ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 56, 28–55 (1993)

The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality

JONATHAN J. KOEHLER

The University of Texas at Austin

This paper is concerned with the influence of scientists' prior beliefs on their judgments of evidence quality. A laboratory experiment using advanced graduate students in the sciences (study 1) and an experimental survey of practicing scientists on opposite sides of a controversial issue (study 2) revealed agreement effects. Research reports that agreed with scientists' prior beliefs were judged to be of higher quality than those that disagreed. In study 1, a prior belief strength × agreement interaction was found, indicating that the agreement effect was larger among scientists who held strong prior beliefs. In both studies, the agreement effect was larger for general, evaluative judgments (e.g., relevance, methodological quality, results clarity) than for more specific, analytical judgments (e.g., adequacy of randomization procedures). A Bayesian analysis indicates that the pattern of agreement effects found in these studies may be normatively defensible, although arguments against implementing a Bayesian approach to scientific judgment are also advanced. © 1993 Academic Press, Inc.

Much research has been conducted on the interplay between people's beliefs and their reactions to new evidence that bears upon those beliefs. It is often suggested that people fail to update their beliefs in normatively appropriate ways. Edwards (1968) showed that people are conservative information processors, failing to revise their prior beliefs to accommodate new information as much as required by Bayes's theorem. Pitz and his colleagues conducted a series of bookbag-and-poker-chip experiments

This article is based on a doctoral dissertation submitted to the University of Chicago. The research was supported by a Dissertation Improvement Grant SES-8710191 awarded by the Decision, Risk, and Management Science Program of the National Science Foundation, and a William Rainey Harper Fellowship awarded by the University of Chicago. I thank my dissertation advisor Joshua Klavman for his guidance and insightful contributions. I am also grateful to Scott Armstrong, Jon Baron, Ulf Bockenholt, Jeanne Enders, Brian Gibbs, William Goldstein, Steve Hoch, Robin Hogarth, George Loewenstein, Amy McCready, Jeff Rachlinski, Milton Rosenberg, Pat West, and Charlsetta Wren for comments on earlier versions of this paper and/or critical assistance with various aspects of this research. Bob Josephs deserves a unique thanks for pointing out the obvious. The members and associates of the parapsychological and skeptical groups who participated in the survey are also gratefully acknowledged. The groups and participants shall remain nameless here and in the text for reasons of confidentiality. Correspondence and reprint requests should be addressed to Jonathan J. Koehler, MSIS Department, CBA 5.202, University of Texas, Austin, TX 78712.

0749-5978/93 $5.00

to show that people's confidence in their beliefs is often unshaken by disconfirming evidence (Geller & Pitz, 1968; Pitz, Downing, & Reinhold, 1967). Similarly, Lord, Ross, and Lepper (1979) showed that strong beliefs are highly resistant to change even in the face of a thorough discrediting of their evidential basis.

In some instances, a reluctance to revise prior beliefs may be related to the difficulty people experience retrieving and accurately recalling disconfirming information (see e.g., Kleck & Wheaton, 1967; Koriat, Lichtenstein, & Fischhoff, 1980; Ross, McFarland, & Fletcher, 1981; Snyder & Uranowitz, 1978). However, this explanation is at best incomplete. Even where information recall is not a factor, people often ignore, underweigh, or reinterpret disconfirming evidence to agree with their beliefs (Batson, 1975; Festinger, Riecken, & Schachter, 1956; Kuhn, Amsel, & O'Loughlin, 1988; Nisbett & Ross, 1980; Pitz, 1969).

Ross and Lepper (1980) proposed that people respond to new evidence in a theory-biased manner. People's judgments about the probative value of evidence may depend, in part, on whether or not the outcomes or conclusions implicated by the evidence are congruent with one's personal beliefs. Thus, belief-confirming evidence may be regarded as relatively more probative than belief-disconfirming evidence for having yielded a belief-consistent answer.

A recent study on "outcome bias" provides independent support for this idea. Baron and Hershey (1988) showed that people take outcome information into account when judging the quality of various medical and monetary decisions. Decisions that were associated with successful outcomes received higher ratings than those associated with less successful outcomes.

