based approach, but if this fails to work they switch to a pattern-based approach. Indeed there is some evidence in support of this suggestion (Olsson, Enqvist, & Juslin, 2005; Juslin, Karlsson, & Olsson, 2005). They argue for an adaptive division of labour between additive (cue-based) and multiplicative (pattern-based) processing depending on the structure of the task environment.

In short, the linear nature of the current task does not allow us to rule between a multi-cue strategy (pattern-based) and a cue-based model. This is an intriguing question for future research. However, our findings do appear to undermine the claim that people use singleton or one-cue strategies.

Overshooting and Maximizing

Another benefit of the regression analyses is that they allow us to compare an individual's actual learning profile with that of an ideal learner exposed to the same data. The main finding from the correlational analyses was that people's implicit cue weights tracked those of an ideal learner across the course of the task. However, the analysis also showed that these implicit weights tended to overshoot the ideal weights as the task progresses (see Figure 6 & 13, lower panels).

One explanation for this is that towards the end of the task people use a maximizing rather than a matching choice rule (Friedman & Massaro, 1998; Nosofsky & Zaki, 1998). This amounts to them choosing the most probable outcome (relative to the cue weights) on every occasion, rather than matching their choices to the outcome probabilities. This would account for the observed overshooting, because a maximizer will look (in terms of their revealed weights) as if they place excessive weights on the cues. This is vividly demonstrated if we construct a learning profile for an ideal maximizer who has pre-knowledge of the objective probabilities for each cue pattern. Their overall regression weights for each cue (computed across all 200 trials) are [-38, -19, +19, +38], which clearly overshoot the objective weights [-2.1, -0.6, +0.6, +2.1].

So a plausible explanation for the overshooting observed in these experiments is that towards the end of the task participants tend to maximize rather than match; presumably once they are surer about the correct responses. Indeed we predict that if participants were exposed to sufficiently long training their revealed cue weights would tend towards the extreme values of an ideal maximizer. Note that this hypothesized shift from matching to maximizing fits with the gradual shift from multi-match to multi-max observed in the strategy analyses (see Figures 8 & 14). Again we predict that in the long run the majority of participants would be fit by the multi-max strategy.

Relevance to Probabilistic Category Learning

Although we have discussed the WP task from the perspective of multi-cue judgment, it can also be considered as a category learning task in which cues are integrated to determine which category (rain or fine) they are diagnostic of. Numerous categorization studies employ stimuli characterized as sets of features, with participants inferring category membership from those features. The only substantive difference is that in multi-cue judgment tasks the cues are not explicitly described as being features of an object but instead are presented as predictors of the outcome (Juslin, Olsson, & Olsson, 2003). Bearing in mind this close resemblance, an important question for future research will be to use online measures of insight – such as the trial-by-trial cue utilization measure employed here in Experiment 2 – to evaluate awareness in categorization. So long as the cues are separable or discrete, such measures should be straightforward to obtain. It has been argued, largely on the basis of neuropsychological data, that category learning is procedural (Ashby & Maddox, 2005; Knowlton & Squire, 1993). The present findings, of course, suggest that this conclusion might be premature (see also Kinder & Shanks, 2001; Nosofsky & Zaki, 1998). Direct evidence concerning awareness and insight in such tasks is very limited, and hence application of the present methods to categorization tasks would be illuminating.

Relevance to Neuropsychological Research

Probabilistic category learning is a fundamental cognitive ability, and one that can be impaired through brain dysfunction. Consequently, tasks like the WP are often used to study the cognitive deficits that arise in Parkinson's disease (Knowlton et al., 1996), amnesia (Knowlton, Squire & Gluck, 1994; Hopkins et al., 2004), and Huntington's disease (Knowlton et al., 1996). This widespread usage heightens the need to develop a clear understanding of how people learn these tasks. Our current findings question several standard assumptions in the literature: that such tasks engage an implicit procedural learning

system (inaccessible to awareness); that people lack insight when learning; and that they adopt suboptimal strategies.

For example, the standard account for why Parkinson's disease patients perform poorly on probabilistic learning tasks is that damage to the basal ganglia impairs their procedural learning system (Knowlton et al., 1996). This kind of explanation, however, is open to question. It seems, in normal participants at least, that multi-cue learning takes place with full awareness and insight on the part of the participant. In terms of the standard dichotomy this implicates an explicit learning system. This problem might be circumvented by supposing that Parkinson's patients tackle the task in a very different way from normal participants. However, a simpler explanation is that they suffer from a generalized learning decrement (cf. Nosofsky & Zaki, 1998).

More recently Shohamy, Myers, Onlaor, and Gluck (2004) have claimed that Parkinson's patients under-perform on the WP task because they are restricted to using single cue strategies. In their study Parkinson's patients and controls completed the same WP task three times on consecutive days. Relative to the controls the Parkinson's patients were impaired on the task, although their learning was significantly above chance, and improved gradually throughout the three days. Indeed their overall performance at the end of day 3 closely resembled the overall performance of controls at the end of day 1.

To account for this pattern of results Shohamy et al. (2004) argued that damage to the basal ganglia restricts patients to singleton or single-cue strategies rather than an optimal multi-cue strategy. Such strategies enable above chance performance, but place limits on the maximum achievable performance. Their main evidence for this claim was that even by day 3 the majority of patients were best fit by a singleton strategy, whereas the majority of controls, after the same amount of training, were best fit by a multi-cue strategy.

However, the experimental data are equally well (perhaps better) explained in terms of a generalized learning deficit. After all, learning in Parkinson's patients is known to be slower on a variety of cognitive tasks. And a general decrement in learning would explain their reduced performance relative to controls, as well as the fact that the strategy model fits for Parkinson's patients on day 3 were identical to the model fits for the controls on day 1. The apparent shift from singleton to multi-cue strategies is again explicable by the gradual learning of each cue-outcome association. As noted above, the early stages of learning may be well fit by a singleton model, but this could reflect imperfect multi-cue learning rather than a singleton strategy. Preliminary support for this account comes from a Parkinson's patient study that shows a high correlation between learning performance and explicit awareness, with the majority of patients best fit by the multi-match strategy (Jahanshahi, Wilkinson, Gahir, Dharmaindra & Lagnado, submitted).

On this alternative account, then, the degradation in performance seen in Parkinson's patients results from a general drop in learning rate, not a specific inability to engage in multi-cue learning.[6] Of course more research is needed to decide definitively between these alternatives, but it is important to acknowledge that a viable, and indeed more parsimonious, alternative exists. A natural extension would be to test Parkinson's and other patients using the online measures of cue usage introduced in the current study.

Conclusions

In this paper we have shown that multi-cue probability learning is accompanied by explicit task knowledge and self-insight. Although this accords well with commonsense, it is a position that has been questioned in recent research. In addition we have argued that individual learning profiles are well explained in terms of a general learning mechanism that integrates multiple cue-outcome contingencies. A dynamic rolling regression model gives a good qualitative fit to the human data, and thus provides us with a convincing as-if model to

describe human behaviour. There are a variety of psychologically plausible mechanisms that could instantiate this computational level description (e.g., connectionist networks; Stone, 1986). Finally, we have conjectured that a single declarative learning system is sufficient to explain people's online predictions, their probability judgments, and their explicit reports about their own cue usage. In short, a successful learner is not only sensitive to the statistical structure of the environment, but is sensitive to this sensitivity.

References

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5,* 204–210.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149-178.

Ashby, F. G., Noble, S., Filoteo, J. V., Waldron, E. M., & Ell, S. W. (2003). Category learning deficits in Parkinson's disease. *Neuropsychology, 17,* 115–124.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275,* 1293 - 1295.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50,* 255–72.

Brehmer, B. (1979). Preliminaries to a psychology of inference. *Scandinavian Journal of Psychology, 20,* 193-210.

Brehmer, B. (1980). In one word: not from experience. *Acta Psychologica, 45,* 223–241.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Cooksey, R. (1996). *Judgment analysis: theory, methods, and applications.* Academic Press, London.

Dayan, P., Kakade, S., & Montague, P.R. (2000). Learning and selective attention. *Nature Neuroscience, 3,* 1218-1223.

Doherty, M. E., & Kurz, E. (1996). Social judgment theory. *Thinking and Reasoning, 2,* 109-140.

Doherty, E.D. & Balzer, W.K. (1988). Cognitive feedback. In B. Brehmer & C.R.B. Joyce (Eds.), *Human Judgment: The SJT View.* New York: North-Holland, 115-162.

Ericsson, K. and Simon, H. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251.

Evans, J. St. B. T., Clibbens, J., Cattani, A., Harris, A., & Dennis, I. (2003). Explicit and implicit processes in multicue judgment. *Memory & Cognition, 31,* 608-618.

