Explicit and implicit processes in multicue judgment

JONATHAN ST. B. T. EVANS, JOHN CLIBBENS, ALLEGRA CATTANI,
ANITA HARRIS, and IAN DENNIS
University of Plymouth, Plymouth, England

In two experiments, a multicue probability learning task was used to train participants in relating judgments to a criterion, on the basis of several cues that could or could not be relevant. The outcome feedback had 25% added noise to simulate real-world experience-based learning. Judgmental strategies acquired were measured by individual multiple linear regression analyses of a test phase (with no feedback) and were compared with self-ratings of cue relevance. In a third experiment, participants were instructed explicitly on cue relevance, with no training phase. The pattern of results suggested that both implicit and explicit cognitive processes influenced judgments and that they may have been sensitive to different task manipulations in the learning phase. On more complex tasks, despite weak explicit learning, explicit processes continued to influence judgments, producing a decrement in performance. These findings explain why studies of expert judgment often show only moderate levels of self-insight, since people have only partial access to the processes determining their judgments.

Expert judgment is often exercised in situations in which multiple pieces of information are available and relevant in varying degrees. For example, a doctor may be attempting to diagnose the likely illness of a patient on the basis of symptoms, medical history, clinical examination, and results of imperfectly diagnostic tests, as well as relevant demographic information, such as gender, age, and occupation. To take another example, a personnel manager might be judging the suitability of a candidate for a position, taking into account work experience, qualifications, interview performance, and results of psychometric tests. These are examples of what are termed *multicue judgments*, where the various dimensions of information are the cues, whose values have particular and distinct instantiations for individual cases.

What makes multicue judgment of psychological interest, as well as of practical importance, is the potential for bias when such judgments are made *intuitively*. If experts lack self-insight into the processes underlying these judgments, they may be unconsciously biased. For example, it may be contrary to law or company policy for the personnel manager to take into account age or gender in his or her decision making. If no explicit algorithm is followed, however, we may have nothing better than self-report from the decision maker to establish which cues were actually used and with what relative weighting. For this reason, a

tradition of research based on *social judgment theory* was developed some years ago to capture the tacit policies that decision makers use when making multicue judgments. Social judgment theory has its origins in the psychology of Egon Brunswick and is rooted in the ecological approach to human cognition—specifically, by relating the way in which the available information is modeled by the environment, on the one hand, and by the judge, on the other. For the history of the theory, the reader is referred to Doherty and Kurz (1996) and, for details of the associated *lens model* methodology, to Cooksey (1996).

Social judgment theory uses multiple linear regression to measure policies that are implicit in the judgments that people make—a technique known as *policy capture*. In such studies, participants—or *judges*—are asked to make wholistic judgments about a series of cases in which information about the same set of multiple cues is available. The implicit policy can then be inferred by performing a multiple linear regression analysis from the cues onto the judgments. The resulting beta weights tell us which cues influenced the judge and to what relative extent. In some studies, an objective criterion of correct performance is also available. In the case of personnel selection decisions, for example, we may be able to assess job performance at a later date. In a full lens model analysis, a second multiple regression is performed from the cues onto the criterion. A simplified (ignoring non-linear and nonadditive aspects) version of the lens model equation is then given as follows:

$$\text{achievement} = \text{knowledge} \times \text{task predictability} \times \text{cognitive control}.$$

Achievement is the correlation between the judgments and the criterion across the different cases considered.

This research was supported by Grant R000223121 from the Economic and Social Research Council of the United Kingdom. We thank Arthur Reber and Michael Doherty for comments on and criticisms of an earlier draft of this article. Correspondence should be addressed to J. St. B. T. Evans, Centre for Thinking and Language, Department of Psychology, University of Plymouth, Plymouth PL4 8AA, England (e-mail: j.evans@plym.ac.uk).

Task predictability is the multiple correlation of the cues with the criterion (or environment), and cognitive control is the multiple correlation of the cues with the judgments. Task knowledge—also known as matching—is the correspondence between the judge's model and the environment.

To see how this works, imagine a judge whose achievement is poor—his judgments do not correspond well to the criterion. Lens model analysis allows us to distinguish three possible explanations. Perhaps the criterion is simply not predictable from the cues identified (task predictability is low). Or perhaps the judge is unreliable, making inconsistent judgments (cognitive control is low). The interesting case is the one in which the task is predictable and the judge is reliable but achievement is still low. This can happen because the judge has an inaccurate model that does not correspond to reality (task knowledge is low). For example, our personnel manager might be biased, consistently preferring young males, when these attributes actually have little or no predictive value for job performance. A further advantage of this methodology is that implicit policies—inferred from the regression analyses—can be compared with explicit policies derived from self-report. For example, the judge can be asked, on completion of the task, to rate the cues that influenced his or her judgment. The degree of correspondence between the two is then taken as a measure of self-insight.

When criterion scores are not available, it is still possible to examine one side of the lens by regressing cues onto judgments. This is known as *judgment analysis* (Stewart, 1988). Judgment analysis studies still allow tacit policies to be captured and compared with explicit ratings. Both the full lens model and judgment analysis treatments have been extensively applied in studies of expert judgment, especially in the medical domain. The results of such studies typically show that tacit policies vary widely among judges, however expert, and also that self-insight can be very limited, with poor correspondence between tacit and explicit policies (Wigton, 1996). However, it has been shown recently that the traditional methodology might underestimate insight, because people are quite good at picking out their own policies from those of others in a recognition test (Harries, Evans, & Dennis, 2000; Reilly & Doherty, 1992). Another curious feature in this literature is that judges tend to identify as relevant cues that they did not use but rarely use cues that they did not name (Harries, Evans, Dennis, & Dean, 1996). There is evidence that judges report the cues to which they attend but are unaware that some of this attended information is not actually used in their judgments (Harries et al., 2000; York, Doherty, & Kamouri, 1987).

The present research was motivated by dual-process theories of reasoning (Evans & Over, 1996; Sloman, 1996; Stanovich, 1999), which posit a clear distinction between pragmatic or associative reasoning processes, on the one hand, and abstract analytical processes, on the

other. Typically, the former are referred to as implicit or preconscious, and the latter as explicit, with clear links to dual-process theories of learning (Berry & Dienes, 1993; Dienes & Perner, 1999; Reber, 1993). Implicit processes are normally defined in such a way as to indicate inaccessibility to self-report, although the value of the implicit/explicit distinction in the learning field has recently been questioned by some authors (Whittlesea, 2002; Whittlesea & Brooks, 1990). However, we believe that a plausible account of the limited self-insight shown by expert judges is that they have learned their judgmental strategies by experience in an implicit manner. To paraphrase the title of a famous article (Nisbett & Wilson, 1977), they cannot tell what they do not know. Our application of dual-process theory to judgmental processes in the present article accords with recent developments in approaches to the heuristics and biases research program (Gilovich & Griffin, 2002; Kahneman & Frederick, 2002).

