European Journal of Cognitive Psychology

ISSN: 0954-1446 (Print) 1464-0635 (Online) Journal homepage: www.tandfonline.com/journals/pecp20

The role of alternative hypotheses in the integration of evidence that disconfirms an acquired belief

Frederic Vallee-Tourangeau, David M. Beynon & Staunton A. James

To cite this article: Frederic Vallee-Tourangeau, David M. Beynon & Staunton A. James (2000) The role of alternative hypotheses in the integration of evidence that disconfirms an acquired belief, European Journal of Cognitive Psychology, 12:1, 107-129, DOI: 10.1080/095414400382226

To link to this article: https://doi.org/10.1080/095414400382226

Published online: 21 Sep 2010.

EUROPEAN JOURNAL OF COGNITIVE PSYCHOLOGY, 2000, *12* (1), 107–129

The role of alternative hypotheses in the integration of evidence that disconfirms an acquired belief

Frédéric Vallée-Tourangeau, David M. Beynon, and Staunton A. James

University of Hertfordshire, Hatfield, UK

Two experiments examined the role of alternative hypotheses in the recognition of belief-incongruent evidence and the consequent attribution of probative value to that evidence. Using a contingency judgement and prediction task, subjects monitored multiple predictor-outcome contingencies. In a subset of three of these contingencies the evidence strongly endorsed a positive contingency in a first phase but strongly endorsed a negative contingency in a second phase. In Experiment 1 the negative evidence was presented, in part, in terms of an alternative contingency involving a new predictor or a new outcome, or in terms of no alternatives. The presence of alternative hypotheses did not influence the recognition of the negative evidence but significantly reduced the subjects' persistence in predicting the outcome in the presence of the predictor. Using the same positive–negative contingencies, Experiment 2 replicated this effect but also demonstrated that error in the feedback during the negative phase strengthened the perseverance in outcome predictions even when subjects acknowledged the negative nature of the evidence. Results from these two experiments indicate that prior beliefs do not bias the recognition of belief-incongruent evidence but the integration of that evidence is determined by the nature of the alternative hypotheses available to the reasoner.

The covariation between a candidate cause and an effect is an important cue that informs causal inferences. Early research on covariation detection concluded that judgements were mostly determined by the frequencies of cause-effect co-occurrences (e.g., Smedslund, 1963); however, more recent research paints a diametrically different picture which indicates that judge-

Requests for reprints should be addressed to Frédéric Vallée-Tourangeau, Department of Psychology, University of Hertfordshire, Hatfield, UK, AL10 9AB.

Email: psyqfv@herts.ac.uk

The research presented here was in part supported by Grant R000222542 from the Economic and Social Research Council awarded to FVT. We thank Mike Dougherty and Bradley Franks for thoughtful comments on a previous version of this manuscript.

http://www.tandf.co.uk/journals/pp/09541446.html

ments are in many circumstances a function of the actual degree of covariation (Shanks, 1995; Vallée-Tourangeau, Hollingsworth, & Murphy, 1998).

Alloy and Tabachnick (1984) argued that in situations where reasoners evaluate covariation data with certain preconceptions, their judgements are likely to concord with their prior beliefs. For example, in their 1969 paper Chapman and Chapman note that some Rorschach interpretations correlate with certain patient characteristics, in this case homosexuality. However, psychodiagnosticians almost never report the valid interpretations but rather cite popular, but uncorrelated associates (e.g., interpretations of feminine clothing). Chapman and Chapman then presented undergraduates with fictitious patient records where the unpopular but valid and popular but invalid interpretations were paired equally often with certain characteristics. The popular interpretations were estimated to occur more often with the target characteristic (viz., homosexuality) than the unpopular interpretations.

Nisbett and Ross (1980; see also Baron, 1994; Gilovich, 1991) suggest that "illusory" correlations arise out of the tendency to favour belief congruent evidence and discount belief-incongruent evidence. Although memory biases have been implicated (e.g., Rothbart, Evans, & Fulero, 1979; Trolier & Hamilton, 1986), Nisbett and Ross's account need not necessarily implicate a distorted perception or distorted recollection of the evidence. Subjects may perceive the evidence correctly, and even recollect it accurately, but the subsequent attribution of probative value will determine what subjects believe the data endorse. Nisbett and Ross argue that this asymmetric consideration of the data may render certain beliefs, in particular covariation expectations, resistant to change. For example, if a reasoner expects two binary variables to be positively correlated, then observations from cells A and D of the 2×2 contingency table (see Figure 1) may be attributed greater probative weight than observations from cells B and C. Even if a reasoner encounters data that oppose a positive-correlation expectation (i.e. many observations from cells B and C), the asymmetric consideration of the evidence might considerably weaken the impact of the evidence. In a (typically) gloomy passage, Nisbett and Ross (p. 169) wrote: "[covariation] theories [are] almost impervious to data and people who subscribe to a given covariation theory before encountering evidence that should have served to overturn the theory, often emerge from such encounters with the theory unblemished and unscathed".

EXPERIMENT 1

The present study investigated the impact of a prior expectation of the evaluation of new evidence that opposed the previously discussed expecta-

	E	–E
C	A	B
–C	C	D

Phase 1	Comet	–Comet
Team	10	2
–Team	2	10

Phase 2	Comet	–Comet
Team	2	10
–Team	10	2

Figure 1. The left 2×2 table shows a summary table for the event frequencies that make up the contingency between two binary variables. Cell A = cause present, effect present. Cell B = cause present, effect absent. Cell C = cause absent, effect present. Cell D = cause absent, effect absent. $\Delta P = P(\text{Effect}|\text{Cause}) - P(\text{Effect}|\text{No Cause})$, or the difference between the following ratios of cell frequencies: $A/A + B - C/(C + D)$. The middle and right tables show the event frequencies making up the positive (Phase 1) and negative (Phase 2) contingencies in the three experimental conditions of Experiments 1 and 2. Note: C = Cause, –C = No Cause, E = Effect, –E = No Effect.

tion. Using a contingency judgement task, subjects learned in a first phase that an outcome variable was contingent upon a predictor variable. In a subsequent phase the contingency became strongly negative. Thus, in lieu of beliefs acquired outside the laboratory, subjects acquired a contingency expectation on the basis of the repeated co-occurrences of a predictor and an outcome variable. The nature of the evidence that suggested the original belief as well as the nature of the ensuing disconfirmatory evidence was controlled. On every trial subjects were required to predict the presence or the absence of the outcome. Thus, the strength of the contingency expectations could be measured by recording the number of outcome-present predictions on trials where the predictor variable was also present. Subjects were also asked to judge the contingency at regular intervals.