Friedman, D., Massaro, D. W., Kitzis, S. N., & Cohen, M. M. (1995). A comparison of learning models. *Journal of Mathematical Psychology, 39,* 164-178.

Friedman, D. & Massaro D.W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review, 5,* 370–389.

Gavanski, I., & Hoffman, C. (1987). Awareness of influences on ones own judgments: The roles of covariation detection and attention to the judgment process. *Journal of Personality and Social Psychology, 52,* 453-463.

Gigerenzer, G., Todd, P.M., & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Gluck, M., Shohamy, D., & Myers, C. (2002). How do people solve the "Weather Prediction" task? Individual variability in strategies for probabilistic category learning. *Learning & Memory, 9,* 408 – 418.

Goldstein, W. M. (2004). Social Judgment Theory: Applying And Extending Brunswik's Probabilistic Functionalism. In D. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making.* Blackwell.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255-262.

Harries, C. & Harvey, N. (2000). Taking advice, using information and knowing what you are doing. *Acta Psychologica, 104,* 399–416.

Harries, C., Evans, J.St.B.T. and Dennis, I. (2000) Measuring Doctors' Self-Insight into their Treatment Decisions. *Applied Cognitive Psychology, 14,* 455-477.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world.* Thousand Oaks, CA: Sage.

Hertwig, R. & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24,* 383–451.

Hopkins, R. O., Myers, C. E., Shohamy, D., Grossman, S., & Gluck, M. (2004). Impaired probabilistic category learning in hypoxic subjects with hippocampal damage. *Neuropsychologia, 42,* 524–535.

Jahanshahi, M., Wilkinson, L., Gahir, H., Dharmaindra, A. & Lagnado, D. A. (submitted). The effects of levodopa on probabilistic category learning in Parkinson's disease patients.

Juslin, P., Karlsson, L., & Olsson, H. (2005). Additive integration of information in multiple-cue judgment: A division of labor hypothesis. Manuscript submitted for publication.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 133-156.

Kelley, H., & Friedman, D. (2002). Learning to forecast price. *Economic Enquiry, 40,* 556 – 573.

Kinder, A., & Shanks, D. R. (2001). Amnesia and the declarative/nondeclarative distinction: A recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience, 13*, 648-669.

Kitzis S., Kelley, H., Berg, D., Massaro, D., & Friedman, D. (1998). Broadening the tests of learning models. *Journal of Mathematical Psychology, 42,* 327-55.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Human Judgment: The SJT view* (pp. 115-160). North-Holland: Elsevier.

Knowlton, B., Mangels, J., & Squire, L. (1996). A neostriatal habit learning system in humans. *Science, 273,* 1399 – 1402.

Knowlton, B., Squire, L.R., Paulsen, J.S., Swerdlow, N., Swenson, M., & Butters, N. (1996). Dissociations within nondeclarative memory in Huntington's disease. *Neuropsychology*, 10(4), 1-11.

Knowlton, B., Squire, L. and Gluck, M. (1994). Probabilistic classification leaning in amnesia. *Learning & Memory, 1,* 106 – 120.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, 262, 1747-1749.

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes, 28,* 3–26.

Nisbett, R. and Wilson, T., 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84,* 231–259.

Olsson, A.-C., Enqvist, T., & Juslin, P. (2005). Non-linear multiple-cue judgment tasks. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), Proceedings of the 27th Annual Conference of the Cognitive Science Society (pp. 1672-1677). Mahwah, NJ: Lawrence Erlbaum.

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

Rappoport, L., & Summers, D. A. (eds.) (1973). *Human judgment and social interaction.* New York: Holt, Rinehart and Winston.

Reber, P. F., Knowlton, B., & Squire, L. R. (1996). Dissociable properties of memory systems: Differences in the flexibility of declarative and non-declarative knowledge. *Behavioral Neuroscience, 110,* 861–871.

Reber, P. F., & Squire, L. R. (1999). Intact learning of artificial grammars and intact category learning by patients with Parkinson's disease. *Behavioral Neuroscience, 113,* 235–242.

Reilly, B., & Doherty, M. (1992). The assessment of self-insight judgment policies. *Organizational Behavior and Human Performance, 53,* 285–309.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42A,* 209–237.

Shanks, D. R., & St John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367–447.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15,* 233-250.

Shohamy, D., Myers, C. E., Onlaor, S., & Gluck, M. A. (2004). Role of the Basal Ganglia in category learning: How do patients with Parkinson's disease learn? *Behavioral Neuroscience, 118, 4,* 676–686.

Shohamy, D., Myers, C. E., Grossman, S., Sage, J., Gluck, M. A., & Poldrack, R. A. (2004). Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. *Brain, 127,* 851-859.

Slovic, P. and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649–744.

Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A. M., Suzuki, W. A., Brown, E. N. (2004). Dynamic Analysis of Learning in Behavioral Experiments. *Journal of Neuroscience, 24,* 447-461.

Squire, L. R. (1994). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. In D. L. Schacter & E. Tulving (Eds.), *Memory systems* (pp. 203–231). Cambridge, MA: MIT Press.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart, J.L. McClelland & the PDP Research Group (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition* (pp. 444–459), 1. Cambridge, MA: MIT Press.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology, 71,* 680-683.

West R. F., & Stanovich K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition, 31,* 243-251.

Wigton, R.S. (1996). Social judgment theory and medical judgment. *Thinking and Reasoning, 2,* 175-190.

Wilkinson L., Lagnado D.A., Quallo M., Jahanshahi M. (submitted). The effect of corrective feedback on non-motor probabilistic classification learning in Parkinson's disease.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious.* Harvard University Press.

York, K. M, Doherty, M. E., & Kamouri, J. (1987). The influence of cue unreliability on judgment in a multiple cue probability learning task. *Organizational Behavior and Human Performance, 39,* 303 – 317.

Zaki, S. R. (2005). Is categorization performance really intact in amnesia? A meta-analysis. *Psychonomic Bulletin & Review,* 11, 1048-1054.

Appendix

Strategy Analysis (adapted from Gluck et al., 2002)

We followed the strategy analysis procedure introduced by Gluck et al. (2002). We investigated four strategies: *multi-max*, which uses all four cues and selects the most probable outcome on each occasion; *multi-match*, which also uses all four cues, but selects an outcome in proportion to its probability; *one-cue*, which selects on the basis of just one cue; and *singleton*, which just learns patterns that contain a single cue.

For each individual participant we constructed ideal response sets for each of the four strategies, defined as the expected pattern of responses if the participant was reliably following that strategy. This was compared with the participant's actual responses by computing a model score:

$$\textit{Score for model } M = \sum(\#rain\_expected_{p,m} - \#rain\_actual_p)^2 / \sum(\#present_p)^2$$

where P = pattern $A \ldots N$; $\#present_p$ is the number of times pattern P appears in the trial set, $\#rain\_expected_{p,m}$ is the number of rain responses expected to pattern P under model M, and $\#rain\_actual_p$ is the actual number of rain responses the participant made in the trial set. The resultant score was a number between 0 and 1, with 0 indicating a perfect fit between model M and the participant's actual responses. The best fitting model (lowest score) was selected for each participant. These were computed over the whole 200 trials, and separately for each block of 50 trials (see Figures 8 & 14).

Footnotes

1. A relevant exception is a study conducted by Reber, Knowlton & Squire (1996) with a 50 trial version of the WP task. They used explicit measures of task knowledge that were very similar to the probabilistic questions used in our experiments (although only administered at the end of the task, not during the task as in our studies). Interestingly, they found that normal controls showed high levels of learning and awareness. This contrasts with the general consensus that the WP is an implicit task, but coheres with the findings in our studies.
2. The window size of 50 was not based on prior model fits, but was the minimum size required to compute stable logistic regression coefficients. Ongoing work uses an online regression model that is not so constrained. However, the essential point is that a moving window model appears to capture the dynamics of individual learning.
3. A considerable body of research in social cognition has studied the extent to which individuals have privileged, conscious, access to the variables that affect their own judgments (e.g., of how much they would like a person). Although it is commonly argued that people often lack insight and are no more accurate than others who observe their behavior, evidence in fact suggests the opposite. Gavanski and Hoffman (1987), for example, found that actors were much more accurate than observers.
4. One problem with Reber et al.'s study is its reliance on the questionable claim that amnesics performed normally in the learning phase. This was based on the fact that amnesics showed no significant difference from controls in terms of % correct predictions. However, they did perform numerically worse than controls, and, as we strive to show in the current article, % correct predictions (averaged across participants and trials) is a crude measure of learning performance (cf. Zaki, 2005).