It is difficult, however, to examine the role of implicit and explicit processes in real experts, because they have already done the relevant learning. In this research, we therefore set out to study judgmental strategies acquired by experience in computer-generated laboratory tasks, in an attempt to simulate the acquisition of expertise. If the strategies underlying multicue judgments are learned by experience, this will most likely have been by making judgments and receiving outcome feedback. However, in the real world, such feedback is unlikely to be noise free. There is an established paradigm that simulates this kind of learning that is known as *multicue probability learning*. In this paradigm, judges are presented with two or more cues that predict some criterion. Cue validities can vary between 0 and 1, according to the degree to which each cue is predictive. When outcome feedback is computed, however, a certain amount of random error, or noise, is first added. The participants in these experiments have to make use of this imperfect outcome feedback to improve their judgmental accuracy.

One of the pioneers of this methodology (Brehmer, 1980) wrote a famous review drawing very pessimistic conclusions about people's ability to learn from experience in both probabilistic and deterministic environments. However, as Klayman (1984) pointed out, many of these early studies had used nonlinear cue functions, whereas results with linear predictors were much more encouraging. This accords with our own reading of the literature: Articles we have examined that provided outcome feedback training with *linear* cues have resulted in significant evidence of learning, even though most added noise to the feedback (Adelman, 1981; Lindell, 1976; Taylor, Hall, Cosier, & Goodwin, 1996; York et al., 1987). The success of outcome feedback continues, however, to be played down in this literature, due to the generally reported superiority of cognitive feedback (Balzer, Doherty, & O'Connor, 1989), which involves giving people explicit information about their performance and/or the normative model of the environment. In the most effec-

tive form, judges can be shown a picture of their own judgment strategy as compared with the ideal strategy. In lens model terminology, this gives them direct feedback of task knowledge, or the matching between their own model and reality.

Although cognitive feedback is clearly more effective than outcome feedback in improving performance on experimental multicue judgment tasks, it has little external validity for our purposes. What we are interested in here is the acquisition of judgmental strategies in a form that simulates the development of expertise. Clearly, noisy outcome feedback is the best that most real-world judges are going to receive from their everyday experience. Of course, experts will also have explicit knowledge derived from lectures, textbooks, and so on in their training. However, if this knowledge forms the main basis for their judgments, it is hard to explain the main findings in the expert judgment literature just cited. One would expect to find relatively similar judgmental policies across experts and good correspondence between inferred and stated policies—the opposite of the typical results.

In the present study, we used the multicue probability learning method to train judgmental strategies. This involved presenting multicue problems for evaluation, with feedback based on a linear model of the predictive cues plus added noise. Our methodology differed from typical studies in this literature, such as those cited above, in two respects. First, we used both relevant and irrelevant cues. The feedback was computed from a linear model in which the weights assigned to relevant cues were either 1 (positive predictor), -1 (negative predictor, used in Experiments 2 and 3), or 0 for irrelevant cues. Hence, relevant cues were maximally predictive in either a positive or a negative direction, and irrelevant cues were entirely nonpredictive. We were interested in discovering people's ability to learn to distinguish relevant from irrelevant cues. To make the task more realistic—and to increase the chances of implicit learning—we wanted a substantial noise component. The outcome included 25% of added noise in all cases (extensive pilot experimentation suggested that this would still permit learning to occur). Noise was incorporated by computing the pure criterion value from a linear model with weights of 1 and 0 for relevant and irrelevant cues, respectively, and then adding a random constant from a Gaussian distribution calculated to reduce task predictability to .87 ($R^2 = .75$). The scores were then normalized to a distribution appropriate to the output scale shown on the screen.

The second difference in our methodology was to use both training trials (80–100 per task) and separate test trials (40). Outcome feedback was presented only on training trials. We measured the acquired strategy by analyzing the test trials only. Our first measure of performance was, hence, not the usual achievement correlation that is computed by correlating judgments with a noisy criterion. Rather, we correlated judgments on the 40 test trials with pure criterion (noise free)—a measure described as the *performance* correlation by Lindell (1976). Obvi-

ously, people cannot learn noise, so this was the best measure of the accuracy of the acquired strategy that we could take. It also had the advantage of providing the still simpler form of the lens model equation:

$$\text{performance} = \text{task knowledge} \times \text{cognitive control.}$$

By measuring the performance correlation, we removed the noise in the environment, making it perfectly predictable. This also allowed us to measure task knowledge, using a statistic known as G . To do this, we performed a multiple linear regression analysis for each separate task given to each participant, regressing all presented cues (relevant or not) onto their judgments. This gave us the multiple correlation coefficient (R^2) that corresponded to cognitive control. We simply divided R^2 into the performance correlation to obtain G scores. In addition, we were able directly to examine the beta weights produced from this analysis to see how well they differentiated relevant from irrelevant cues.

In summary, the purpose of this article is to elucidate the role of implicit and explicit cognitive processes in multicue judgment. The principal means by which we will do this (Experiments 1 and 2) is to look at the relation between three variables: (1) the criterion against which people are trained, (2) the judgments made in the test phase and the policy that is implicit in them, and (3) the explicit account of cue relevance given in a post-task interview. Hence, we will distinguish *performance*, which is the relation between judgments and criterion, *self-insight*, which is the relation between explicit ratings and judgments, and *explicit knowledge*, which is the relation between explicit ratings and the criterion. The last tells us the extent to which the outcome feedback has generated knowledge of the environment that can be expressed explicitly. This is independent of traditional insight measures, which relate what people say to what they do, regardless of the accuracy of their implicit policies.

EXPERIMENT 1

In Experiment 1, we compared performance on a standard multicue judgment task—which we call the *single task*—with a forced-choice task. In the single task, participants were shown the results of several ability tests ranging from *very poor* to *very good* and were asked to predict a criterion of success in a job. Pilot experimentation suggested that people might be biased by prior beliefs when real attributes were named, so in this study we used arbitrary labels—Test A, Test B, and so on—to avoid such biases, albeit at the cost of some realism in the task. In the forced-choice tasks, two candidates were compared on each test, with results shown side by side. Here, the task was to predict which of the two candidates would be better overall. To our knowledge, this is the first study to utilize multicue probability learning with this kind of task.