THE SCIENTIST

If a dependence between outcomes and quality judgments is regarded as a bias, one might assume that some decision makers would be less susceptible to it than others. For example, it would seem less likely that professional scientists and others trained in the use of the scientific method would allow their beliefs about what constitutes a good or desirable outcome to affect their judgments about the quality of scientific research. This is because the classical model of science with which most scientists are familiar requires emotional neutrality, and unbiased observation and interpretation of phenomena (Merton, 1942/1973; Scheffler, 1967). The expectations, attitudes and desires of individual scientists should not, and presumably do not, affect their judgments and decisions.

However, some empirical evidence suggests that this model does not accurately describe scientific conduct. Ian Mitroff (1983) conducted a series of detailed interviews with 42 eminent Apollo moon scientists and reported that most were emotionally involved in their work. Furthermore, those who held very strong beliefs about the nature of the moon appeared most anxious to dismiss evidence that contradicted their personal theories. Similarly, Mahoney (1977) studied a group of 75 scientific journal reviewers and found that they were strongly biased against manuscripts that reported results contrary to their strong behaviorist perspective. In short, judgments about the quality of scientific research appear to be quite dependent on the fit between a scientist's own beliefs and the conclusions supported by the research, particularly when the beliefs are strongly held.

THE MODEL

The present study explores the generality of Mitroff's and Mahoney's findings, and tests a descriptive model of evidence quality judgments among scientists. It is hypothesized that these judgments will deviate from some neutral standard of quality by the extent to which the evidence meets or opposes the individual scientist's expectations and the strength with which the expectations are held. Evidence such as a scientific study that agrees with prior beliefs will receive higher quality ratings than would be given by neutral scientific observers, and evidence that disagrees with beliefs will receive lower ratings than that given by neutral observers. In addition, the magnitude of the agreement effect will be positively related to the strength of the prior beliefs. That is, scientists who have strong prior beliefs will rate belief-confirming studies higher and belief-disconfirming studies lower than will scientists who have weaker prior beliefs.

These predictions are captured in the following model: $Jq = Nq + AP$, where Jq = judgments of evidence or study quality, Nq = quality as judged by neutral observers who are aware of the design but not the results of the study, A = agreement of the data with prior belief, and P = prior belief strength. The model is paramorphic in the sense that it is designed to predict scientists' judgments, but no claim is made that it is descriptive of the psychological process scientists use to arrive at their judgments (e.g., anchoring and adjustment).

As Fig. 1 shows, the model predicts that scientists who hold relatively weak prior beliefs will make quality judgments that differ only slightly from those of neutral judges. But as prior belief strength increases, the multiplicative term of the model increases and judgments are expected to depart from those of neutral judges.

THE NORMATIVE ISSUE

Mahoney (1976, 1977, 1979), Baron and Hershey (1988) and Lord et al. (1979) are among the few who address the normative status of an effect of prior beliefs on judgments about the quality of evidence that bears upon

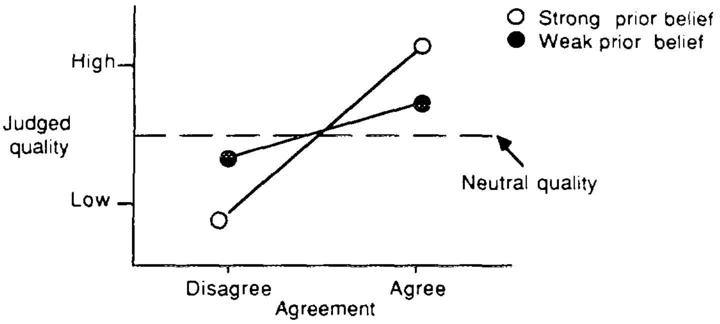

FIG. 1. Ordinal predictions for scientific judgments of evidence quality as a function of neutral quality (Nq), agreement (A), and strength of prior belief (P) as given by the model Jq = Nq + AP.

those beliefs. Mahoney (1977) finds such dependencies to be among "the most pernicious and counterproductive elements in the social sciences" (p. 173). Lord et al. (1979) disagree and argue as follows:

> [T]here can be no real quarrel with a willingness to infer that studies supporting one's theory-based expectations are more probative than, or methodologically superior to, studies that contradict one's expectations. When an objective truth is known or strongly assumed, then studies whose outcomes reflect that truth may reasonably be given greater credence than studies whose outcomes fail to reflect that truth. Hence the physicist would be "biased" but appropriately so, if a new procedure for evaluating the speed of light were accepted if it gave the "right answer" but rejected if it gave the wrong answer (p. 2106).