5. An important difference between our experiments and Gluck et al.'s is that we paid participants in relation to their performance. This is known to improve learning in a variety of judgment and learning tasks (Hertwig & Ortmann, 2001). So another reason why many more of our participants were fit by multi-cue strategies might lie in the increased incentive to solve the task optimally.
6. A recent study by Shohamy et al. (2004) might be interpreted as evidence against this claim. They found that Parkinson's disease patients were impaired on a standard feedback version of a probabilistic learning task, but not on an observational (paired-associate) version. This seems consistent with the claim that the standard version of the WP task involves implicit learning, and this is disrupted in Parkinson's patients. However, the Shohamy et al. study used non-equivalent tests for the two versions of the task. When this was corrected, and a more sensitive within-subject design used, no difference was found between feedback and paired-associate versions (Wilkinson, Lagnado, Quallo & Jahanshahi, submitted).

Author Note

David A. Lagnado, David R. Shanks and Steven Kahan, Department of Psychology, University College London, UK; Ben R. Newell, School of Psychology, University of New South Wales, Australia.

This work was part of the programme of the ESRC Research Centre for Economic Learning and Social Evolution, and was also supported by the Leverhulme Trust/ESRC Research Programme “Evidence, Inference and Enquiry” (David Lagnado), and the Australian Research Council (Ben Newell). We thank A. R. Jonckheere for advice on statistics, Magda Osman for helpful discussions, and Martijn Meeter for comments on an earlier draft.

Correspondence concerning this article should be addressed to David Lagnado, Department of Psychology, University College London, Gower Street, London WC1E 6BT. E-mail: d.lagnado@ucl.ac.uk

Table 1

Learning Environment for Experiments 1, 2 & 3.

Pattern	Cards present	Total	P(pattern)	P(fine|pattern)
A	0001	19	0.095	0.895
B	0010	9	0.045	0.778
C	0011	26	0.13	0.923
D	0100	9	0.045	0.222
E	0101	12	0.06	0.833
F	0110	6	0.03	0.500
G	0111	19	0.095	0.895
H	1000	19	0.095	0.105
I	1001	6	0.03	0.500
J	1010	12	0.06	0.167
K	1011	9	0.045	0.556
L	1100	26	0.13	0.077
M	1101	9	0.045	0.444
N	1110	19	0.095	0.105
Total		200	1.00	

Note: 0 = card present, 1 = card absent.

Table 2

Mean rank correlation coefficients (Spearman r_s) between implicit weights, ideal weights and explicit ratings for Experiments 1, 2 and 3.

		Trial 50	Trial 100	Trial 150	Trial 200	Linear trend *F*
EXP 1	implicit - ideal	0.60**	0.66**	0.74**	0.85**	5.65*
	ideal – explicit	0.52**	0.52**	0.63**	0.75**	4.98*
	implicit -explicit	0.40*	0.58**	0.59**	0.77**	5.49*
EXP 2	implicit - ideal	0.71**	0.71**	0.89**	0.85**	2.25
	ideal – explicit	0.59**	0.69**	0.74**	0.83**	4.85*
	implicit -explicit	0.63**	0.66**	0.71**	0.83**	5.25*
EXP 3	implicit - ideal	0.60**	0.82**	0.78**	0.80**	1.22
	ideal – explicit	-	-	-	0.71**	-
	implicit -explicit	-	-	-	0.63**	-

Note: ** = significant at $p < 0.01$. * = significant at $p < 0.05$.

The final column shows the *F* values for linear trend tests for the preceding four mean correlations.

Figure Captions

Figure 1. Probabilistic environment in the Weather Prediction task for Experiments 1, 2 & 3.

Figure 2. Screen presented to participants in the learning phase of Experiments 1, 2 & 3.

Figure 3. Learning performance measured by mean proportion correct predictions (±SE) in Experiment 1 (the optimal level is 0.83).

Figure 4. Explicit ratings in Experiment 1. Upper panel: Blocked mean probability ratings (±SE) for rain given each card (objective probabilities for card 1 = 20%, card 2 = 40%, card 3 = 60%, card 4 = 80%). Lower panel: Blocked importance ratings (±SE) for strong and weak cards.

Figure 5. Implicit regression weights for each card for best performer (upper panel) and worst performer (lower panel) in Experiment 1.

Figure 6. Group analyses in Experiment 1. Upper panel: Mean implicit regression weights for each card. Lower panel: Mean implicit and ideal regression weights for cards 1 and 4.

Figure 7. Mean implicit regression weights (absolute values) for the strong cards (1&4) and weak cards (2&3) in Experiment 1.

Figure 8. Strategy analysis by block for Experiment 1.

Figure 9. Learning performance measured by mean proportion correct predictions (±SE) in Experiment 2.

Figure 10. Explicit ratings in Experiment 2. Upper panel: blocked mean probability ratings (±SE) for rain given each card. Lower panel: blocked importance ratings (±SE) for strong and weak cards.

Figure 11. Upper panel: Mean trial-by-trial explicit cue usage ratings for strong cards (1&4) and weak cards (2&3) in Experiment 2. Lower panel: Mean implicit regression weights (absolute values) for strong and weak cards.

Figure 12. Implicit regression weights for each card for best performer (upper panel) and worst performer (lower panel) in Experiment 2.

Figure 13. Group analyses in Experiment 2. Upper panel: Mean implicit regression weights for each card. Lower panel: Mean implicit and ideal regression weights for cards 1 and 4.

Figure 14. Strategy analysis by block for Experiment 2 (upper panel) and Experiment 3 (lower panel).

Figure 15. Two possible models of the relation between learning and behavioural responses. A: single process model; B: dual-process model (adapted from Lovibond and Shanks, 2002).

Figure 1.

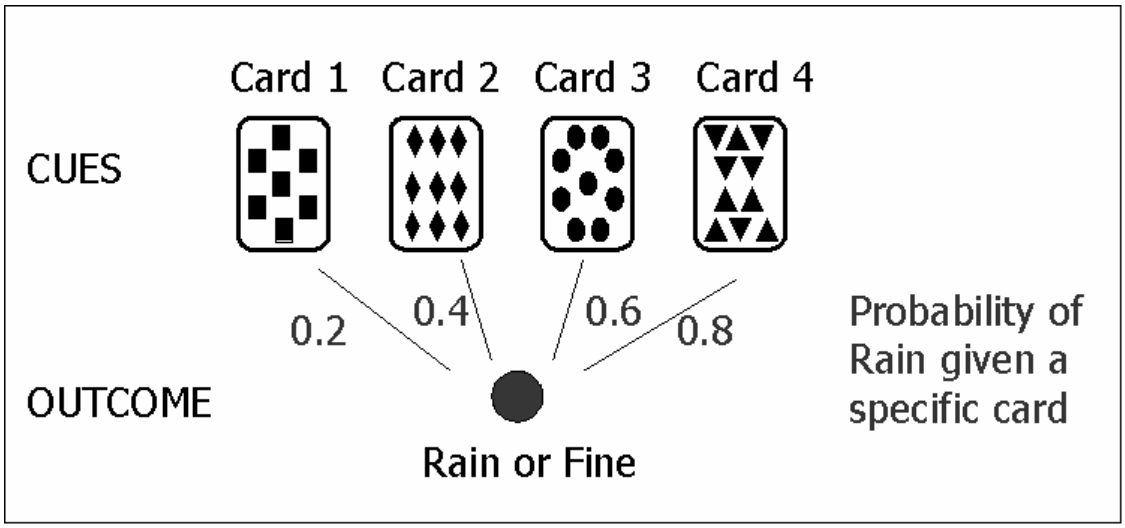

Card 1
Card 2
Card 3
Card 4
CUES
0.2
0.4
0.6
0.8
Probability of Rain given a specific card
OUTCOME
Rain or Fine

Figure 2.