The object of the experiment was to investigate the extent to which acquired strategies, as measured on the test

trials, were implicit or explicit in nature. For this reason, we also included self-rating of the cues used at the end of each task. Measuring subjective weights in multicue probability learning tasks is not usual, but one study in which this was done was York et al. (1987). Their findings suggested again at least a partial dissociation between judgmental strategy and explicit knowledge. Their learning task involved two positively predictive cues, one of which had a substantially higher cue validity than the other (i.e., it more strongly predicted the criterion). The regression weights in the judgment analysis revealed that people learned to give more weight to the more predictive cue following outcome feedback. However, 50% of the participants rated the two cues as equally important in the self-ratings taken after the experiment. Of those who did differentiate, the majority favored the stronger cue. Our own design involved a stronger form of this manipulation, because the weaker cues were wholly irrelevant.

We also wished to see whether the degree of self-insight depended on the complexity and difficulty of the learning task. Our hypothesis was that more complex tasks would make implicit learning more likely to occur (Reber, 1993) and self-insight measures more likely to drop. We believed that the forced-choice task would provide more complexity than the single task, because more information has to be processed. To manipulate complexity within each task, we also introduced two variables. First, the total number of cues presented (relevant and irrelevant) was either four or six. Second, the predictive cues were either a majority or a minority. Thus, the four problems (implemented for both single and forced-choice tasks) had two predictive cues out of six, four out of six, one out of four, or three out of four. We abbreviate these tasks as 2(6), 4(6), 1(4), and 3(4). Our hypotheses were (1) that the task would be more difficult when the total number of cues was higher (six) rather than lower (four) and (2) that majority learning would be harder than minority learning. For example, learning which cue predicts in the 1(4) task should be easier than learning which cue does *not* predict in the 3(4) task. One reason for this is that assuming all four cues to be relevant when only three are will still lead to relatively high performance in this condition, so the irrelevant cue might go unnoticed.

Method

Participants

Twenty-four undergraduate students at the University of Plymouth participated on a paid volunteer basis, 12 in each group.

Design

The participants were divided into two groups according to whether they received the tasks with four or six total cues. Within each group, each participant was given either single-task problems in the first session and forced-choice problems in the second session or vice versa. Within each session, they completed a majority learning task (3[4] or 4[6], depending on the group), followed by a minority learning task (1[4] or 2[6]), or vice versa, with orders being determined randomly and independently for each participant. The position of relevant and irrelevant cues was also determined randomly and independently for each task.

The values for each displayed cue were generated separately for each task presented to each participant. These were generated randomly and independently with equal mean and variance. Hence, the cues were theoretically uncorrelated but, due to random processing, might have small intercorrelations in any particular generated set. For purposes of feedback, the displayed values of only the relevant cues were combined in a linear model and then adjusted by random addition or subtraction of an error constant in such a way as to add 25% of random noise with a normalized standard deviation. Due to random generation, actual beta values for the set of cases inspected could vary somewhat across tasks but would result in large positive values for relevant cues.

Procedure

Each participant attended for two 90-min sessions in successive weeks, completing two of the four tasks in each session. All the problems were presented on a computer using custom-written software. The participants were tested in a small room on individual workstations, with up to 6 being run in parallel. The main task was carried out entirely by interaction with the computer, with the participants using a mouse to set values with a slider (of constant length in all tasks and experiments) and click confirmation of judgments. However, the self-rating task (see below) was administered as a pencil-and-paper task following completion of each computer task.

Single task. The participants were asked to judge a candidate's suitability for a job (unspecified) by using the mouse to drag a slider along a line labeled from *poor* to *good*. They were provided first with the results of either four or six ability tests, whose results were displayed verbally at one of five levels: *very poor*, *poor*, *average*, *good*, or *very good*. For example, a screen might display the following:

Test A	Good
Test B	Very Good
Test C	Poor
Test D	Very Good

followed by a slider on which they could drag a bar to the point between poor and good that represented their judgment (always initially set in the center). On the training trials, once they clicked to fix their judgment, an identical slider immediately appeared beneath this one, with the bar fixed to indicate the actual suitability of the candidate, providing the (noisy) outcome feedback. The participants were told that some, and possibly all, of the information given would be relevant. They were told that they would have to guess at first but that the feedback should help them improve their judgments.

Forced-choice task. The forced-choice task was similar, except that participants were asked to judge which of two candidates was better on the basis of comparative test results. A typical screen display might show

	Case 1	Case 2
Test A	Good	Good
Test B	Poor	Very Good
Test C	Very Poor	Poor
Test D	Good	Average

In this case, the sliders used for judgment and feedback were labeled Case 1 on the left and Case 2 on the right. Preference for either candidate was indicated by pulling the bar right or left from the central starting point.

Feedback on the forced-choice task was computed by first calculating noise-added criterion values for each candidate, as on the single task. The Case 1 value was then subtracted from the Case 2 value to compute a difference score distributed either side of zero. Finally, this feedback score was normalized to the same standard deviation as on the single task before being presented on the screen.

Self-rating task. After completing both the training and the test phase tasks for each problem, the participants were asked to rate the relevance of each test on the problem just completed on a scale from 1 (*less relevant*) to 7 (*more relevant*). Note that although cues

had actually been set to one of two values, totally predictive or totally nonpredictive, the participants were not made aware of this.

Results

Performance

The mean performance scores are shown in the top of Table 1. These scores are simply the correlation between judgments and criterion over the 40 test trials. It will be noticed that all scores are positive and well above zero. A three-way mixed model analysis of variance (ANOVA) was performed with performance scores as the dependent variable, after application of the Fisher z transformation to improve distribution properties. The factors were size (four or six cues, between groups), task (single task or forced choice, within groups), and pattern (majority or minority learning, within groups). There was a highly significant effect of task [$F(1,22) = 21.74$, $MS_e = 0.72$, $p < .001$] due, as was expected, to performance being much better on the single task (mean score, .72) than on the forced-choice task (.56), but neither size [$F(1,22) = 3.85$, $MS_e = 1.19$] nor pattern [$F(1,22) = 1.56$, $MS_e = 0.56$] had any significant effect in the ANOVA, and there were no significant interactions.