Just as this study sought to investigate the extent to which an expectation can impede the assimilation of new conflicting evidence, this study also investigated how new hypotheses can facilitate the processing of disconfirming evidence. There is a marked preference in cognitive psychology and social cognition to investigate or stress the "biasing" influence of prior beliefs. Examples abound, to cite but a few: Rothbart et al. (1979, p. 344) looked at how "events succumb to the distorting influence of prior beliefs" and Gilovich (1991, p. 50) reviewed "the tendency of our expectations to influence the interpretation of new information [such that] people are inclined to see what they expect to see and conclude what they expect to conclude". Lord, Lepper, and Preston (1984, p. 1231) wrote: "modern psychology has provided substantial empirical evidence to buttress the argument that our beliefs pervasively color and bias our response to subsequent information, evidence or argumentation." Theories are like mist on eyeglasses: They obscure facts (to adapt a proverb cited in Gould, 1991).

Yet for all the experimental demonstrations (or suggestions) that beliefs are obstacles to the assimilation of new information, philosophers and psychologists have praised the role of beliefs and expectations in guiding research and discovery: In Popper's searchlight metaphor, science is the "indefatigable examination" of the power of hypotheses and expectations "to throw light on experience" (1976, p. 361). The notion that hypotheses individuate and segregate experience in relevant categories is ubiquitous in the history and philosophy of science,[1] and sometimes surfaces in cognitive psychology: "the diagnostic impact of data is defined only in the context of particular hypotheses ... in the absence of any hypotheses, the collection of data is merely idle stockpiling" (Fischhoff & Beyth-Marom, 1983, p. 242). Even those who research the "biasing" influence of prior beliefs acknowledge that "there can be no argument with the proposition that people—scientists and laypersons alike—would be ill-advised to eschew the use of preconceived theories and become unbridled empiricists" (Ross & Lepper, 1980, p. 18).

The present study sought to determine the influence of an alternative hypothesis in the assimilation of new evidence that opposed an acquired belief. Dawes (1988) argues that plausible alternative explanations are necessary to permit the proper assimilation of new evidence that contradicts a hypothesis. Dawes claims that negative evidence alone is often psychologically insufficient to compel reasoners to abandon a hypothesis, especially if that hypothesis is too plausible to abandon (one of his examples is the projective rationale underlying the use of the Rorschach which insulates the test against the absence of empirical validation; Losee (1993) provides telling examples on belief perseverance from the history of science). Dawes (p. 243, original emphasis) goes on to suggest that the "most effective way of attacking a belief based on plausibility is not presenting disconfirming evidence, but rather providing a *new* plausible hypothesis". In the absence of a plausible alternative hypothesis "we tend to minimise the impact of evidence against a hypothesis we have held on the grounds of previous belief or plausibility" (p. 247). Einhorn and Hogarth (1986) suggest that the strength of an explanation for a given phenomenon is a function of the explanation's current strength minus the strength of the current alternatives which can also explain the phenomenon. In line with this, Dougherty, Gettys, and Thomas (1997, Exp. 2) found that people's evaluation of the likelihood of a focal cause produ-

[1] William Whewell (1858/1989, pp.139–140) wrote amply on the importance of ideas in guiding induction: "the want of Ideas leaves the mind overwhelmed, bewildered, and stupefied by particular sensations with no means of connecting the past with the futures, the absent with the present, the example with the rule; open to the impression of all appearances, but capable of appropriating none."

cing an effect was mediated by the likelihood that alternative causes could have produced the effect: The more likely these alternatives seemed to be, the less the likelihood attributed to the focal cause.

The present study provided a judgement task which tested the influence of alternative explanations in the evaluation of evidence that opposes an original hypothesis. Baron (1994) argues that belief persistence in the face of incongruent evidence may be due to a failure to engage in open-minded thinking: Some reasoners are more open-minded than others, and as such are more likely to abandon a belief. Although this may be an interesting individual difference, the study aimed to determine whether persistence could also be a feature of the context in which reasoning takes place. Thus, the same subjects evaluated evidence that disconfirmed an acquired belief either in a context which supplied an alternative explanation for that evidence or in a context that did not. Experiment 1 used a contingency judgement task in which subjects monitored concurrently multiple predictor-outcome contingencies. In a sub-set of three contingencies, the evidence strongly endorsed a positive-contingency hypothesis in a first phase and then the evidence strongly endorsed a negative-contingency hypothesis in a second phase. In two of these three conditions, the negative evidence was presented, partly, in terms of positive evidence for an alternative contingency involving either a new predictor variable or a new outcome variable. In a third condition, no alternative contingencies were provided. The theoretical considerations in Einhom and Hogarth (1986) and Dawes (1988) suggest that the negative evidence presented in a context that provides an alternative explanation will be attributed greater probative value, resulting in a greater adjustment of the positive-contingency expectation.

Method

Subjects. Twenty-four undergraduates from the University of Hertfordshire participated in this experiment.

Task scenario. Subjects were asked to imagine the existence of an international astronomical agency which monitored comet activity in adjacent galaxies. The agency had six observatories, and at each was stationed a team comprising one astronomer and one technician. Each observatory produced 48 observation reports in which the presence or absence of the team was recorded and whether its assigned comet had been detected or not (automated tracking equipment could detect the comet in the absence of the team). The subjects' task was to evaluate the effectiveness of the team in detecting their comet on the basis of these 48 reports. With six stations, the subjects were presented with 288 reports.

All names were fictitious and associations with actual astronomers or observatories were unintentional. The six observatories were "Paris", "Dallas", "Osaka", "Munich", "Florence", and "Dublin". The pairs of names for the different teams were "Prousot/Randieu", "Woods/North", "Kurasawa/Masuto", "Heinimann/Schmidt", "Ricci/Alighieri", and "Flannagan/Murphy". The comet names were "Cartwright's", "Plavka's", "Toulmin's", "Vespar's", "Herschell's", and "Minsky's". Although the nationality of the team was linked to the home observatory, the associations between the comet names and the observatories were arbitrary and as such the subjects had to learn them during the task.