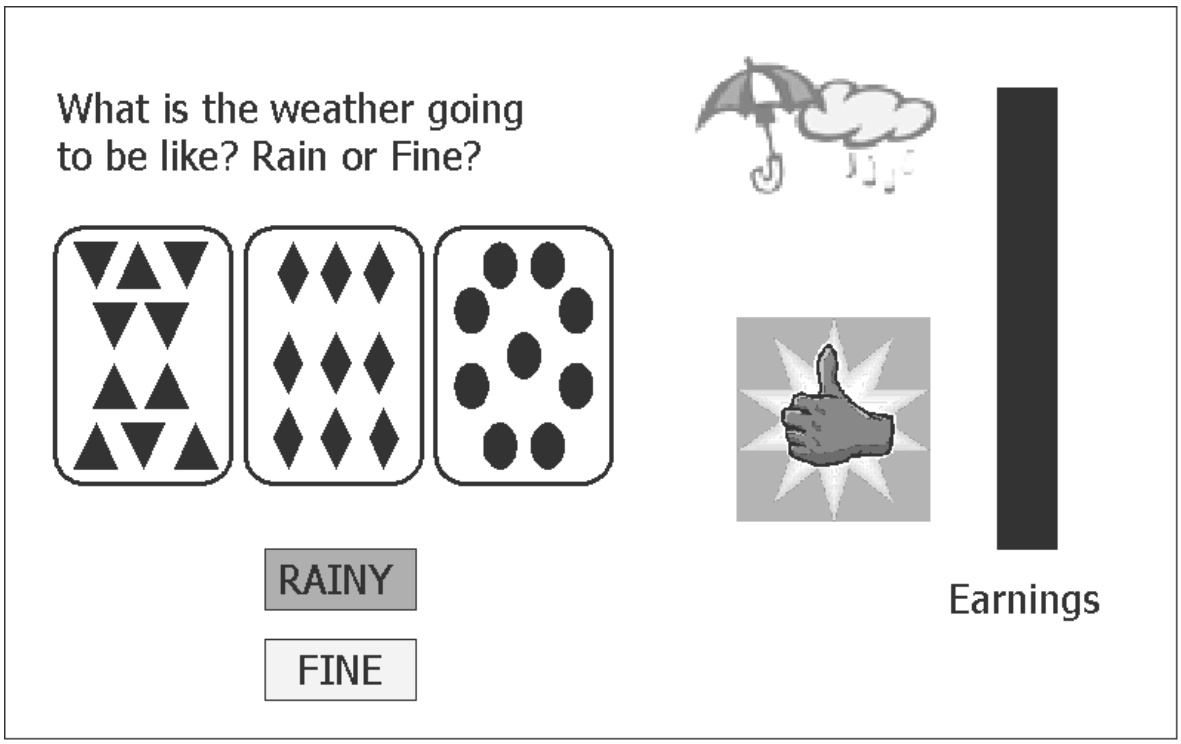

What is the weather going to be like? Rain or Fine?
RAINY
FINE
Earnings

Figure 3.

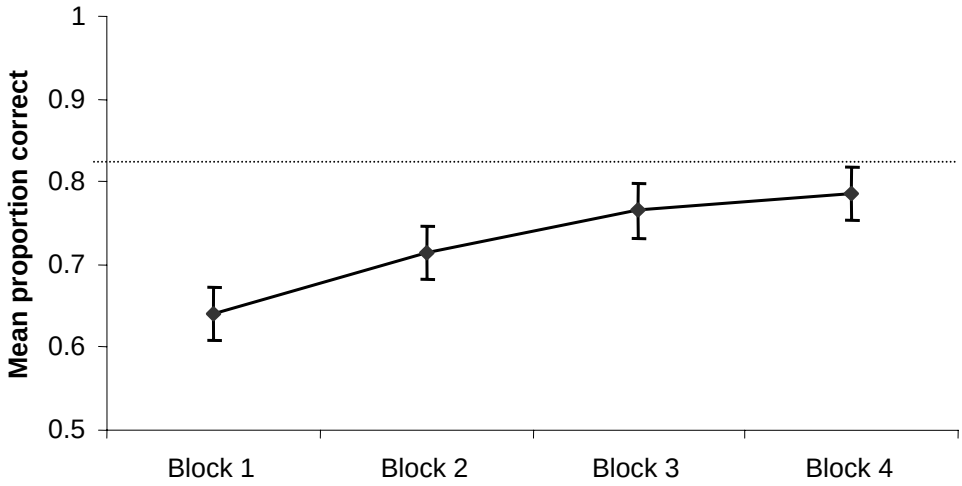
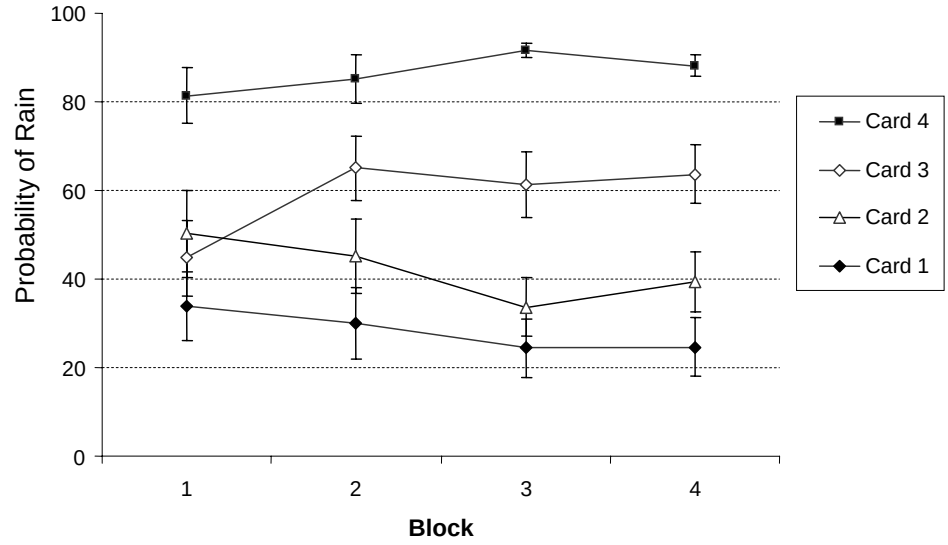

Mean proportion correct
1
0.9
0.8
0.7
0.6
0.5
Block 1
Block 2
Block 3
Block 4

Figure 4.


100
80
60
40
20
0
Probability of Rain
1
2
3
4
Block
Card 4
Card 3
Card 2
Card 1

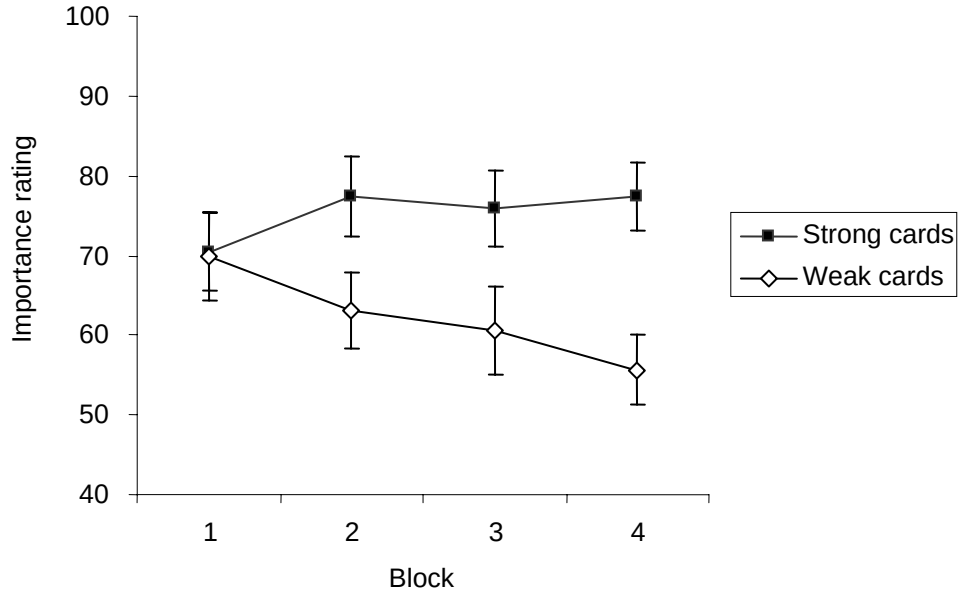
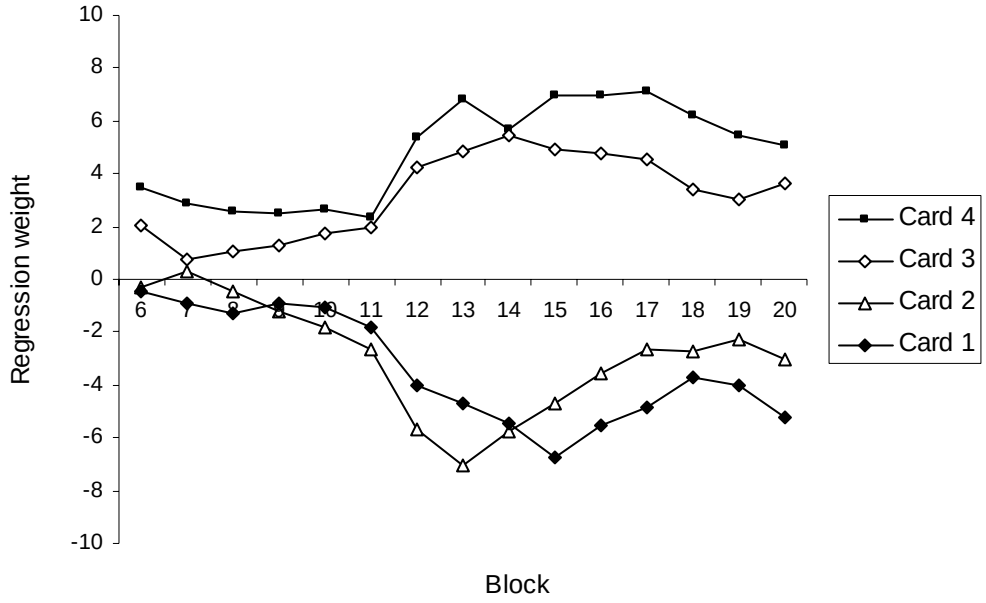

100
90
80
70
60
50
40
Importance rating
1
2
3
4
Block
Strong cards
Weak cards

Figure 5.