To compute task knowledge (G) scores, we ran individual multiple regression analyses on the four tasks completed by each participant, regressing the cue values onto the judgments made in the 40 test trials. G scores were then computed by dividing the performance correlations for each task by the multiple correlation coefficient, in line with the equation given in the introduction. The G scores are shown at the bottom of Table 1. These are effectively corrections of the raw performance scores, which remove any nonlinear effects of performance and, therefore, measure directly the correspondence between the linear weights that participants apply to cues (the beta weights in the regression analysis) and the linear weights in the criterion model. The G statistic removes both unreliability in judgments and the ecological unreliability of the added noise. This correction generally raises scores and corrects for differences in linear fit be-

tween conditions. We ran an ANOVA on the G scores similar to that run on the performance scores. As before, task produced a significant effect [mean, single task = .84, forced choice = .71; $F(1,22) = 7.22$, $MS_e = 0.054$, $p = .013$], and pattern was not significant [$F(1,22) = 2.70$, $MS_e = 0.040$]. However, in this analysis, the effect of size was marginally significant [$F(1,22) = 4.10$, $MS_e = 0.037$, $p = .055$]. This reflects higher mean scores when the number of cues was smaller—that is, four (.81) rather than six (.73). This suggests that adding extra irrelevant cues reduces the amount that people learn.

Self-Insight

Insight was measured by correlating the beta weights in each individual participant's regression analysis with their self-rating of cue usage across the four or six cues used in each task. Note that these correlations were independent of the performance measures. They measured the correspondence between people's explicit judgments of the relevance of cues and the use of cues they actually made, regardless of how accurate (against criterion) their actual cue use was. The mean insight scores are shown in Table 2. We ran an ANOVA on insight scores (with Fisher z transformation), with the same factor structure as that for the ANOVAs just reported. This analysis yielded only one significant effect: Insight was significantly higher [$F(1,22) = 4.58$, $MS_e = 0.44$, $p = .044$] for the single task (mean, .58) than for the forced-choice task (.38). There was also a significant interaction between task and learning [$F(1,22) = 4.87$, $MS_e = 0.46$, $p < .05$]. Inspection of Table 2 reveals that insight was higher for minority learning on the single task but was a little lower on the forced-choice task.

Explicit Knowledge

To measure the extent to which training resulted in explicit knowledge of the criterion, we calculated the mean self-ratings for the cues that were actually relevant (predictive in the training phase) and subtracted the mean ratings for the cues that were actually irrelevant. We call this measure the *explicit knowledge index* (EKI) and show the mean values for the different conditions in Table 3. It will be seen from the table that the EKI values were substantially above zero. The presence of explicit knowledge was confirmed by a set of one-tailed one-sample t tests for each condition, combining the four and six cue groups for added power. In all cases, the EKI scores were significantly above zero [single task, majority learning, $t(23) = 3.49$, $p < .001$; single task, minority learning, $t(23) = 3.42$, $p < .001$; forced choice, majority learning, $t(23) = 3.41$, $p < .001$; forced choice, minority learning, $t(23) = 2.85$, $p < .005$]. To see whether explicit knowledge was influenced by the kind of task that was learned, we performed a three-way ANOVA on the EKI scores, with the number of cues, type of learning (majority or minority), and task (single or forced choice) as the factors. Surprisingly, there was no main effect of task [$F(1,22) = 1.19$, $MS_e = 1.83$] or of any other factors, de-

Table 1
Experiment 1:
Mean Performance (r_p) and Task Knowledge (G) Scores

Task	Number of Cues	Minority Learning		Majority Learning	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Performance Scores					
Single	4	.80	.17	.79	.06
	6	.59	.27	.69	.15
Forced choice	4	.53	.30	.64	.15
	6	.52	.17	.56	.23
<i>G</i> Scores					
Single	4	.89	.10	.89	.05
	6	.74	.23	.83	.06
Forced choice	4	.66	.32	.81	.15
	6	.67	.18	.69	.32

Table 2
Experiment 1: Mean Insight Correlations

Task	Number of Cues	Minority Learning		Majority Learning	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Single	4	.67	.53	.45	.62
	6	.65	.36	.54	.54
Forced choice	4	.38	.74	.40	.73
	6	.27	.48	.47	.49

spite the fact that the forced-choice task was substantially more difficult, as measured by performance and task knowledge (*G*) scores, and was associated with poorer self-insight in the analyses reported above.

Discussion

Both performance and task knowledge scores were well above zero for both the single task and the forced-choice task, apparently indicating that learning of these rather complex problems had occurred. However, because all the cues were either positive predictors or non-predictors, there was a possibility of spurious performance correlations. For example, if the participants had assumed that all the cues were predictive, or had guessed a random subset without actually having learned from the feedback trials, they still would have achieved overall positive correlations with the criterion. Hence, as a check, we computed the difference between the mean beta weightings that the participants gave to the relevant cues (predictive in the training phase) and those given to the irrelevant cues. These difference scores were, on average, well above zero (.32 for the single task and .22 for the forced-choice task). Hence, as was expected from our survey of the literature (Adelman, 1981; Lindell, 1976; Taylor et al., 1996; York et al., 1987), training the participants with outcome feedback on a multicue probability learning task with 25% added noise did produce clear evidence of learning.

The issue of interest is the nature of the learning that has taken place. To investigate this, we also asked the participants to give explicit ratings of cue relevance after performance on each task. On this basis, we were able to determine that at least some of the learning that had occurred was explicit, because our explicit knowledge indices were significantly above zero. However, we do not believe that explicit learning gives the whole story. This is because the more complex forced-choice task significantly reduced judgmental performance without apparently affecting the explicit knowledge measure. Overall, the explicit knowledge score was only marginally higher for the single task (1.50) than for the forced-choice task (1.37). It is possible that there is a small effect here that a study of larger statistical power could detect. However, we should note that the comparison on performance correlations, with a similar degree of power, produced a substantially significant effect ($p < .001$).

Turning to the insight scores, there are two points of interest. First, in a literature in which poor to moderate

self-insight is typically claimed (Harries et al., 2000; Wigton, 1996), the correlations shown in Table 2 are relatively high—well above zero in most conditions. Second, insight scores dropped significantly on forced-choice tasks, in parallel with the drop in performance, but in contrast with the explicit knowledge scores. This is an interesting finding, but it would be premature to discuss it prior to a replication. In Experiment 2, we attempted to do this by introducing within-task factors that we hoped would affect complexity.