Team-comet contingencies. The presence or absence of a team (which acted as the predictor variable) and the presence or absence of a comet detection (which acted as the outcome variable) created a 2×2 contingency table that defined the contingency between the team and the detection of its comet. There were four different contingency conditions. The first one, called Positive–Negative, is represented in Figure 1. For the first 24 reports a team was a good predictor of the detection of its comet, that is there were 10 comets detected on the 12 reports where the team was present or $P(\text{comet} \mid \text{team}) = .83$ and there were 2 comets detected on the 12 reports where the team was absent or $P(\text{comet} \mid \text{no team}) = .17$. The difference between these two conditional probabilities, or ΔP, equalled .67 (see Figure 1). For the next 24 reports the same team was a very poor predictor of the detection of its comet and became a much better predictor of the absence of detection: $P(\text{comet} \mid \text{team}) = .17$ and $P(\text{comet} \mid \text{team}) = .83$, for a contingency of –.67 (the cumulative contingency was 0). Such a positive–negative contingency created a situation where subjects acquired a positive expectation about the effectiveness of one variable (the team) in causing the occurrence of another (the comet detection), which would be confronted with negative evidence in the second half of the task.

This positive–negative contingency was programmed at three observatories to evaluate the role of the presence of an alternative explanation on the evaluation of evidence that contradicts an acquired belief. At the first observatory, the presence of a second team was programmed on all the cell C trials during Phase 2. That is, on the 10 trials in which the comet was detected but the original team was absent, a new team was presented.[2] This was called the Alternative Team condition.

[2]As the assignment of observatory and team labels were counterbalanced across condition, the names of the members of the alternative team were dependent on which of the 6 observatories was assigned to the Alternative Team condition for a given subject. The names of the alternative teams were "Lalande/Colet", "Curtis/Brody", "Suziko/Nakajima", "Knecht/Hesse", "Giardini/Renaldo", "O'Donnelly/Connor".

At the second observatory where the positive–negative contingency was programmed, the detection of a new comet occurred on all the cell B trials during Phase 2. Thus, on the 10 trials in the second half where the team was present but the original comet was not detected, a new comet was detected.[3] This condition was called Alternative Comet. At the third observatory where the positive–negative contingency was programmed, neither cell B nor cell C were coded in terms of the presence of a new team or a new comet; in that condition, called No Alternative, the negative evidence was not supplemented with additional information.

To mask the positive–negative shifts in these three conditions, subjects were also asked to monitor three foil conditions. In the first, the contingency was zero in the first phase, but was positive in the second; in the second and third foil conditions, the contingency was zero and negative in both phases of the task respectively. The positive–negative contingency was programmed at three observatories, the zero-positive, positive, and negative contingencies were programmed at each one of the remaining three observatories.

Procedure. Subjects were tested one at a time in an experimental cubicle. The instructions and the reports were presented on a computer monitor. After having read about the scenario, subjects were informed that they would see 48 reports about the performance of each of six teams of astronomers at detecting one of six comets, for a total of 288 reports. Reports from the six observatories were mixed together and the order of their presentation randomised for each subject. Two constraints operated on the randomisation process: (1) after each block of 72 trials, exactly 12 reports from each observatory had been presented and (2) of those 12 reports, 6 involved the presence of the team and 6 involved the absence of the team.

After each block of 72 trials, subjects would be asked to rate the effectiveness of each team in detecting its respective comet, using a scale from –100 to 100. The order in which subjects were prompted for their estimates was randomised after each block. A negative estimate meant that the comet was detected more often in the absence of the team than in its presence; a positive estimate meant that the comet was detected more often in the presence of the team than in its absence. After each block of 72 trials the subjects would be reminded of their previous estimate for each team. During Phase 2, the estimate screen for the Alternative Team condition would read: "At *Observatory*, *original team name*

[3]The alternative comet name was taken randomly from the following set: "Ashbourne's", "Collin's", "Phrebus's", "Gillespie's", "Grampton's".

received competition from the new team of *alternative team name*" where the italicised items refer to the labels used for that particular subject. For example, "At Dublin Observatory, Flannagan/Murphy received competition from the new team O'Connelly/Connor".

The 288 reports were presented one at a time on the computer screen. The report number was indicated in the middle of the top row. In the middle of the screen the labels "Observatory", "Astronomer" and "Technician" were printed under which the name of the observatory along with the names of the members of the team if they were present on that report (when they were absent ellipses were printed underneath the labels). In the bottom left of the screen appeared the names of the comets that could be detected, each paired with a number. To the right of the comet "menu", subjects were prompted to predict which comet would be detected on every trial by entering the appropriate comet number on the keyboard or "0" if they thought there would be no comet detection on that report. Prediction accuracy was stressed in the instructions: "Accuracy is important: If you feel that a comet will not be detected on a given trial then you can make a 'no detection prediction' by entering '0' on the keyboard". One second after the prediction was entered, feedback was presented below the subject's prediction. Feedback informed the subjects which comet if any had been detected on that report. Feedback remained on the screen until subjects depressed the spacebar for another report.

After the 144th trial, or midway through the task, the game paused and subjects were informed that at one observatory a new team would compete with the current team to detect the comet and at another observatory the current team would begin to monitor a second comet. To ensure that subjects would recognise the new team as the competing team, the new team's names were always printed in capital letters. As well, whenever a report concerned the observatory with the two competing teams a message was printed below the report number reminding the subjects that at that observatory two teams were competing to detect the same comet.

Measures. Subjects' evaluation of the evidence was measured first with the effectiveness ratings for each of the six team–comet relationships. Note that subjects were *not* asked to rate the effectiveness of the second team in the Alternative Team, nor were they asked to rate the relationship involving the new comet in the Alternative Comet condition. The strength of the subjects' contingency expectation was measured with the proportions of comet detection predictions on trials where a team was present.