Regression weight
10
8
6
4
2
0
-2
-4
-6
-8
-10
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Block
Card 4
Card 3
Card 2
Card 1

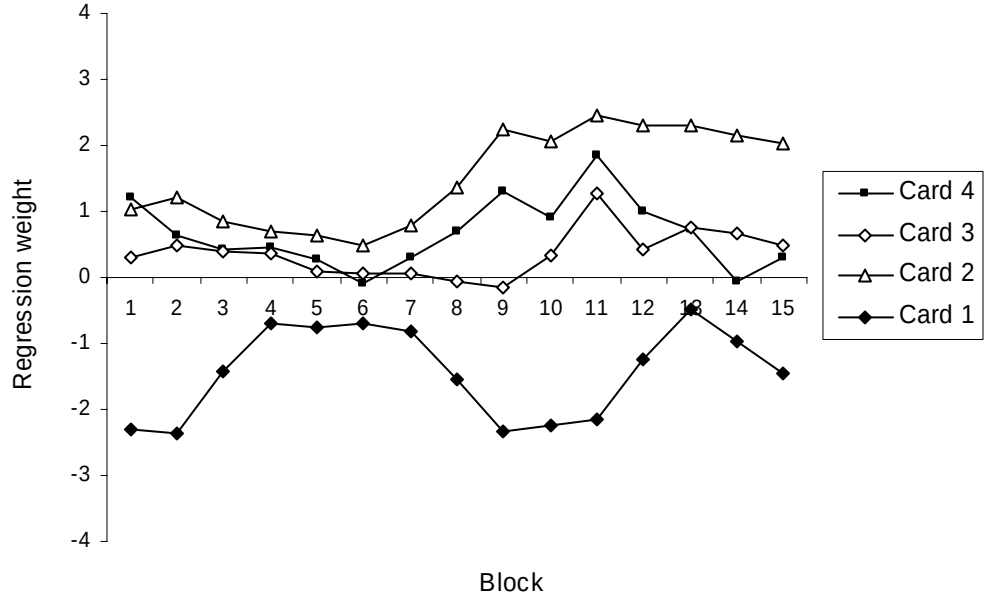
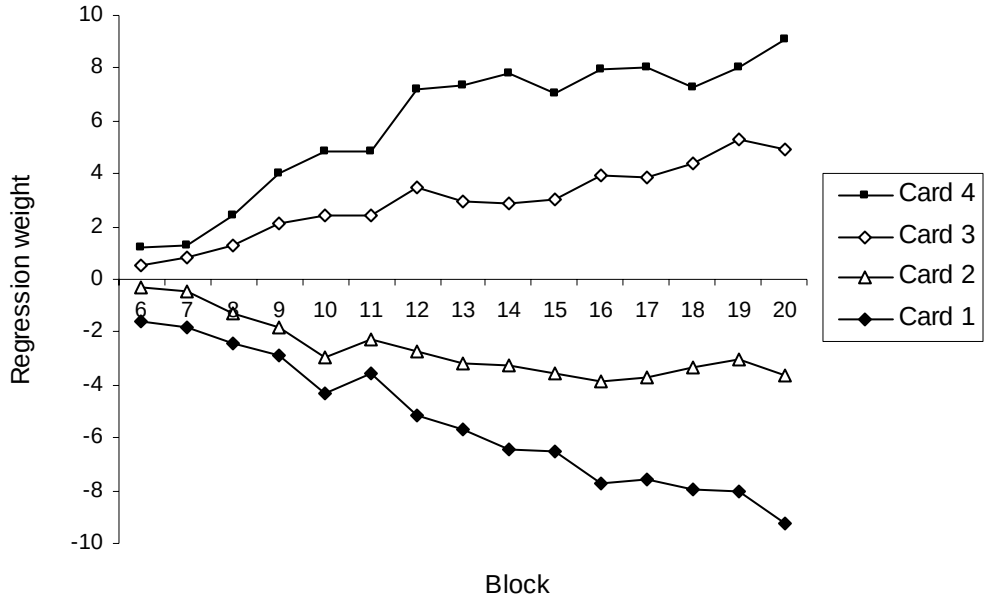

Regression weight
4
3
2
1
0
-1
-2
-3
-4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Block
Card 4
Card 3
Card 2
Card 1

Figure 6.


Regression weight
10
8
6
4
2
0
-2
-4
-6
-8
-10
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Block
Card 4
Card 3
Card 2
Card 1

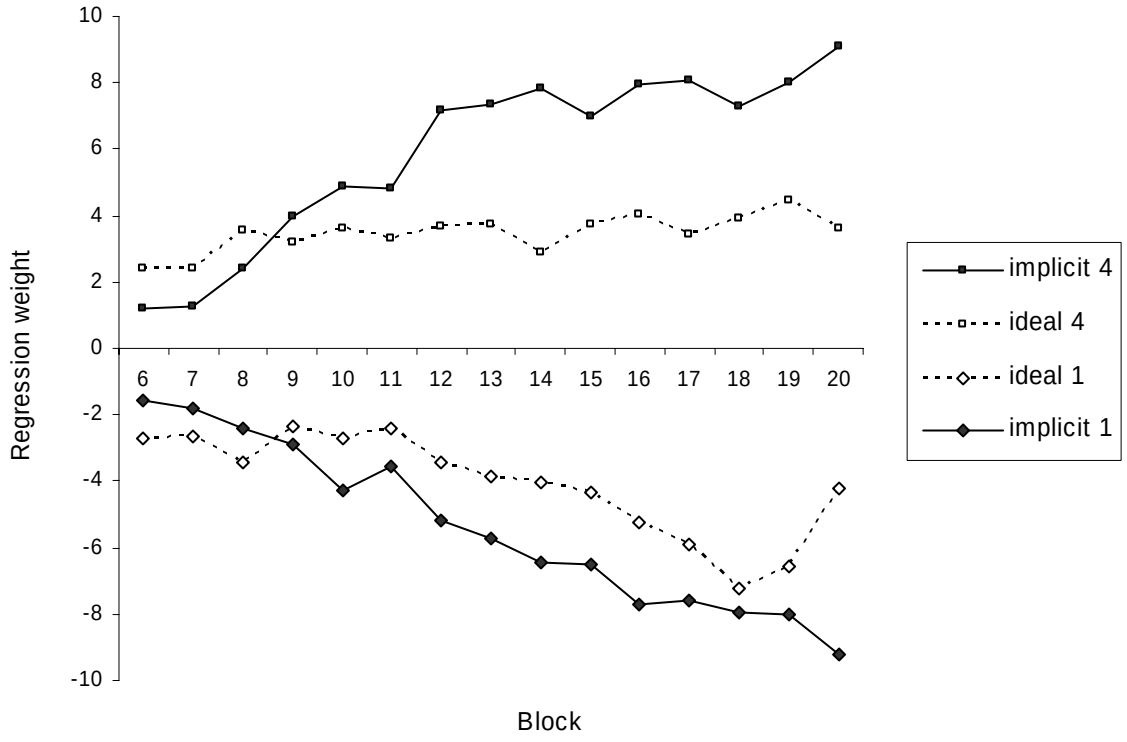

Regression weight
10
8
6
4
2
0
-2
-4
-6
-8
-10
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Block
implicit 4
ideal 4
ideal 1
implicit 1

Figure 7.