EXPERIMENT 2

In Experiment 2, we decided to concentrate on just the single task, since it differs in several respects from the forced-choice task. What we wanted was to introduce factors within this task that would affect difficulty. We had hoped that the majority/minority learning factor would have this effect, but there was no evidence of an effect in Experiment 1. Hence, we introduced a new factor into the single-choice task that we had reason to expect would influence its difficulty. Instead of having cues that predicted either positively or not at all, we also included cues that predicted the criterion negatively. Negative cues are known to be harder to learn than positive cues (see, e.g., Brehmer, 1974; Lindell, 1976). To use negative predictors, we had to modify the scenario slightly. It was still described as a personnel selection problem, but the tests reported were now described as personality, rather than ability, tests. Hence, for some tests, a higher score might predict ability in the job, whereas for other tests a lower score might predict performance. In Experiment 2, the participants were unsure before training as to whether particular test results would predict positively or negatively or would prove irrelevant. This had the further advantage that the possibility of spurious performance, discussed in relation to Experiment 1, was effectively eliminated. The participants could not achieve good performance simply by assigning higher criterion scores to higher values of some cues, because predictive cues were equally likely to predict negatively, so that higher values would reduce the criterion score.

Method

Participants

Thirty-six undergraduate students at the University of Plymouth participated in the experiment on a paid volunteer basis.

Table 3
Experiment 1: Measures of Explicit Knowledge
(Explicit Knowledge Index Scores)

Task	Number of Cues	Minority Learning		Majority Learning	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Single	4	2.45	2.21	0.72	2.14
	6	0.87	2.37	1.94	1.37
Forced choice	4	1.08	2.26	1.14	2.05
	6	1.19	1.70	1.37	1.59

Design

As in Experiment 1, the participants were tested in two separate groups according to whether a total of four or six test results were displayed for each case. They were tested over two sessions, completing six separate tasks (all using the single-task method) in a random order, three in each session. The second session followed within 3–10 days of the first. There were two within-group factors: polarity (positive or negative cuing) and pattern. The patterns labeled positive are represented as +, +++, and ++-. In the first case (+), only one cue (out of four or six, depending on the group) was relevant and predicted the criterion in a positive manner. The second pattern was +++, with three relevant cues all predicting positively. The final pattern was ++-, a mixed case, but with two of the relevant cues being positive and the last negative. The three negative patterns were similarly -, ---, and --+. In all cases, the position of the relevant cues was determined randomly and independently for each task assigned to each participant.

Procedure

The procedure was similar to that of Experiment 1, with the single-task format. The main differences were (1) that the scenario referred to personality, rather than ability, tests, (2) that the labels used to describe the cue values were *very low*, *low*, *average*, *high*, or *very high* (rather than *very poor*, etc.), and (3) that the participants were told that either high or low scores for a particular test might predict better performance in the job. The self-rating task was modified so that the participants rated cues from -3 (*predict negatively*) through 0 (*irrelevant*) to +3 (*predict positively*).

Results

Performance

Performance scores are shown at the top of Table 4. They were generally lower than those in Experiment 1, even in the positive cue conditions. We assume that prior uncertainty about the direction of prediction of the cues increased learning difficulty generally. A three-way ANOVA was carried out, with size (four or six cues) as a between-group factor and with polarity (positive or negative) and pattern (three values) as within-group factors, with Fisher z transformation. In this experiment, the size of the cue set had a substantial effect on task difficulty [mean for four cues = .44, for six cues = .26; $F(1,33) = 9.25$, $MS_e = 0.35$, $p < .005$]. There was also a significant main effect of polarity [mean for positive condition = .43, for negative = .27; $F(1,33) = 19.01$, $MS_e = 0.18$, $p < .001$] and a significant interaction between polarity and pattern [$F(2,66) = 7.08$, $MS_e = 0.14$, $p < .0005$]. Inspection of Table 4 indicates that this was

due to the three-cue condition's yielding an advantage for positive cues only.

As in Experiment 1, individual multiple regression analyses were carried out, and task knowledge (G) scores were computed. The means of these are shown at the bottom of Table 4. The ANOVA produced three significant main effects similar to those for the performance scores: size [four cues = .57, six cues = .29; $F(1,34) = 12.63$, $MS_e = 0.33$, $p = .001$], polarity [positive = .51, negative = .35; $F(1,34) = 6.72$, $MS_e = 0.22$, $p = .014$], and pattern [one cue = .40, three cues = .56, mixed cues = .33; $F(2,33) = 6.16$, $MS_e = 0.16$, $p = .013$]. The pattern of significant interactions differed from that observed with the performance scores, however.

First, there was no significant interaction between polarity and pattern ($F < 1$), because the advantage of the three-cue condition could now be seen for negative as well as positive cues. However, there was a significant interaction between polarity and cues [$F(1,34) = 5.97$, $MS_e = 0.22$, $p = .02$] that was not present in the performance ANOVA. As an inspection of the G scores in Table 4 reveals, task knowledge was substantially higher for positive than negative cues only when the total number of cues was six rather than four. Since the G statistic can be taken as the more accurate psychological measure of learning, for the reasons explained above, we prefer to draw conclusions from these results. Hence, the difficulty of extra irrelevant cues is seen as restricted to negative cue patterns in this experiment, whereas the facilitation of learning with three relevant cues of the same polarity is seen as general to both positive and negative learning. In fact, six cue sets containing negative cue patterns seemed to result in almost no learning at all.

Self Insight

Mean insight scores are shown in Table 5. As in Experiment 1, these were substantially above zero. An ANOVA on transformed scores revealed significant main effects of polarity [mean for positive cues = .60, negative = .44; $F(1,34) = 8.15$, $MS_e = 0.75$, $p < .01$], size [four cues = .64, six cues = .40; $F(1,34) = 14.42$, $MS_e = 1.18$, $p < .001$], and pattern [one predictor = .49, three predictors (same) = .49, three predictors (mixed) = .58; $F(2,68) = 3.76$, $MS_e = 0.60$, $p < .05$] but no significant interactions. Hence, In Experiment 2, not only were per-

Table 4
Experiment 2:
Mean Performance and Task Knowledge (G) Scores

Cues	+		+++		++-		-		---		--+	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Performance Scores												
4	.46	.33	.69	.16	.31	.47	.42	.29	.42	.19	.33	.30
6	.41	.33	.44	.19	.30	.31	.18	.36	.11	.39	.09	.34
G Scores												
4	.52	.37	.79	.19	.40	.53	.56	.33	.72	.21	.40	.50
6	.50	.40	.53	.30	.32	.48	.07	.53	.18	.57	.18	.46

Table 5
Experiment 2: Mean Insight Correlations

Number of Cues	+		+++		++-	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
4	.67	.35	.79	.24	.81	.44
6	.40	.48	.44	.46	.46	.55
	-		---		--+	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
4	.63	.38	.34	.64	.58	.54
6	.21	.55	.39	.47	.48	.53

formance and task knowledge scores significantly lower when the total number of cues was higher and when the training was based on negative rather than positive cues, but also the same trends were observed in the insight analysis.