Lord et al. contend that it is not only understandable, but appropriate, for prior beliefs to influence judgments of evidential diagnosticity. Approached from a Bayesian perspective, prior beliefs constitute useful information which, in combination with other evidentiary components (obtained in part from a careful reading of other elements of the study), *should* affect judgments of quality. Moreover, because these quality judgments are closely related to Bayesian likelihood values, the Lord et al. recommendation is—in Bayesian terms—equivalent to allowing prior odds to influence likelihood ratios.

In the context of scientific belief updating, Bayes's theorem indicates that the change in one's beliefs given the results of a study should depend upon the likelihood ratio, i.e., the probability of observing those results if the beliefs are true, $P(D|H)$, relative to the probability of observing those results if the beliefs are false, $P(D|\text{-}H)$. Following the suggestion of Lord et al., a decision maker faces the dual task of using the likelihood ratio $P(D|H)|P(D|\text{-}H)$ to revise prior beliefs, *and* using prior beliefs to estimate the likelihood ratio (or a primary component thereof). For example, implausible results serve both as evidence against current beliefs, *and* evidence that the outcome obtained using this method may not be closely linked to the truth. The Bayesian analysis discussed below and detailed in Koehler (1989) supports this conclusion.

Agreement Effects

The analysis begins with several simplifying assumptions. First, the issue under consideration is dichotomous, such as whether or not a proposition is true or which of two measurements is greater. Second, the study produces results that are consistent with one side of the issue and inconsistent with the other. Third, the study's method is prone to error, although there is no a priori reason to believe that the method is biased in favor of one side or the other. These characteristics—which are reasonable in many cases—allow for a relatively simple analysis of how the influence of a scientific study's results on quality judgments can be represented within a Bayesian framework.

Prior to being exposed to a study, the scientist believes that one of two competing hypotheses, H or -H, is more likely to be true. If the scientist favors H and subsequently learns that the study's results favor H, he or she estimates, in effect, $P(G|A)$, the probability that the study is good given that it agrees with his or her prior beliefs. If the study's results favor -H, the scientist estimates $P(G|\text{-}A)$, the probability that the study is good given that it disagrees with his or her prior beliefs. The question concerning the effect of outcome on judgments of quality may now be restated as follows: Under what conditions, if any, should $P(G|A)$ be greater than $P(G|\text{-}A)$, and by how much?

Scientists and others who evaluate studies may have a set of prior beliefs that pertain to three propositions: A = "This study gives results that agree with my hypothesis or prior belief"; T = "This study gives results that are congruent with the true state of nature"; G = "This is a good quality study." For simplicity once again, A, T, and G may be treated as dichotomous; results either agree or disagree with prior beliefs, results are either congruent or incongruent with the true state of nature, and a study either does or does not meet the chosen criterion for being a "good" study. Four "primitive prior beliefs" derived from these propositions are as follows: $P_1 = P(A|T)$, $P_2 = P(T|G)$, $P_3 = P(T|\text{-}G)$, and $P_4 = P(G)$.

P_1 is the probability that a study that yields true results will agree with one's prior beliefs. For unbiased studies (i.e., studies in which the probability of obtaining the correct result does not depend on which of two exhaustive and mutually exclusive hypotheses is true), $P(A|T)$ is identical to one's degree of belief in one's hypothesis, $P(H)$. A proof is given in

Appendix A.[1] Further, because H is the hypothesis believed to be more likely a priori, P(H), and hence P_1, is greater than .5. Neutral belief is indicated by $P_1 = .5$.

P_2 and P_3 are the conditional probabilities of truth given good and bad quality studies respectively. Good quality studies are those that meet or exceed some standard of quality, while bad quality studies do not. A standard of quality may be defined any way at all provided that good quality studies are more likely to get true results than are bad quality studies. Although the dichotomization of a quality standard may at first seem artificial, the practical consequences of quality judgments often are dichotomous: Grants are awarded or denied, manuscripts submitted for publication are accepted or rejected, etc.

Of course, the likelihood that a study will yield accurate information or true results depends on a multitude of factors, including the difficulty of getting at the truth. In some domains, even high quality studies may often produce false results, while in other domains it may be so easy to produce true results that even very poor quality studies can arrive at basic truths. Thus, P_2 and P_3 range from 0 to 1.

P_4 is one's prior belief (i.e., before learning the results) that the study in question meets a good quality standard. P_4 also ranges from 0 to 1.