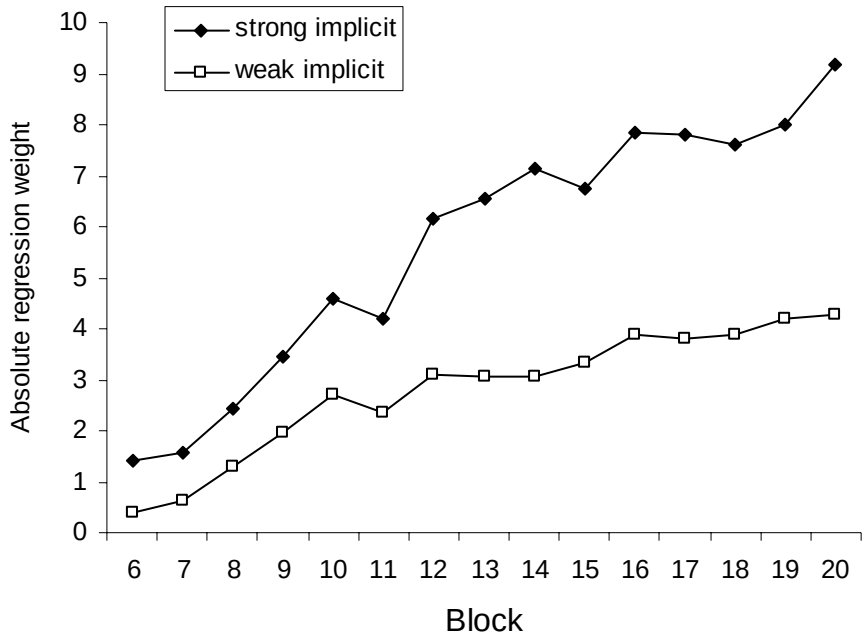

strong implicit
weak implicit
Absolute regression weight
0
1
2
3
4
5
6
7
8
9
10
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Block

Figure 8.

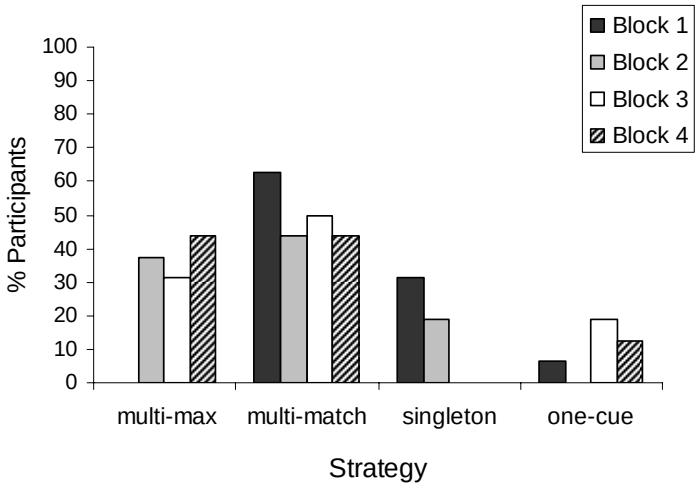

Block 1
Block 2
Block 3
Block 4
100
90
80
70
60
50
40
30
20
10
0
% Participants
multi-max
multi-match
singleton
one-cue
Strategy

Figure 9.

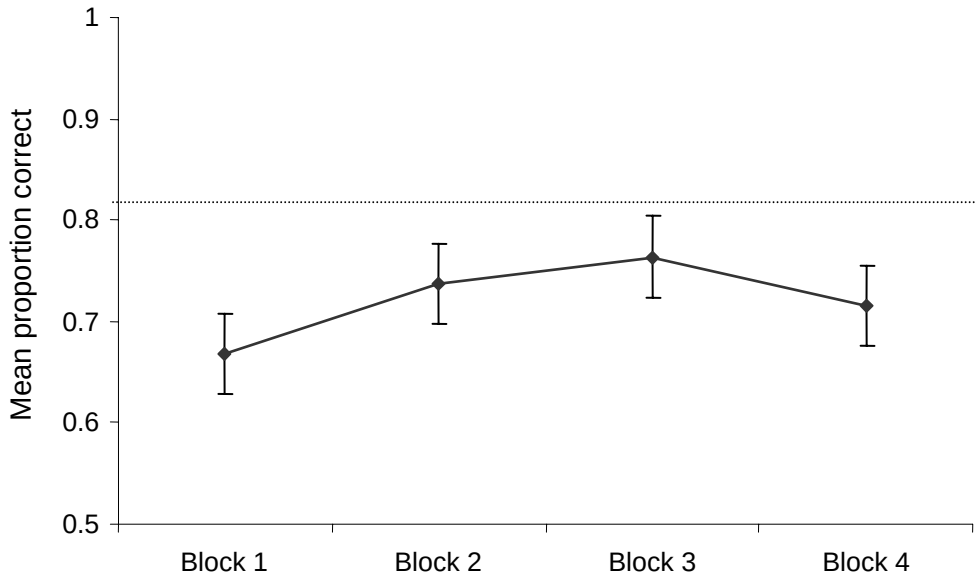
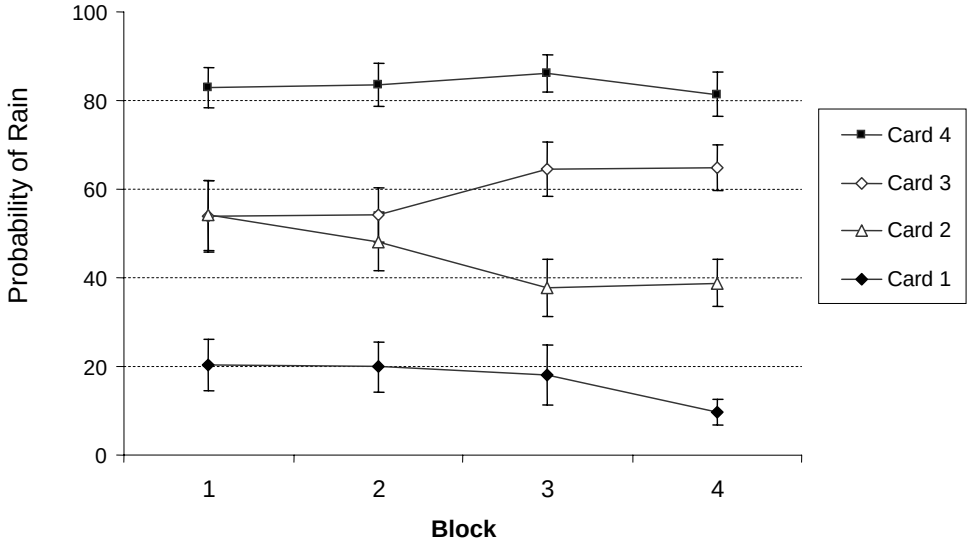

1
0.9
0.8
0.7
0.6
0.5
Mean proportion correct
Block 1
Block 2
Block 3
Block 4

Figure 10.


100
80
60
40
20
0
Probability of Rain
1
2
3
4
Block
Card 4
Card 3
Card 2
Card 1

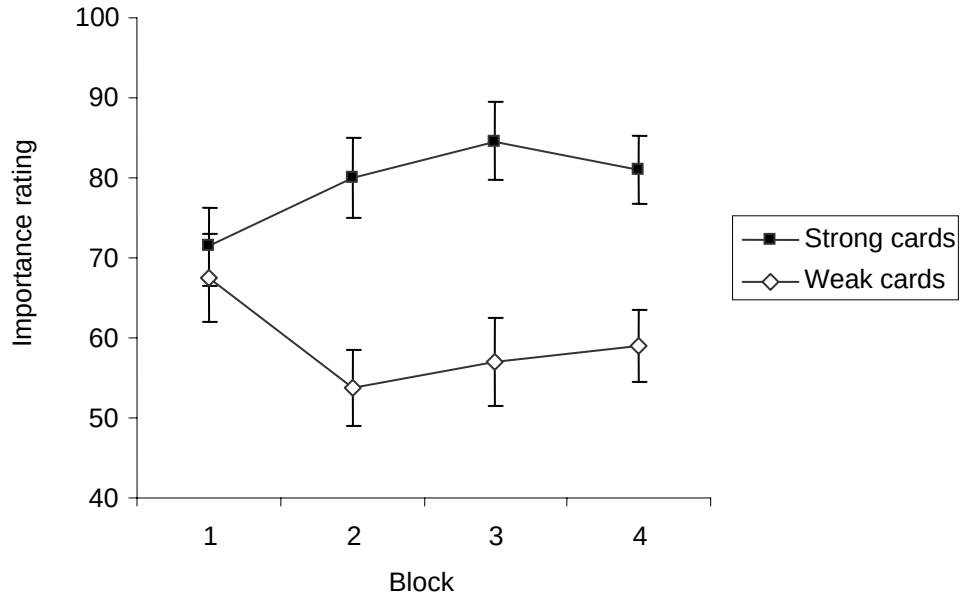
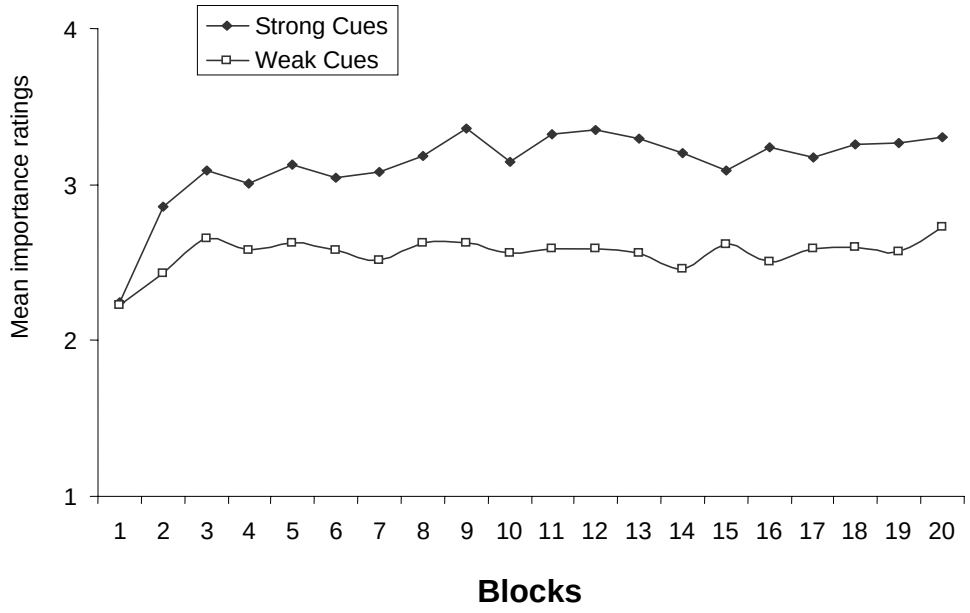