Explicit Knowledge

The analysis of explicit knowledge for Experiment 2 was a little more complicated than that for Experiment 1, due to the presence of both positive and negative cues. If the participants acquired explicit knowledge, we would expect their ratings of positive cues to be more positive than those of irrelevant cues and their ratings for negative cues to be more negative than those of irrelevant cues. We therefore measured two indices subtracting mean ratings in this way, which we call the positive knowledge index (PKI) and the negative knowledge index (NKI), respectively. The PKI was measured across the +, +++, ++-, and --+ tasks, and the NKI was measured for the -, ---, ---+, and +++- tasks, so that the mixed cue tasks provided data for both indices. The mean scores for each index are shown in Table 6.

It will be recalled that the introduction of cues with unpredictable direction of influence made Experiment 2 generally a lot more difficult than Experiment 1 (compare Tables 1 and 4). A reduction in explicit learning might be a factor in this performance drop, since the explicit knowledge measures are also much lower than those in Experiment 1. One-sample *t* tests revealed that only one of the four conditions shown in Table 6 had an index significantly different from zero [NKI for the six-cue condition; $t(17) = -2.01, p < .05$], suggesting that little explicit learning had occurred. We have seen already that both the number of cues and their polarity affected task difficulty and self-insight measures in parallel. It is, therefore, of considerable theoretical interest to determine whether these factors also influenced explicit knowledge acquisition. To assess this, we reversed the sign of the NKI and put all the indices into a two-way ANOVA with number of cues and polarity as factors. There were no significant effects in this ANOVA. Hence, neither of the factors affecting task difficulty had a detectable effect on our explicit learning measure. It is particularly surprising that there was no greater explicit knowledge for positive than for negative cues (in fact,

the means are slightly in the wrong direction), given the massive effect of this factor on performance.

Discussion

We have compelling evidence that performance in Experiment 2 was largely mediated by implicit learning. Although performance and task knowledge scores were substantially lower than those in Experiment 1, they were still well above zero, indicating that learning had occurred. Moreover, the level of these measures was substantially affected by two factors that affected the difficulty of the task: the use of additional irrelevant cues in the display and the presence of cues that were negatively, rather than positively, related to the criterion. By contrast, there was little evidence of explicit learning and no indication that these task difficulty factors had a significant effect on the explicit learning indices.

As in Experiment 1, self-insight measures were affected by the factors influencing task difficulty, whereas explicit knowledge measures were not. This is an intriguing and, we believe, theoretically important finding for which a clear interpretation will be offered in the General Discussion section. It is, however, relevant at this point to consider what may have caused the drop in performance with more cues and with negative patterns. This could happen because people were not learning and were simply guessing more (lowering cognitive control values), because the participants were acquiring judgmental strategies that were less accurately matched to the criterion (lower task knowledge or *G* scores), or possibly because of both. The fact that *G* scores were also significantly lower with more cues and with negative cue patterns indicates that the latter was definitely involved: People acquired inaccurate models. To check whether linear models also accounted for less variance when performance was lower, we ran an additional ANOVA on the *R*²s adjusted values for each participant in each condition, taken from the multiple regression analyses. In contrast with the large effect of cues on *G* scores, this analysis revealed no significant effect of size [mean for four cues = .49, six cues = .47; $F(1,34) < 1$]. However, a significant effect of polarity was observed [mean for positive cues = .52, negative = .44; $F(1,34) = 9.18, MS_e = 0.04, p = .005$]. So it seems that additional (irrelevant) cues do not prevent people from acquiring linear judgment models but simply render these models less accurately matched to the criterion. Negative cues, however,

Table 6
Experiment 2: Explicit Knowledge Measures

Number of Cues	PKI		NKI	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
4	-0.16	1.88	-0.03	1.38
6	0.31	0.96	-0.41	0.85

Note—PKI, positive knowledge index; NKI, negative knowledge index.

reduce both the amount and the accuracy of the learning that takes place.

EXPERIMENT 3

The purpose of the present research was to investigate the extent to which multicue probability learning involves implicit or explicit cognitive processes. In Experiment 3, we took a different approach to this issue from that in the previous experiments. We wanted to see whether the factors that affect task difficulty when strategies have to be acquired by outcome feedback learning would also affect performance of the task presented in a clearly explicit manner. To do this, we dispensed with the training trials entirely and instructed the participants explicitly as to which were the relevant cues. We did this for exactly the same conditions as those investigated in Experiment 2, running only the 40 (test) trials in which no feedback was presented. This is quite similar to the procedure followed in some studies of so-called cognitive feedback in the literature (e.g., Adelman, 1981).

Method

Participants

Thirty-six undergraduate students at the University of Plymouth participated in the experiment on a paid volunteer basis.

Design

The design was based on that of Experiment 2, with three differences. First, no training trials were provided, but *test* trials (no feedback) were administered in a fashion identical to that used in Experiment 2. Second, prior to these trials, the participants were told which cues were relevant and in which direction. The final difference from Experiment 2 was that the self-rating test was omitted. It would be pointless to ask people which cues were relevant when they had been told this at the outset.

Procedure

The participants were given the personnel selection problem with personality test results as used in Experiment 2. Instead of the training phase, they were instructed to make judgments in line with explicit information about the relevance of the tests. For each task, they were shown a sheet like the following:

Test A = 0
Test B = +
Test C = -
Test D = 0

The participants were told that + tests predict positively the candidate's suitability for the job, that - tests predict negatively, and that 0 tests were irrelevant. This sheet of paper remained in front of them throughout their performance of the computer task.