Results

Effectiveness ratings. The top-left quadrant of Figure 2 shows the mean effectiveness ratings in the three Positive–Negative conditions. Ratings increased in the first half reflecting the positive contingency between the teams and comet detection but decreased in the second half, reflecting the negative contingency. After Block 2 the estimates did not differ ($F < 1$) nor did they differ after Block 4 [$F(2, 46) = 1.15$]. To assess the magnitude of adjustment as a function of the presence of an alternative contingency, the difference between the ratings made after Block 4 and the ratings made after Block 2 was calculated for each subject. The most pronounced adjustment was in the Alternative Team condition (mean of –58.3, Se = 15.6), followed by the adjustment in the No Alternative condi-

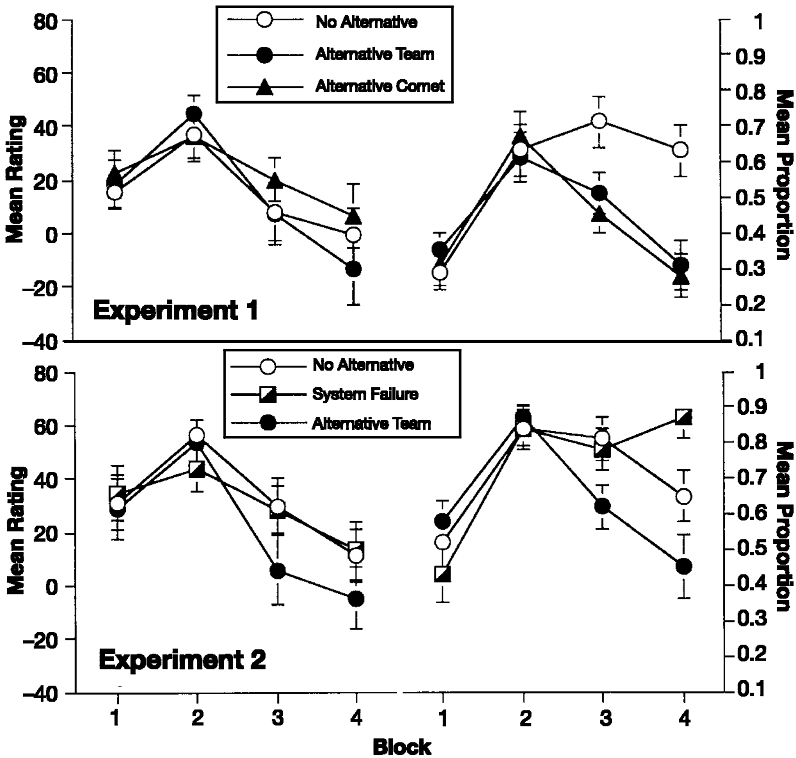

Figure 2. Mean effectiveness ratings (with their standard errors) of the predictor variable (team of astronomers) in the three Positive–Negative conditions of Experiments 1 and 2 (top- and bottom-left quadrants, respectively), and mean outcome (comet) prediction proportions (with their standard errors) in the same conditions (top- and bottom-right quadrants).

tion (–37.5, Se = 14.1), followed in turn by the adjustment in the Alternative Comet condition (–29.4, Se = 14.8). All mean adjustments were reliably smaller than zero [smallest reliable $t(23) = -1.99$ for the Alternative Comet condition]. However, a repeated-measures analysis of variance (ANOVA) on these mean differences was not reliable [$F(2, 46) = 2.18$]. Thus, although ratings were in line with the change in the nature of the original team–comet contingency, the magnitude of the adjustment was statistically equivalent in the three conditions.

Prediction responses. Mean comet prediction proportions in the positive–negative conditions are reported in the top-right quadrant of Figure 2. In both the Alternative team and Alternative Comet conditions the proportions of comet predictions in the presence of the team rose over the first half to decrease over the second half reflecting the nature of the changes in the number of comet detections (namely, comets were detected on 83% of the trials during which the team was present in the first half but detected only on 17% of those trials in the second half). However, in the No Alternative condition, comet detection predictions rose even during the third block when at that point the evidence had changed drastically; over the fourth block, the proportions returned to the level observed over Block 2, namely .63. The adjustment in the proportion of comet detection predictions was measured by taking the difference between the Block 4 and Block 2 proportions. There was no adjustment in the No Alternative (mean of 0, Se = .07) condition but in both conditions with alternatives, adjustment was substantial [mean of –30, Se = .11 in the Alternative Team condition and –.39, Se = .09 in the Alternative Comet condition]. The mean adjustments in both alternative conditions were reliably smaller than zero [$t(23) = -2.66$, $p < .05$ for the Alternative Team condition, and $t(23) = -4.49$, $p < .001$ for the Alternative Comet condition]. Furthermore, in a repeated-measures ANOVA the difference between the means of the three conditions was significant [$F(2, 46) = 9.63$, $p < .001$]. Individual mean comparisons revealed that the adjustment in the Alternative Team and Alternative Comet conditions were reliably more pronounced than in the No Alternative condition [$F(1, 23) = 9.14$, $p < .01$ and $F(1, 23) = 18.7$, $p < .001$ respectively] but adjustment in the two alternative conditions was not reliably different [$F(1, 23) = 1.02$].

Discussion

In this judgement task subjects monitored, concurrently, six predictor-outcome contingencies. Given the task demands, the subjects' ratings demonstrated that subjects could discriminate between different levels of contingency and could track changes in contingencies. These results support the growing contention (e.g., Shanks, 1995) that people are accurate judges of contingency, even in situations where they monitor multiple contingencies on the basis of trial-by-trial information.

In the three conditions where the evidence was positive in a first phase and then turned negative in a second phase, subjects adjusted their ratings in line with the evidence. Importantly, the availability of alternative explanations for the negative evidence did not enhance the magnitude of the adjustment in the ratings. The terminal ratings in these three conditions were in line with the cumulative zero contingency in these conditions.

In general, the proportions of comet detection predictions were also in line with the actual proportions. In the three Positive–Negative conditions the actual proportion of comet detections was .83 in the first phase but abruptly decreased to .17 in the second phase. In the No Alternative condition, subjects were very conservative in the adjustment of their prediction responses. Over the third block of trials, prediction responses were not adjusted at all (in fact the mean proportion *increased* over Block 3). Yet the effectiveness ratings in that condition were sharply adjusted after Block 3, demonstrating that subjects clearly recognised the nature of the negative evidence yet did not adjust their prediction responses. Over the fourth block, prediction responses were adjusted somewhat, although in the overall analysis they returned to the proportion recorded over the second block, where the evidence supported a positive contingency. Effectiveness ratings, however, were much lower after Block 4 than after Block 2. Thus, although effectiveness ratings in the No Alternative condition indicated a recognition of the negative evidence, subjects couched their predictions predominantly in terms of the presence of the outcome. This suggests that even if subjects' positive contingency expectation was disconfirmed by new evidence, and subjects acknowledged the disconfirmation, the expectation survived as a "heuristic structure" (Anderson, 1982) with which to generate new predictions.