100
90
80
70
60
50
40
Importance rating
1
2
3
4
Block
Strong cards
Weak cards

Figure 11.


Strong Cues
Weak Cues
Mean importance ratings
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Blocks

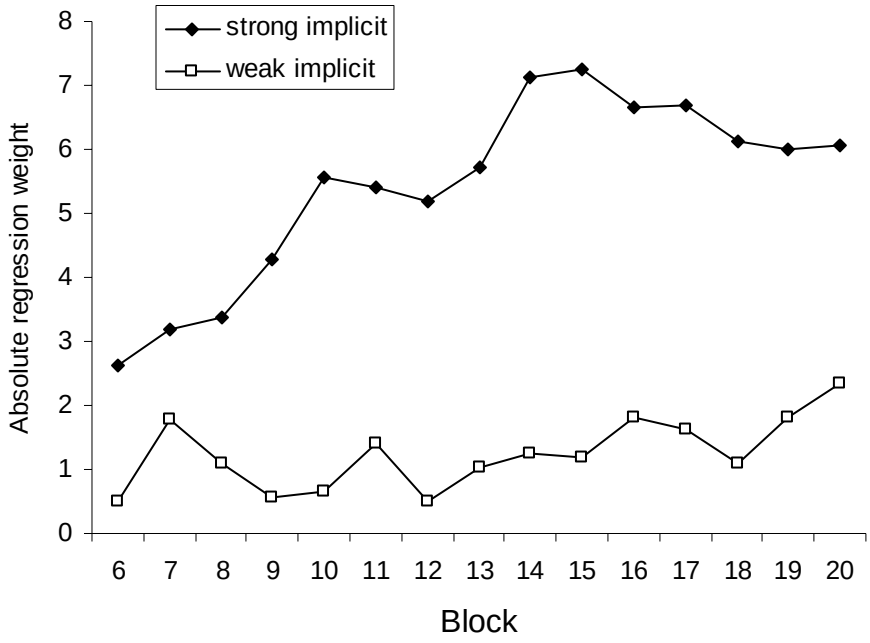

strong implicit
weak implicit
Absolute regression weight
8
7
6
5
4
3
2
1
0
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Block

Figure 12.

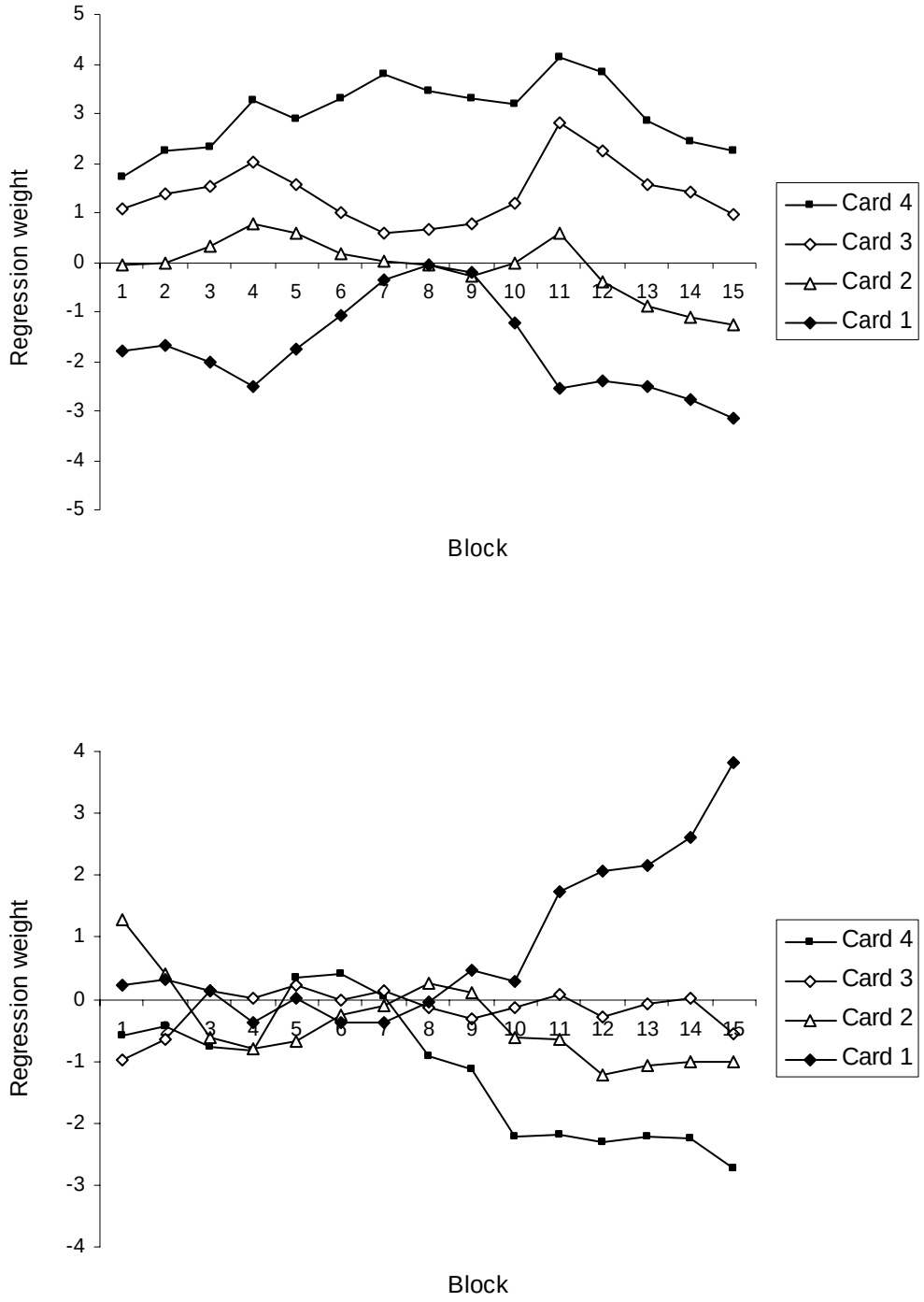
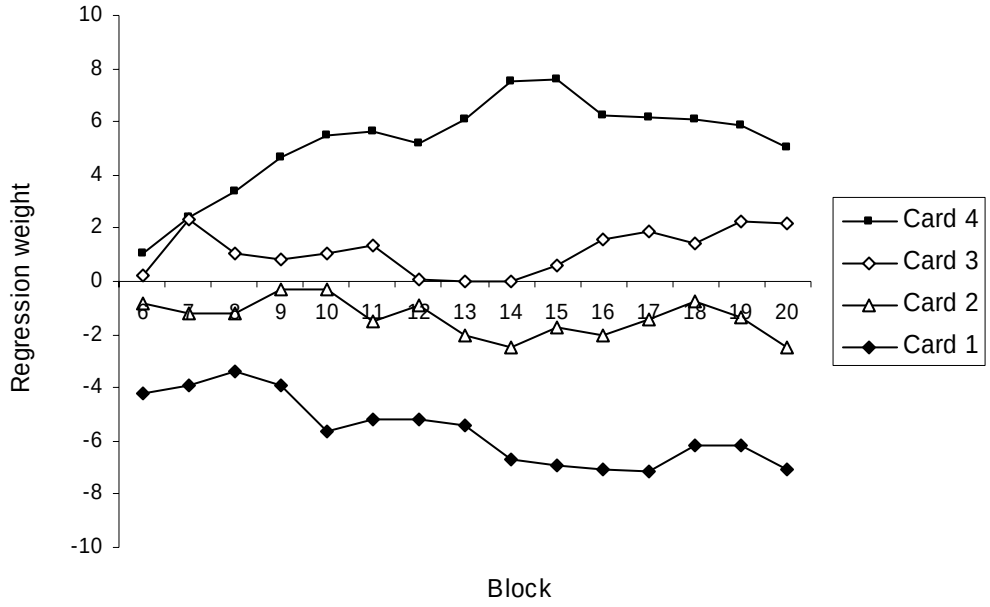


Figure 13.