The details of this Bayesian analysis are presented in Appendix B. It shows that the probability that a study is good quality is higher when the study has produced results that agree with the judge's prior beliefs than when the study has produced results that disagree with the judge's beliefs (i.e., $P(G|A)$ is always greater than $P(G|-A)$). It further shows that the magnitude of this agreement effect increases as the strength with which a prior belief is held increases. In other words, this normative analysis is consistent with the descriptive model proposed earlier in which the outcomes of a study influence judgments about its quality, and do so to the extent that the judge's prior beliefs are strong (i.e., near 0 or 1).

Judgments about Specific Study Features

Quality judgments may be made both about very specific "analytical" features of a study and its more global "evaluative" features (Einhorn & Koelb, 1982). Thus, one may judge the adequacy of a randomization technique as well as a study's overall methodological quality. However,

[1] One way to intuitively grasp $P(H) = P(A|T)$ is to consider that as prior belief increases, one becomes more sure that one's own belief is true, i.e., that true results would agree with one's prior belief. Thus, assume that a scientist has some degree of belief in a hypothesis H, captured by P(H). Now suppose that a study relevant to H is conducted that is guaranteed to produce true results. What should be the scientist's probability that the results of this study will agree with his or her belief? In unbiased cases, this probability can only be P(H).

the normative status of a link between beliefs and quality judgments is more questionable when the judgments are made about specific features of the study rather than its overall quality.

Consider, for example, a scientist who is asked to make general, evaluative judgments about a study that gives implausible results. The scientist may reason as follows: "Something is probably wrong with this study somewhere. The problem may lie in one or more of the specific features of the study that I am explicitly called upon to evaluate, one or more features that I consider but am not explicitly called upon to evaluate, or one or more features of the study that I didn't even think to consider." Notice, then, that there are at least three major sources of potential flaws in a study that gives surprising results.

When judges are asked to evaluate a specific feature of such a study, similar, but more restrictive, avenues may be explored for potential flaws. Here the judge may infer that the feature is flawed either for detectable or undetectable reasons. However, there is less reason to believe that any *particular* feature of a study is flawed than there is to believe that flaws must exist in at least *some* features. Moreover, to the extent that a presumedly flawed study has many features, the likelihood that any single feature is flawed becomes small (provided it is assumed that the number of flaws in a flawed study is not directly proportional to its number of features). Consequently, the influence of a study's results on scientific judgments of quality might reasonably be smaller when the judgments concern particular features of the study when they are more general in scope.

Self-Perceptions and Normative Views

Regardless of whether analytical and evaluative judgments of quality are or should be influenced by prior beliefs, it is interesting to explore scientists' intuitions about these effects. Nisbett and Wilson (1977) have argued that we do not have complete access to our cognitive processes and, consequently, that we are unaware of factors that influence our judgments and actions. In a similar vein, it is predicted that scientists in the present study will deny that their evaluations of scientific studies were influenced by the reported results. It is further predicted that because scientists generally believe in the classical scientific model (Mulkay & Gilbert, 1982), they will maintain that the results of scientific studies *should not* influence judgments about the quality of the studies.

GENERAL APPROACH

This research employed two methodologies. First, a laboratory experiment was conducted using advanced graduate students in the sciences. Subjects evaluated the quality of studies that agreed or disagreed with their induced expectations about each of two fictitious scientific controversies. Second, two groups of practicing scientists having strong, opposite beliefs about a scientific issue were surveyed and asked to make quality judgments about a hypothetical, but representative research report that either supported or opposed their beliefs.

STUDY 1

Method

Subjects. The subjects were 297 science graduate students at the University of Chicago. Subjects completed an average of 2.3 years of graduate study and were recruited through advertisements in campus newspapers. The task lasted an hour and each subject was paid $10.

Fifty-four percent of the subjects were trained in the natural sciences, 36% were trained in the social sciences, and 10% were trained in other scientific areas. Fifty-four percent of the subjects reported having taken courses in scientific methodology. These demographic variables did not significantly influence the observed pattern of responses.

Materials and procedure. The stimuli consisted of a 20- to 35-page booklet that detailed two fictitious scientific issues. The issues concerned the existence of heat-sensing organs in the fictitious Canadian Stripeneck bird and an electromagnetic ray called a K-ray. The booklets contained background information on the issues, one detailed experimental research report related to each issue, and a series of questions about the reports.