In contrast, the pattern of prediction response adjustment in the Alternative Team condition paralleled the adjustment in effectiveness ratings. In that condition, the negative evidence originating from cell C of the contingency table, namely trials where the predictor (the team) was absent but the outcome occurred (a comet was detected), was presented in terms of an alternative team preceding the detection of the comet. Subjects could replace their original positive-contingency expectation with a new one involving a new team. The reduction in the number of outcome predictions in the presence of the predictor variable occurred, presumably, because the heuristic properties of the original expectation were greatly diminished by the replacement explanation.

The adjustment in the prediction responses recorded in the Alternative Comet condition, resulted mostly from response competition and as such is of lesser theoretical interest. In this condition, starting with Phase 2, a new comet was detected by the team. Subjects quickly learned to predict the occurrence of the new comet (35% over Block 3 and 62% over Block 4) and consequently their proportion of (original) comet detection predictions decreased over Phase 2. The presence of a competing response could have masked the true nature of the subjects' appreciation of the team's loss of diagnosticity, and thus it is unclear to what extent the adjustment in the prediction responses reflected a change in the nature of the contingency expectation. In fact the adjustment in the effectiveness ratings in the Alternative Comet condition was the lowest of the three Positive–Negative conditions.

To summarise the results of Experiment 1, in all three Positive–Negative conditions the increasing diagnostic value of the team was equally recognised during the first phase, as over 60% of the trials where the team was present, a comet detection was predicted. As the team lost its diagnostic value over the second phase, subjects readily adjusted their prediction responses in the Alternative Team condition (and in the Alternative Comet condition), but not in the No Alternative condition, even if in all three conditions the effectiveness ratings indicated a recognition of the negative evidence. Thus, the extent to which a positive contingency expectation was adjusted in light of disconfirming evidence was a function of the presence of a discrete alternative causal candidate, as offered in the Alternative Team condition. In the No Alternative condition subjects may have attributed the comet detections occurring in the absence of the original team of astronomers during the second phase to the automated tracking equipment. The fact that this hypothesis did not compel our subjects to abandon their original covariation expectation suggests that the constant presence of the automated equipment (on all trials, in all conditions) may have been treated as an "enabling condition" (Cheng & Novick, 1992) rather than as a discrete candidate cause that did not offer sufficient grounds for adjusting the causal importance of the original candidate cause (the team of astronomers).

A cognitive dilemma. Alloy and Tabachnick's (1984) "interactional model" specifies that in cases where prior beliefs and covariation evidence are diametrically opposed, reasoners experience a "cognitive dilemma" and generally seek to resolve this imbalance by making "covariation assessments biased in the direction of their initial expectation" (p. 116). On the other hand, if the opposing evidence is sufficiently salient "covariation assessment [will be] pulled in the direction of current situational information" (p. 116). Along with Alloy and Tabachnick, Brewer and Lambert (1993) argue that weak or ambiguous evidence is often filtered in a way that support the prior expectation (or "schema") but, in contrast with Alloy and Tabachnick, they argue that unambiguous evidence, regardless of the strength of the expectation, can rarely be denied or misperceived.[4] Billman, Bornstein, and Richards (1992) have also argued that prior beliefs do not affect the reasoner's ability to discriminate between different contingencies. The task used in Experiment 1 provided unambiguous evidence: Subjects could look at each trial for as long as they wanted and the presence and the absence of either the predictor or outcome variable was clearly identified. When in the No Alternative condition the new evidence was composed of a preponderance of cell B and cell C observations, subjects acknowledged the nature of the evidence as shown by the adjustment in their ratings. The covariation assessment was not distorted to favour the previously acquired expectation.

An advocate for Alloy and Tabachnick's (1984) interactional model might argue that the prior expectation in the No Alternative condition was simply not strong enough or that subjects were not sufficiently motivated (Kunda, 1990) to influence the "perception" of covariation (Alloy and Tabachnick use "covariation perception", "covariation inference", and "causal attribution" interchangeably). In which case the cognitive dilemma should be resolved by abandoning the expectation. However, the lack of adjustment in the prediction responses in the No Alternative condition indicates that subjects did not abandon the expectation, at least not to the extent dictated by their concurrent recognition of the disconfirming nature of the new evidence. These results then imply at least a third possible reaction to the cognitive dilemma, namely, embracing the *status quo*. When the dilemma is not resolved, the data is held in "abeyance" (cf. Chinn & Brewer, 1993). Unambiguous disconfirming evidence can be acknowledged but its probative force can remain unheeded. There need not be distortion of the evidence for a theory to survive exposure to contradictory data.

[4]In their classic analysis of belief persistence, Festinger, Riecken, and Schachter (1956/1964) assumed as a starting point that "people are in touch with reality", that they do not "blind themselves" to the fact that the predicted outcome has not occurred. Festinger et al.'s dissonance reduction analysis anticipates in many ways the *cognitive dilemma* analysis described by Alloy and Tabachnick (1984). Thus, it suggests that dissonance between disconfirming data and a belief can be reduced by aligning the belief with the data, and in an interesting prediction (confirmed in their naturalistic study) that departs from Alloy and Tabachnick's model, by seeking to enhance the plausibility of the belief by proselytising to other believers and to sceptics. As Festinger et al.'s study demonstrates, this occurs only within a sufficiently supportive network of people who experience the same dissonance.

EXPERIMENT 2

We have argued that the probative value of evidence is granted by a specific hypothesis and disconfirming data also needs to be anticipated by a hypothesis to be meaningful. Even if the logic of (naive) falsificationism dictates the rejection of a hypothesis upon the observation of a falsified prediction, the interface between a hypothesis and the experimental world usually consists of a network of auxiliary assumptions about the circumstances in which a hypothesis is tested (Lakatos, 1970). Therefore, a falsified prediction underdetermines which component(s) of the conjunction made of the focal hypothesis and the auxiliary assumptions should be rejected. However, in the face of disconfirmatory evidence, lay reasoners and scientists alike seek to explain the new data by examining the testing circumstances before questioning the validity of the focal hypothesis (Gorman, 1992; Kern, 1982; Tweney, 1985); the consequent adjustment to the negative evidence may thus be conservative or non-existent. This default scrutiny of the auxiliary assumptions is turned into a prescriptive principle by Pazzani and Flowers (1990, p. 416): "We claim that the general principle of scrutinizing evidence extremely carefully only when it contradicts an expectation is an important part of the discovery process." Similarly, Tweney (1990) proposed that computational models of scientific discovery should be built such as to distrust disconfirming data.