10
8
6
4
2
0
-2
-4
-6
-8
-10
Regression weight
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Block
Card 4
Card 3
Card 2
Card 1

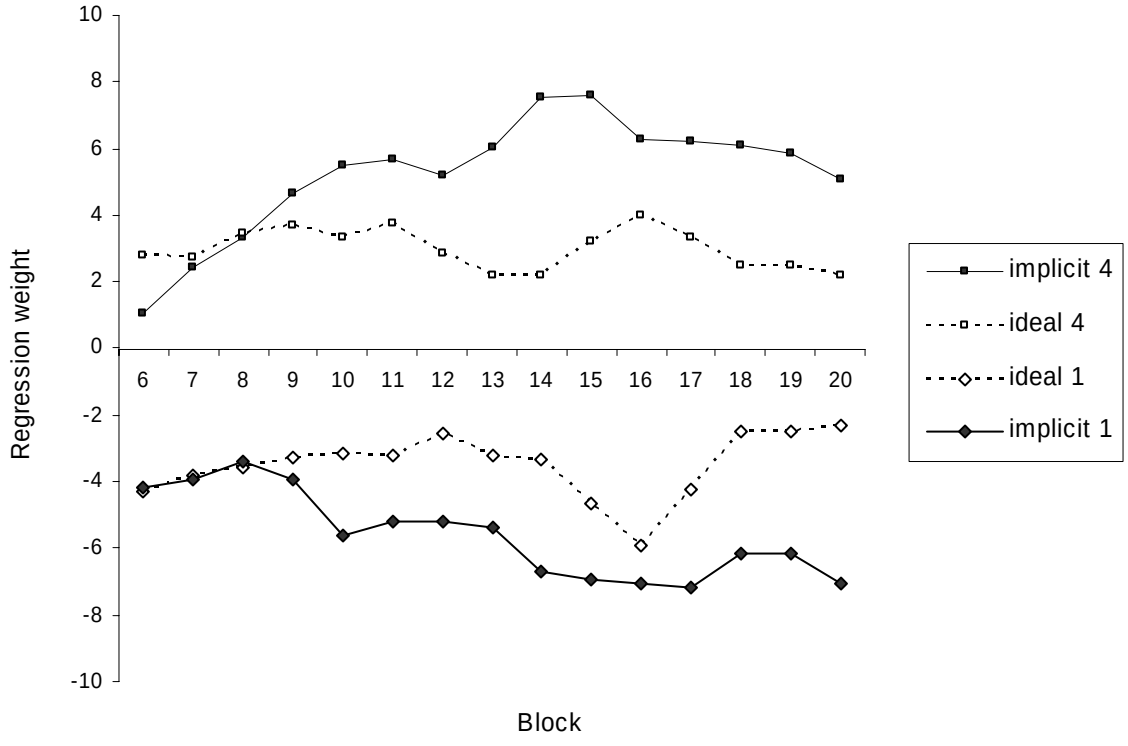
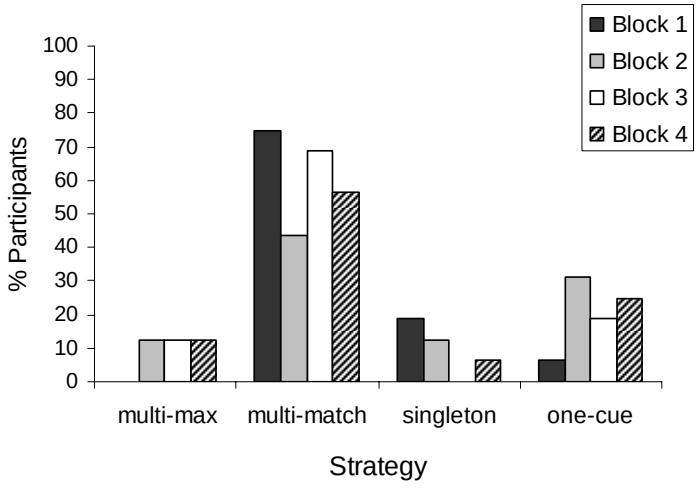

10
8
6
4
2
0
-2
-4
-6
-8
-10
Regression weight
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Block
implicit 4
ideal 4
ideal 1
implicit 1

Figure 14.


Block 1
Block 2
Block 3
Block 4
100
90
80
70
60
50
40
30
20
10
0
% Participants
multi-max
multi-match
singleton
one-cue
Strategy

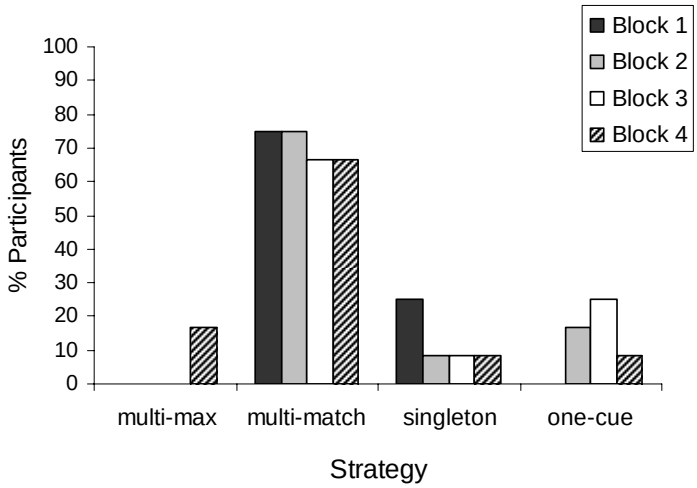
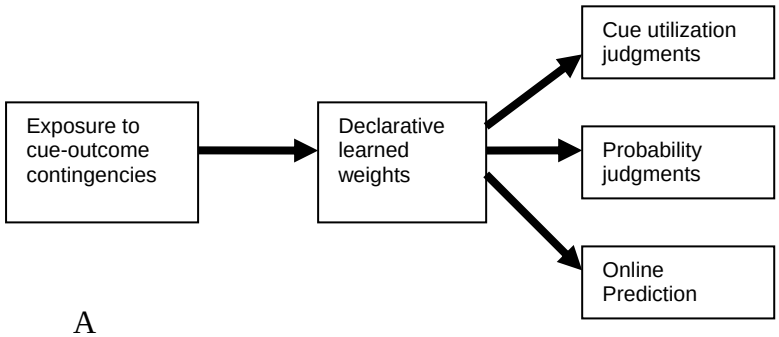
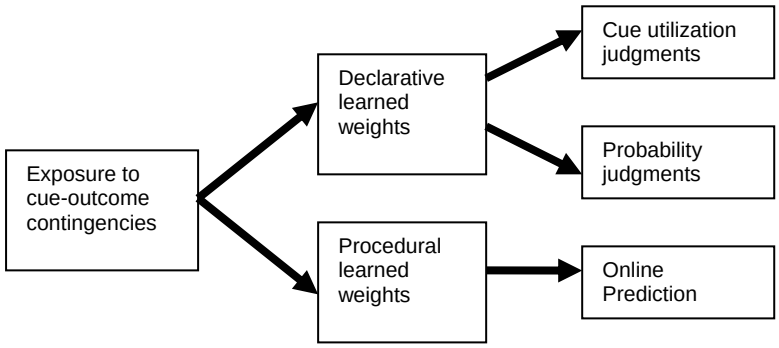

Block 1
Block 2
Block 3
Block 4
100
90
80
70
60
50
40
30
20
10
0
% Participants
multi-max
multi-match
singleton
one-cue
Strategy

Figure 15.


Exposure to cue-outcome contingencies
Declarative learned weights
Cue utilization judgments
Probability judgments
Online Prediction

A


Exposure to cue-outcome contingencies
Declarative learned weights
Procedural learned weights
Cue utilization judgments
Probability judgments
Online Prediction

B