Subjects were randomly assigned to one of five groups, four experimental and one neutral group. The experimental groups differed in induced strength of prior belief (strong, weak) and research report quality (high, low).

Subjects in the experimental groups read two-page summaries about each issue prior to reading the research reports. The purpose of the summaries was to induce a prior belief that the Stripeneck and K-ray hypotheses were either correct or incorrect. The summaries given to the strong prior belief groups were more detailed and provided stronger arguments than those given to the weak prior belief groups. After reading the summaries, subjects indicated their degree of belief in the hypotheses on a 0–100 scale in which the endpoints were labeled "very unlikely" and "very likely." A manipulation check showed that subjects in the strong prior belief groups held more extreme beliefs than did subjects in the weak prior belief groups on both sides of the hypotheses, $p < .001$ in each (pro-hypothesis: $M_{\text{strong}} = 81.5$, $M_{\text{weak}} = 66.6$; con-hypothesis: $M_{\text{strong}} = 16.8$, $M_{\text{weak}} = 28.4$).

After indicating their beliefs, subjects were presented with detailed experimental research reports related to the Stripeneck and K-ray issues.

Subjects received either two high quality or two low quality research reports. The high and low quality reports were similar in content and length, although their methodologies differed in several important respects. For example, the high quality Stripeneck report employed trained scientists and sophisticated bird tracking instrumentation; the low quality report involved untrained volunteers who tracked the birds with imprecise equipment.

Results and discussion sections were appended to each report such that subjects' induced beliefs were supported in one issue, but opposed in the other. After studying these reports, subjects made six analytical (i.e., specific) and three evaluative (i.e., general) judgments about each report using seven point Likert-type scales. For instance, analytical judgments were made about the appropriateness of conducting the Stripeneck tracking study during the winter, and the sufficiency of the number of angles from which the K-ray was measured. Evaluative judgments were made about the relevance, methodological quality, and clarity of the research reports.

Next, subjects completed a demographics form and answered a series of questions about whether their judgments were and should have been influenced by the results of the reports.

Subjects in the neutral group were not provided with summary pages or the results of the research reports they evaluated. Thus, there was no manipulation of agreement or disagreement with prior beliefs within this group. The evaluative question pertaining to the clarity of the results section was, of course, omitted for this group.

Design. Issue (Stripeneck, K-ray) and agreement (results agree/disagree with induced beliefs) were within-subjects variables for the four experimental groups. Issue order and agreement order were assigned at random. Prior belief strength (strong, weak), and research report quality (high, low) were the primary between-subjects variables.

Prior belief manipulation and analyses. Subjects' prior beliefs were successfully manipulated about two-thirds of the time, where a successful manipulation is defined as a probabilistic report by subjects (prior to reading the research report) that they held a belief in the direction favored by the summary pages. Successful belief manipulations were nearly twice as common in the strong prior belief groups as in the weak prior belief groups. The fact that it was more difficult to induce weak directional beliefs is not surprising because even small amounts of random variation near the 50% belief level are likely to result in crossovers to the other side.

One consequence of the difficulties surrounding the manipulation of prior beliefs in this experiment is that not all subjects received one study that agreed with their induced beliefs, and one that disagreed. Some may have received studies that agreed or disagree with their prior beliefs about both issues. Therefore two analyses were conducted. One analysis was conducted on the data from subjects whose beliefs were successfully manipulated on both issues. In this analysis, agreement was treated as a within-subjects variable as planned. A second analysis was conducted on data from the first issue evaluated by all 297 subjects. In this analysis, agreement and issue were treated as between-subjects variables. These analyses produced extremely similar results. The data below reflect the results of the second analysis. Noteworthy discrepancies in the significance levels found in the two analyses are given in footnotes 2 and 3.

Results

Separate 2 × 2 × 2 repeated measures MANOVAs were performed on the evaluative and analytical judgments. Composite indexes of evaluative and analytical quality were obtained by taking the mean of the individual judgments.

Evaluative judgments. As predicted, an agreement main effect, $F(1,280) = 10.51$, $p < .001$, and a prior belief strength × agreement interaction $F(1,280) = 6.53$, $p < .02$ were found. Subjects gave higher ratings to reports that agreed with their prior beliefs than to those that disagreed ($M_{\text{agree}} = 5.1$ ($n = 146$), $M_{\text{disagree}} = 4.6$ ($n = 150$)). The agreement effect was stronger among subjects who were induced to hold strong prior beliefs.