In Experiment 1 subjects in the Alternative Team condition were able to couch the new data as a competition between a new team that was performing very well and an original team that was performing very poorly. Belief persistence as measured by the adjustment in the prediction responses was reliably reduced in the Alternative Team condition compared to a condition without such alternative. In more "natural" settings, reasoners may question negative evidence by examining the auxiliary assumptions underlying the testing of the hypothesis. The perseveration in the No Alternative condition of Experiment 1 might have been the product of subjects entertaining such doubts (e.g., questioning the validity of the observation reports) but nothing in the scenario or in the way the trials were presented suggested that they could do so. Experiment 2 provided reasoners with the opportunity to discount negative evidence by indicating that feedback on some trials during the disconfirming phase might be erroneous. In this situation the positive contingency expectation might be maintained or even strengthened, as reflected by an increase in the outcome prediction responses during the second phase. As Gorman (1989, p. 387) wrote: "The possibility of error ... may allow scientists and ordinary people to form hypotheses or mental representations based on positive instances and retain them in the face of evidence that appears to contradict them." Thus, Experiment 2 introduced a Positive–Negative condition where during the negative-disconfirming phase, "system-failure error" (O'Connor, Doherty, & Tweney, 1989) occurred at that observatory such that the feedback about the success of the team at detecting its comet might be misleading. If subjects use the possibility of system failure to explain disconfirming evidence then this may lead to strong hypothesis perseveration during the second phase of the task.

Kern (1982) offered the first report on the effect of system-failure error on hypothesis revision. In her task, subjects (graduate students in physics, chemistry, and biology) sought to determine the location of the boundary that demarcated high and low moisture area where plants could survive on an alien planet. From the "planet's orbit" subjects could launch "probes" with plants on board, which landed on either side of the currently hypothesised moisture boundary (which subjects could adjust from trial to trial). The probes "radioed" feedback about whether the plants lived or died. On the basis of feedback subjects adjusted the moisture line. Subjects in certain conditions were told that system-failure error could occur, namely that 25% of the probes malfunctioned and the resulting feedback could be inaccurate. Subjects in these conditions were significantly less likely to adjust their hypothesis (as measured by moving the moisture line) than subjects in conditions which had no error in the feedback (or only measurement error, i.e., errors in the exact location where the plants landed). Kern called this phenomenon "inferential perseveration".

Gorman (1989), using Wason's (1960) 2–4–6 task, demonstrated that feedback error impeded the subjects' ability to induce the correct rule (the possibility of error also influenced performance in a modified version of Wason's task, and in a different induction task—Eleusis; Gorman, 1986). This occurred because when subjects encountered evidence that disconfirmed their hypotheses they invoked feedback error as an explanation of the recalcitrant data, thereby maintaining an incorrect hypothesis. In other words, feedback error enabled subjects to explain the disconfirmatory data as artefactual and thus having no probative value.

Experiment 2 of this study used the multiple-contingency learning task of Experiment 1. As in that experiment, there were three conditions that showed, in a first phase, strong evidence for a positive contingency between a predictor and an outcome variable, and in a second phase, strong evidence for a negative contingency. In two of these three conditions an explanation for the negative evidence was provided to the subjects. In one condition (the Alternative Team condition) the explanation was meant to accentuate the probative impact of the disconfirming evidence (by introducing a new team which detected the comet), whereas in the other (called the System-Failure condition) the explanation was meant to strip the disconfirming evidence of its probative impact. In a third condition (the No Alternative condition), the disconfirming evidence was not framed in terms of the presence of an alternative team. Adjustment in the prediction responses should be most mitigated in the System-Failure condition, followed by the No Alternative condition and the Alternative Team condition in turn. Effectiveness ratings, on the basis of our results in Experiment 1 should be similar in both the No Alternative and Alternative Team condition, but may lag behind in the System Failure condition if subjects discount a portion of the negative evidence.

Method

Subjects. Twenty-four undergraduates from the University of Hertfordshire participated in this experiment.

Task scenario. The same scenario as in Experiment 1 was used. Subjects were to evaluate the effectiveness of each one of six teams in detecting a comet on the basis of 48 observation reports. Subjects saw 288 reports in total.

Team–comet contingencies. The System-Failure, No Alternative, and Alternative Team conditions were programmed with the positive-negative contingency shown in Figure 1. One of the three foil contingencies was changed, namely the zero-positive foil contingency was replaced by a contingency that was positive in both phases.

Procedure. The same procedure as in Experiment 1 was used. After the 144th trial (or midway through the game), the screen cleared and subjects were informed that at one observatory a new team would compete with the current team to detect the comet, and they were informed about the possibility of feedback error at another observatory. Subjects read: "Expect possible error in the feedback reported from station [the name of the observatory would appear]. That is, the results from the observation made at this station may or may not be accurate. When a report appears suspect, you will be warned. Do not trust the feedback for that report." During the second phase when feedback error was programmed to occur on a given trial the message "Do not trust this report: there is feedback error" would appear in red at the top of the screen. During the second phase the estimate screen for the Alternative Team condition reminded subjects of the competition between the two teams (as in Experiment 1) and the estimate screen for the System-Failure condition reminded subjects about the feedback error in some reports.

There was a .5 probability that the feedback error message would be presented on trials where the presence of the team was paired with the absence of a comet detection. This type of trial (team present–comet detection absent) was chosen because it represents the most salient disconfirming evidence for a positive contingency expectation. The error message occurred on no other trials.

Measures. As in Experiment 1, the evaluation of the evidence was measured with ratings of the effectiveness of the different teams after each of the four blocks of trials and the consequent strength of the contingency expectation was measured with the proportions of comet detection predictions on trials where a team was present across the four blocks of trials.

Results

Due to a programming error, five subjects saw the feedback error message on some trials during Phase 1 in the System-Failure condition. Data for these subjects were removed from the sample. Consequently, all analyses were carried out on the remaining 19 subjects.[5]

Effectiveness ratings. The mean effectiveness ratings in the three Positive–Negative conditions mirrored closely the actual contingencies (see bottom-left quadrant of Figure 1): At the end of the second block of trials, mean ratings were all moderately positive, and did not differ from one another ($F < 1$). Adjustment, calculated as the mean difference in ratings between Blocks 4 and 2, was greatest in the Alternative Team condition (–58.4, Se = 16.8), followed by the No Alternative condition (–45.3, Se = 10.7), and by the System-Failure condition (–30.8, Se = 12.0). All mean adjustments were reliably smaller than zero [smallest reliable $t(18) = -2.58$ for the System-Failure condition], but were not reliably different from one another [$F(2, 36) = 1.1$].