Results

Mean performance scores for Experiment 3 are shown in the top of Table 7. These were substantially higher than those in Experiment 2, as was to be expected given that the participants were told the influence of cues without having to learn this for themselves and in line with general claims that cognitive feedback is more effective than outcome feedback (Balzer et al., 1989). An ANOVA on transformed scores revealed a significant main effect of polarity [mean for positive cues = .76, mean negative = .66; $F(1,34) = 10.71$, $MS_e = 0.36$, $p < .005$] and of pattern (one predictor = .73, one predictor (same) = .70, one predictor (mixed) = .71) and an interaction between polarity and pattern [$F(2,68) = 3.83$, $MS_e = 0.40$, $p < .05$]. The interaction appears to reflect a particularly large polarity effect for tasks with a single predictive cue. *G* scores are shown at the bottom of Table 7. In this case, the ANOVA revealed no significant effects. Mean *G* scores were higher for positive cues (mean = .84) than for negative cues (.76), but the effect was well short of significance [$F(1,34) = 1.64$, $MS_e = 0.19$].

We also ran an ANOVA on the combined performance scores of Experiments 2 and 3, treating experiment as a between-group factor. This revealed an unsurprisingly huge main effect of experiment [mean, Experiment 2 = .35, Experiment 3 = .71; $F(1,67) = 58.77$, $MS_e = 1.31$, $p < .001$]. Of more interest was the significant main effect of polarity [mean positive = .59, negative = .46; $F(1,67) = 23.54$, $MS_e = 0.28$, $p < .001$] and the fact that it did *not* interact with experiment [$F(1,67) = 0.004$, $MS_e = 0.001$]. This analysis produced no main effect of cues but a significant cues $\times$ experiment interaction [$F(1,67) = 5.26$, $MS_e = 1.31$, $p < .05$].

It will be recalled that performance fell significantly when more cues were presented in Experiment 2. The trend was (nonsignificantly) reversed in Experiment 3. Finally, there was a significant main effect of pattern [$F(2,66) = 16.10$, $MS_e = 0.29$, $p < .001$], which also significantly interacted with experiment [$F(2,66) = 7.93$, $MS_e = 0.29$, $p < .001$], and an interaction between polarity and pattern [$F(2,134) = 3.41$, $MS_e = 0.86$, $p < .05$]. In

Table 7
Experiment 3: Mean Performance and Task Knowledge (*G*) Scores

Cues	+		+++		++-		-		---		--+	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Performance Scores												
4	.79	.34	.63	.57	.67	.30	.55	.52	.64	.43	.67	.34
6	.89	.20	.82	.23	.76	.17	.69	.55	.68	.49	.73	.21
<i>G</i> Scores												
4	.81	.42	.69	.63	.79	.35	.52	.70	.72	.57	.80	.40
6	.93	.18	.91	.20	.89	.16	.75	.56	.88	.31	.90	.17

summary, the factors that affected performance in Experiment 2 failed to affect performance in Experiment 3, with the exception of polarity, which had a similar effect. However, the polarity effect did not extend significantly to the analysis of task knowledge scores.

GENERAL DISCUSSION

The purpose of this research was to investigate the relative influence of implicit and explicit cognitive processes in multicue judgment. Studies of expert groups in the social judgment theory tradition typically show low to moderate *self-insight*, as measured by the correlation between self-ratings of cue relevance and beta weights, the latter measuring statistically the influence that cues have on the judgments made. However, this gives only a limited indication of the degree to which implicit and/or explicit processes may be responsible for the judgments. In the present study, we attempted to simulate the everyday process of learning by experience, by using a computer-controlled multicue probability learning task in which noisy outcome feedback was provided throughout a training phase. A subsequent test phase, administered without outcome feedback, allowed us to determine the judgmental policies acquired and to compare these with self-reports of cue relevance.

Research on implicit learning (for reviews, see Berry & Dienes, 1993; Dienes & Perner, 1999; Reber, 1993) has shown on a variety of tasks that people can acquire the ability to control or predict complicated systems and patterns, without being able to report explicitly the rules they have learned. For this reason, we introduced several factors that might affect the complexity of the task. In Experiment 1, we found that the use of a forced-choice task reduced performance and task knowledge (*G*), as compared with a single-choice task. In Experiment 2, we found that performance and task knowledge declined for tasks using negatively, rather than positively, predictive cues and for tasks in which additional irrelevant cues were present. In all these cases, we also found a corresponding decline in our measure of self-insight. The question to be addressed here is why insight measures should drop when the task is more difficult. This is far from self-evident, since the insight measure is statistically independent from the performance measures. It is measured as the correlation between explicit judgments of cue relevance and the actual beta weights used in the judgments, regardless of whether the judgments were accurate. Also, measures of explicit knowledge were *not* affected by the same factors.

When one examines the pattern of correlations closely, it implies a causal linkage that leads us to question the normal interpretation of the *insight* measures. The term implies a measure of the degree of awareness of underlying processes. Our results, however, suggest that it really measures the extent to which explicit processes influence judgments. Consider the findings of Experiment 2.

Our three observable quantities were (1) criterion, (2) explicit ratings of relevance, and (3) judgments made.

We note that 1 correlates with 3, because performance correlations (judgments with criterion) are well above chance, and also that 2 correlates with 3, because this is measured by our self-insight correlations. However, 1 correlates little with 2, because there was only slight evidence of explicit knowledge of the criterion in Experiment 2. A pattern in which 3 correlates with 1 and 2 but in which 1 and 2 do not correlate with each other implies that 1 and 2 have independent causal links to 3. This means that the world (outcome feedback) can influence judgments independently of explicit knowledge via its influence on the inferred implicit learning process. However, this analysis also implies that explicit knowledge independently influences judgments even when such knowledge is *not soundly based*. The superior performance in Experiment 1 may well be due to the fact that, in this experiment, the participants did show evidence of explicit learning of the criterion.

These findings provide strong support for dual-process theories of the kind referenced in the introduction. However, they also show a subtle distinction between a dual-process theory of learning and one applied to judgment or reasoning. The former (somewhat controversial) position is that implicit and explicit learning are distinct processes arising from separate cognitive mechanisms (Dienes & Perner, 1999; Reber, 1993). The latter is that distinct implicit and explicit processes compete for control of responses to inferential and judgmental tasks (Evans & Over, 1996; Sloman, 2002; Stanovich, 1999). For example, in the study of *belief biases* in reasoning, there is evidence that logical and belief-based processes compete within individuals (Evans, Barston, & Pollard, 1983; Stanovich & West, 1997), a finding recently supported by neuropsychological evidence (Goel, Buchel, Frith, & Dolan, 2000).