An unexpected main effect for prior belief strength was also found, $F(1,280) = 5.54$, $p < .02$). Subjects holding strong prior beliefs tended to rate the quality of the research reports higher than did those holding weaker beliefs. It may be that there is a true main effect for strength of prior belief (or a correlate of prior belief such as concern or interest) in which people with strong beliefs on a topic feel more positively disposed toward studies on that topic, regardless of agreement. More likely, perhaps, is that this effect is due to the significant interaction term. As Fig. 2 indicates, subjects in the strong prior belief/agree group gave significantly higher quality judgments to the reports than did subjects in the other groups, including the neutral group ($p < .005$ for all). None of the other groups differed significantly from one another.

Analytical judgments. A marginally significant main effect for agreement was found, $F(1,262) = 2.82$, $p < .10$.[2] Subjects gave higher ratings to reports that agreed with their prior beliefs than to those that disagreed ($M_{\text{agree}} = 4.7$, $M_{\text{disagree}} = 4.5$). However, a prior belief strength × agreement interaction was not found. A series of planned contrasts failed to

[2] A within-subjects analysis on the smaller successfully manipulated data set revealed a stronger main effect for agreement ($F(1,100) = 5.77$, $p < .02$, $M_{\text{agree}} = 4.7$, $M_{\text{disagree}} = 4.4$).

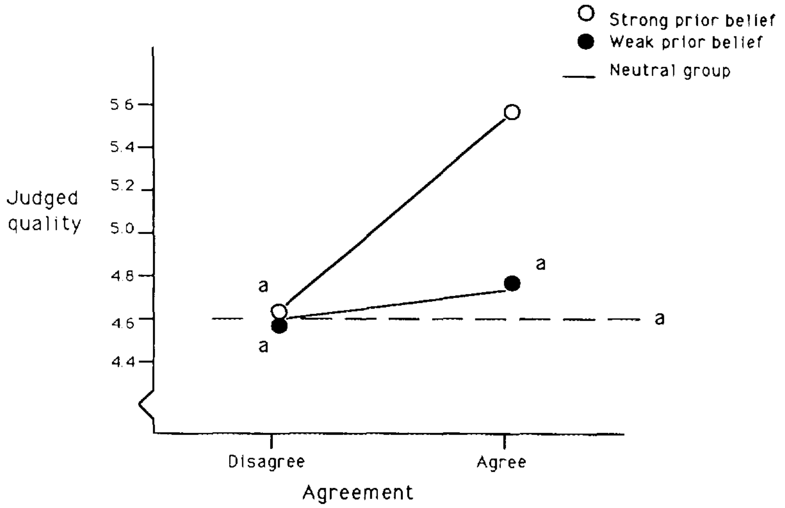

FIG. 2. Study 1 (Experiment). Mean judgments of quality for the evaluative questions as a function of agreement and prior belief strength. (a) Differs from strong prior belief/agree group, $p < .005$ (Bonferroni corrected).

show differences between the analytical judgments made by subjects in the neutral group and those made by subjects in any of the other groups. It is therefore difficult to know whether the agreement effect on the analytical questions is due to inflated ratings in the agree cells, deflated ratings in the disagree cells, or both.

A separate 2 × 2 within-subjects MANOVA was performed with judgment type (evaluative, analytical) and agreement (agree, disagree) to test the prediction that the agreement effect would be smaller for analytical judgments than for evaluative judgments. The judgment type × agreement interaction was marginally significant, $F(1,276) = 2.90$, $p < .09$, indicating that the agreement effect was slightly larger on the evaluative questions than on the analytical questions (evaluative-agree $M = 5.0$, evaluative-disagree $M = 4.6$, analytical-agree $M = 4.7$, analytical-disagree $M = 4.5$).[3]

Self-perceptions and normative views. Sixty-four percent of the subjects felt that their assessments of the methodological quality of the Stripeneck and K-ray studies *were not* influenced by the direction of the results yielded by the studies; 26% felt that their judgments were influenced, and 10% were not sure. A large majority (83%) felt that a

[3] A within-subjects analysis on the successfully manipulated data set revealed a stronger judgment type × agreement interaction, $F(1,107) = 4.30, p < .04$ (evaluative-agree $M = 5.1$, evaluative-disagree $M = 4.5$, analytical-agree $M = 4.7$, analytical-disagree $M = 4.4$).

scientist's assessment of the methodological quality of a study *should not* depend upon the direction of the results yielded by the study; 12% felt that this dependence should exist, and 5% were not sure.