Prediction responses. The mean proportion of comet detection predictions are shown in the bottom-right quadrant of Figure 2: They showed varying amounts of adjustment in the face of the negative evidence during the second phase. The means reported for the System-Failure condition for Phase 2 were calculated over the trials where the error message was *not* presented. We might expect that subjects would dismiss the trials where the feedback error was announced and that, consequently, a truer reflection of their contingency expectation and of the effect of sporadic

[5]Analyses carried out on the full sample revealed the same patterns as the ones reported with N = 19. The full sample data, however, did show lower (although not reliably lower) mean effectiveness ratings in the System-Failure condition after Blocks 1 and 2.

system failure might be measured with their prediction responses on the trials where feedback error did not occur.

During Phase 2 adjustment in the Alternative Team condition was swift; comet detections were predicted on 87% of the (original) team present trials over Block 2 but on only 45% over Block 4. In the No Alternative condition, adjustment was minimal over Block 3 (3% fewer comet detection predictions than during Block 2), but more pronounced over Block 4 where the mean proportion was .65. There was minimal adjustment in the System-Failure condition over Block 3 (6% fewer comet detection predictions than over Block 2) but the mean proportion of comet detection predictions on team present trials *increased* over Block 4 to reach .87. The mean difference in comet detection predictions between Block 4 and Block 2 was calculated. The mean adjustment was –0.19 (Se = 0.07) in the No Alternative condition, 0.03 (Se = 0.08) in the System-Failure condition, and –0.42 (Se = 0.1) in the Alternative Team condition. Only the mean adjustments in the System-Failure condition failed to differ from zero ($t < .001$). The magnitude of adjustment was reliably different across conditions [$F(2, 36) = 8.10$, $p < .001$]. Pairwise comparisons were treated as planned: All were reliable [smallest reliable $t(18) = -2.04$ for the Alternative Team–No Alternative comparison.

Discussion

The main purpose of Experiment 2 was to extend the test of the claim that evidence that disconfirms an acquired expectation is granted meaning by the explanations available to the reasoner, and consequently that its probative impact is a function of the nature of these explanations. Experiment 1 established that a replacement hypothesis accentuated the probative value of disconfirming evidence. Experiment 2 replicated this effect but also established that a hypothesis can diminish the impact of disconfirming evidence. Thus, in the System-Failure condition the possibility of explaining some of the negative evidence in terms of feedback error impeded the adjustment of the contingency expectation. In fact over the last block of trials in that condition, the proportion of outcome predictions in the presence of the predictor variable was the highest recorded proportion in Phase 2 of the Positive–Negative conditions in both Experiments 1 and 2. Yet, effectiveness ratings in the System-Failure condition were adjusted during the second phase, which indicates that subjects did not discount all the negative evidence; in fact the magnitude of adjustment in the ratings in the three Positive–Negative conditions was statistically equivalent. Thus in the face of identical disconfirming evidence, the availability of one explanation (namely, system failure) led to strong hypothesis perseveration, which was reduced if not eliminated by the availability of a different kind of explanation (namely, competition from an alternative team). The important new finding is the dissociation between the recognition of the disconfirming evidence and the adjustment of the contingency expectation, a dissociation that is mediated by the nature of the hypotheses available to interpret the new data.

As logicians, philosophers, and historians of science have documented, a vast, creative and resourceful repertoire of explanations can be brought to bear on the interpretation of new data in a real-world inductive inference problem. Experiments 1 and 2 sampled from this repertoire to demonstrate that the success of an alternative predictive cue at signalling the occurrence of a target effect can facilitate abandoning a discredited predictor, whereas explanations that discredit the evidence itself encourages reasoners to retain an acquired expectation in the face of negative evidence.

Finally, we should note that there was substantially more adjustment in the prediction responses in the No Alternative condition of Experiment 2 than in the one of Experiment 1. We suspect that the context of Experiment 2 facilitated adjustment generally by virtue of the contrast offered by the System-Failure condition (where there was no adjustment at all). Since the experimental design was within subjects, the level of adjustment in the System-Failure condition may have set a benchmark which subjects implicitly used to guide their own responses. In Experiment 1 that benchmark would have been provided by prediction responses in the No Alternative condition. In line with this possibility is the fact that Experiment 2 also fostered greater adjustment in the Alternative Team condition than did Experiment 1. Let us stress that the more important feature of our data is that in both Experiments 1 and 2, prediction response adjustment in the Alternative Team condition was significantly greater than in the No Alternative condition.

Open-mindedness and belief change. An interesting virtue of the within-subjects design employed in these experiments is that it reinforces the fruitfulness of an account of hypothesis perseveration and belief change in terms of the context of reasoning in contrast to accounts couched in terms of the impartiality or open-mindedness of reasoners. For example, Baron (1994) writes:

> Irrational belief persistence involves the failure to search impartially for evidence. (p. 282)
>
> A major mechanism of irrational persistence involves distortion of one's perception of what evidence would mean to an unbiased observer. (p. 289)
>
> The prescriptive policy to avoid [irrational belief persistence] is actively open-minded thinking. (p. 309)

These are, of course, well-intended prescriptions, and Baron also wants to draw attention to the well-documented tendency to seek evidence that supports a currently held hypothesis (e.g., Baron, Beattie, & Hershey, 1988; Beyth-Marom & Fischhoff, 1983; Klayman & Ha, 1987). But these prescriptions suggest that reasoning expertise is a skill independent of the domain or context in which reasoning takes place. In contrast we would argue that "open-mindedness" is a function of domain specific knowledge, that "impartiality" results only if the reasoner evaluates evidence from the perspective of competing but equally well-informed hypotheses, recreating the testing circumstances preached by Platt (1964).

A similar point can be made concerning Lord et al. (1984) "consider-the-opposite rule". They argue that reasoners would be more swayed by disconfirming evidence if they can consider that evidence from the perspective opposite their currently held beliefs. To do this, reasoners must be able to create an alternative hypothesis as rich as the one they currently believe and, as Tweney et al. (1980) demonstrated, this is something reasoners find very difficult (see also, Doherty, Mynatt, Tweney, & Schiavo, 1979). Thus, the prescription to consider the opposite will work only to the extent that the process of considering the opposite is informed by genuine alternative hypotheses.