It has been shown many times in the implicit learning literature that evidence of explicit learning may be weak or absent on complex control tasks (Berry & Dienes, 1993; Reber, 1993). Hence, the drop in explicit knowledge acquisition in Experiment 2 could well be due to the complexity introduced by negative cuing. What the present findings add, however, is that the contribution of explicit processes to subsequent judgments (in combination with implicit processes) may be unaffected by this lack of learning. We can only speculate on how this might impact on real-world judgments. Experience-based implicit learning might be helpfully supplemented by explicit knowledge based on expert study, but it might equally be overridden by prejudicial beliefs and opinions.

Other findings in this study are, first, that implicit learning is impaired by the presence of irrelevant cues in the environment. This factor becomes irrelevant when people are explicitly told which information needs to be heeded (Experiment 3). The second and major influence was the presence of negative predictors that impaired

performance in both Experiments 2 and 3, even though outcome feedback training was used in one and cognitive feedback, or explicit instruction, in the other. Not only are negative predictors harder to learn, but their mere presence increases task difficulty, which is shown by comparing single-task performance between Experiments 1 and 2. The difficulty of negative predictors was known (Brehmer, 1974), but our results are interesting nevertheless and link with a range of evidence for positivity biases in reasoning and judgment (Evans, 1989; Klayman & Ha, 1987). In particular, it seems that cue polarity affected both the implicit learning, which formed the main basis for judgments in Experiment 2, and the ability to comply with explicit instructions in Experiment 3. Although polarity did not affect explicit knowledge acquisition in Experiment 2, it seems there was little explicit learning in this experiment for it to influence.

In conclusion, this study has provided some clear and intriguing answers to the questions we set out with. First, it seems that both implicit and explicit processes influence multicue judgments, in line with strong evidence of dual processing on other reasoning and judgment tasks. Second, the correlation between cue weights and explicit relevance ratings, normally regarded as a measure of self-insight, might more accurately be characterized as the extent to which explicit processes influence judgment. The reason that expert judges are typically reported as having moderate self-insight in the social judgment theory literature (Harries et al., 2000) could well be because part, but only part, of their judgments are explicitly produced. Third, in line with other findings in the implicit learning literature, our results show that explicit knowledge acquisition from experience-based learning is affected much more by task difficulty than is implicit learning. Finally, our findings strongly indicate that explicit processes continue to exert their partial influence on judgments even when little explicit learning has occurred.

REFERENCES

- ADELMAN, L. (1981). The influence of formal, substantive and contextual task properties on relative effectiveness of different forms of feedback in multiple-cue probability learning tasks. *Organizational Behavior & Human Performance*, **27**, 423-442.
- BALZER, W. K., DOHERTY, M. E., & O'CONNER, R., JR. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, **106**, 410-433.
- BERRY, D. C., & DIENES, Z. (1993). *Implicit learning*. Hove, U.K.: Erlbaum.
- BREHMER, B. (1974). Hypotheses about relations between scaled variables and the learning of probabilistic inference tasks. *Organizational Behavior & Human Performance*, **11**, 1-27.
- BREHMER, B. (1980). In one word: Not from experience. *Acta Psychologica*, **45**, 233-241.
- COCKSEY, R. W. (1996). The methodology of social judgement theory. *Thinking & Reasoning*, **2**, 141-173.
- DIENES, Z., & PERNER, J. (1999). A theory of implicit and explicit knowledge. *Behavioral & Brain Sciences*, **22**, 735-808.
- DOHERTY, M. E., & KURZ, E. M. (1996). Social judgement theory. *Thinking & Reasoning*, **2**, 109-140.
- EVANS, J. ST. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, U.K.: Erlbaum.
- EVANS, J. ST. B. T., BARSTON, J. L., & POLLARD, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, **11**, 295-306.
- EVANS, J. ST. B. T., & OVER, D. E. (1996). *Rationality and reasoning*. Hove, U.K.: Psychology Press.
- GILOVICH, T., & GRIFFIN, D. (2002). Introduction—Heuristics and biases: Then and now. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 1-18). Cambridge: Cambridge University Press.
- GOEL, V., BUCHEL, C., FRITH, C., & DOLAN, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, **12**, 504-514.
- HARRIES, C., EVANS, J. ST. B. T., & DENNIS, I. (2000). Measuring doctors' self-insight into their treatment decisions. *Journal of Applied Cognitive Psychology*, **14**, 455-477.
- HARRIES, C., EVANS, J. ST. B. T., DENNIS, I., & DEAN, J. (1996). A clinical judgement analysis of prescribing decisions in general practice. *Le Travail Humain*, **59**, 87-111.
- KAHNEMAN, D., & FREDERICK, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49-81). Cambridge: Cambridge University Press.
- KLAYMAN, J. (1984). Learning from feedback in probabilistic environments. *Acta Psychologica*, **56**, 81-92.
- KLAYMAN, J., & HA, Y. W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, **94**, 211-228.
- LINDELL, M. K. (1976). Cognitive and outcome feedback in multiple-cue probability learning tasks. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 739-745.
- NISBETT, R. E., & WILSON, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231-295.
- REBER, A. S. (1993). *Implicit learning and tacit knowledge*. Oxford: Oxford University Press.
- REILLY, B. A., & DOHERTY, M. E. (1992). The assessment of self-insight in judgment policies. *Organizational Behavior & Human Decision Processes*, **53**, 285-309.
- SLOMAN, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, **119**, 3-22.
- SLOMAN, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 379-398). Cambridge: Cambridge University Press.
- STANOVICH, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- STANOVICH, K. E., & WEST, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, **89**, 342-357.
- STEWART, T. R. (1988). Judgement analysis: Procedures. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgement: The SJT view* (pp. 41-74). Amsterdam: Elsevier, North-Holland.
- TAYLOR, L. A., HALL, P. D., COSIER, R. A., & GOODWIN, V. L. (1996). Outcome feedback effects on risk propensity in an MCPLP task. *Journal of Management*, **22**, 299-311.
- WHITTLESEA, B. W. A. (2002). Two routes to remembering (and another to remembering not). *Journal of Experimental Psychology: General*, **131**, 325-348.
- WHITTLESEA, B. W. A., & BROOKS, L. R. (1990, November). *Factors affecting reliance on analytic and nonanalytic knowledge*. Paper presented at the 31st Annual Meeting of the Psychonomic Society, New Orleans.
- WIGTON, R. S. (1996). Social judgement theory and medical judgement. *Thinking & Reasoning*, **2**, 175-190.
- 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 & Human Performance*, **39**, 303-317.