We did not assess how "open-minded" our subjects were or whether they were familiar with Popper's philosophy of science (as some might have been) or whether they had training in formal logic. The point, however, is as a group our subjects could persevere in using a discredited expectation to generate new predictions and *at the same time*, and on the basis of ostensibly identical disconfirming evidence, abandon a similar contingency expectation. Their "open-mindedness" to the new disconfirming evidence was determined by the richness and the nature of the hypotheses that was used to interpret the evidence.

In summary, this study has demonstrated that the adjustment of a positive contingency expectation in the light of sustained disconfirmation was a function of the nature of the hypotheses with which that negative evidence was interpreted. In a context that did not offer a replacement hypothesis, adjustment was very conservative. In the presence of an explanation that invited reasoners to dismiss portions of the disconfirming evidence, namely feedback error, the expectation was not adjusted; that is, reasoners strongly perseverated in expecting a positive contingency. In the presence of a replacement explanation where the disconfirmation of one expectation simultaneously involved the affirmation of a new one, the strength of the original expectation was greatly reduced. This study also demonstrated that the unbiased recognition of disconfirming evidence and the ensuing change in the contingency expectation can be dissociated. In the experimental task employed in this study, the negative evidence was unambiguous and reasoners always recognised the changes in the nature of the evidence. The consequent adjustment of the belief was not as much a consequence of the characteristics of the reasoner as it was of the characteristics of the context of interpretation.

Manuscript received November 1998
Revised manuscript received May 1999

REFERENCES

Alloy, L.B., & Tabachnick, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91*, 112–149.

Anderson, C.A. (1982). Inoculation and counterexplanation: Debiasing techniques in the perseverance of social theories. *Social Cognition*, *1*, 126–149.

Baron, J. (1994). *Thinking and deciding* (2nd ed.). New York: Cambridge University Press.

Baron, J., Beattie, J., & Hershey, J.C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, *42*, 88–110.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology*, *45*, 1185–1195.

Billman, D., Bornstein, B., & Richards, J. (1992). Effect of expectancy on assessing covariation in data: "Prior belief" versus "meaning". *Organizational Behavior and Decision Processes*, *53*, 74–88.

Brewer, W.F., & Lambert, B.L. (1993). The theory-ladenness of observation: Evidence from cognitive psychology. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 254–259). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Chapman, L.J., & Chapman, J.P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*, 271–280.

Cheng, P.W., & Novick, L.R. (1992). Covariation in natural induction. *Psychological Review*, *99*, 365–382.

Chinn, C.A., & Brewer, W.F. (1993). Factors that influence how people respond to anomalous data. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 318–323). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Dawes, R.M. (1988). *Rational choice in an uncertain world*. City: Orlando, FL: Harcourt Brace Jovanovich.

Doherty, M.E., Mynatt, C.R., Tweney, R.D., & Schiavo, M.D. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 111–121.

Dougherty, M.R.P., Gettys, C.F., & Thomas, R.P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, *70*, 135–148.

Einhorn, H.J., & Hogarth, R.M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3–19.

Festinger, L., Riecken, H.W., & Schachter, S. (1964). *When prophecy fails*. New York: Harper Torchbooks. (Original work published 1956).

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239–260.

Gilovich, T. (1991) *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.

Gorman, M.E. (1986). How the possibility of error affects falsification on a task that models scientific problem solving. *British Journal of Psychology*, *77*, 85–96.

Gorman, M.E. (1989). Error, falsification and scientific inference: An experimental investigation. *Quarterly Journal of Experimental Psychology*, *41A*, 385–412.

Gorman, M.E. (1992). *Simulating science*. Bloomington, IN: Indiana University Press.

Gould, S.J. (1991). *Bully for brontosaurus*. New York: W.W. Norton.

Kern, L.H. (1982). *The effect of data error in inducing confirmatory inference strategies in scientific hypothesis testing*. Unpublished doctoral dissertation. Ohio State University.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Lakatos, I. (1970). Falsification and the methodology of scientific research programme. In I. Lakatos & A. Musgrave (Eds.), *Criticisms and the growth of knowledge* (pp. 91–196). Cambridge, UK: Cambridge University Press.

Lord, C.G., Lepper, M.R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgments. *Journal of Personality and Social Psychology*, *47*, 1231–1243.

Losee, J. (1993). *An historical introduction to the philosophy of science* (3rd ed.). Oxford, UK: Oxford University Press.

Nisbett, R.E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

O'Connor, R.M., Doherty, M.E., & Tweney, R.D. (1989). The effects of system failure error on predictions. *Organizational Behavior and Human Decision Processes*, *44*, 1–11.

Pazzani, M.J., & Flowers, M. (1990). Scientific discovery in the lay person. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 402–435). San Mateo, CA: Morgan Kaufmann Publishers.

Platt, J.R. (1964). Strong inference. *Science*, *146*, 347–353.

Popper, K.R. (1976). *Objective knowledge: An evolutionary approach*. Oxford, UK: Oxford University Press.

Ross, L., & Lepper, M.R. (1980). The perseverance of beliefs: Empirical and normative considerations. *New Directions for Methodology of Social and Behavioral Science*, *4*, 17–36.

Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. *Journal of Experimental Social Psychology*, *15*, 343–355.

Shanks, D.R. (1995). *The psychology of associative learning*. Cambridge, UK: Cambridge University Press.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology*, *16*, 583–588.

Trolier, T.K., & Hamilton, D.L. (1986). Variables influencing judgments of correlational relations. *Journal of Personality and Social Psychology*, *50*, 879–888.

Tweney, R.D. (1985). Faraday's discovery of induction: A cognitive approach. In D. Gooding & F.A.J.L. James (Eds.), *Faraday rediscovered* (pp. 189–209). New York: Stockton Press.

Tweney, R.D. (1990). Five questions for computationalists. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 471–484). San Mateo, CA: Morgan Kaufmann Publishers.

Tweney, R.D., Doherty, M.E., Worner, W.J., Pliske, D.B., Mynatt, C.R., Gross, K.A., & Arkkelin, D.L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, *32*, 109–123.

Vallée-Tourangeau, F., Hollingsworth, L., & Murphy, R.A. (1998). "Attentional bias" in correlation judgments: Smedslund (1963) revisited. *Scandinavian Journal of Psychology*, *39*, 221–233.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140.

Whewell, W. (1989). Novum organon renovatum (3rd ed.). In R.E. Butts (Ed.), *William Whewell, theory of scientific method*. Indianapolis, IN: Hackett Publishing Company, Inc. (Original work published 1858).