Members of the various experimental groups responded similarly to these questions. There were no significant question framing effects (i.e., positive versus negative frames) and subjects' responses to the self-perception questions were not related to the size of the agreement effects: Those who believed that the outcome of the study did not influence their quality judgments were actually influenced by the outcome as much as those who admitted some probable influence.

Discussion

The results of the laboratory experiment support the presence of an agreement effect in scientific judgments of evidence quality. Studies giving results that agreed with subjects' prior beliefs were judged to be higher quality than studies that disagreed. The agreement effect was more pronounced when prior beliefs were strong and, to a lesser extent, when the judgments concerned evaluative rather than analytical aspects of the reports.

Contrary to what might be expected, the agreement effects were not the result of very critical or harsh judgments by those whose prior beliefs were disconfirmed by the research reports. Instead, these effects were driven primarily by subjects in the strong prior belief/agree groups. These subjects gave much more favorable ratings to the research reports than did subjects in the other groups.

One explanation for the observed leniency in the agree conditions is that scientists may differentially scrutinize studies that yield belief-congruent and belief-incongruent results. Studies that are known to have yielded belief-congruent data may be examined less carefully for having obtained the "correct" result, and may be presumed to have been conducted properly. On the other hand, when scientists evaluate studies that are not known to have produced "correct" results—as when evaluating a study in which the results are either unknown or not in line with expectations—their suspicion that something may be wrong with the study is heightened. From a Bayesian perspective, such behavior may be justifiable.

Subjects' responses to the self-perception and normative questions were intriguing. By stating that their beliefs *did not* influence their quality judgments (when the agreement effect suggests otherwise) and by suggesting that scientific judgments of quality *should not* be influenced by the judge's prior beliefs, many subjects unwittingly violated self-imposed normative canons. But, at least in the evaluative domain, Bayes suggests that these subjects may have accidentally landed on normatively high ground.

The somewhat smaller agreement effects in the analytical domain provide further evidence that the scientists made judgments in normatively appropriate ways.

STUDY 2

The laboratory experiment showed that scientific judgments of research quality are influenced by whether the research confirms or disconfirms prior beliefs, and by the strength with which the prior beliefs are held. While the experiment was useful for providing a controlled environment in which to test the model, the issues and studies were artificial, the subjects had little or no knowledge of the relevant areas, and their level of involvement probably was lower than that of practicing scientists in their areas of expertise. To addresses these issues, an experimental survey of practicing scientists on both sides of the extrasensory perception (ESP) controversy was conducted.

The ESP controversy is an ideal forum for studying the influence of scientists' prior beliefs on their judgments of evidence quality. First, there are relatively large numbers of scientists on both sides of the issue, many of whom have strong views about paranormal phenomena. Recently, a pair of target articles on the state of parapsychology in *Behavioral and Brain Sciences* (Alcock, 1987; Rao & Palmer, 1987) attracted dozens of spirited responses from interested scientists. Many of these scientists are members or affiliates of parapsychological or "skeptical" organizations. Some of these organizations hold regular conferences, investigate paranormal claims, and publish magazines or journals in which parapsychological studies are reported. The scientists who participated in this survey were solicited from the membership and consultant lists of several of these organizations. Each scientist was asked to evaluate a hypothetical (but representative) ESP study that either agreed or disagreed with his or her prior beliefs.

It was predicted that the survey results would be similar to those of the laboratory experiment. Studies that confirmed scientists' prior beliefs about ESP would be rated higher than disconfirming studies. In addition, there was some interest in investigating the relative size of the agreement effects in these two scientific communities. Skeptics frequently accuse ESP believers of being insufficiently critical of evidence that supports their positions while dismissing disconfirming evidence (Bauslaugh, 1981; Hansel, 1980; Marks, 1986). Parapsychologists counter that it is the skeptics who prejudge the quality of ESP research (Child, 1985; Kelly, 1979; Rao, 1979; Rockwell et al., 1978). An examination of the group × agreement interaction may shed some light on this controversy.

Method

Subjects. Surveys were mailed to 195 parapsychologists and others belonging to various professional organizations in the field, and to 131 scientists and consultants affiliated with various skeptical organizations.[4] Postcard follow-ups were sent 2 weeks after the initial mailing. Seventy-five parapsychologists (38%) and 39 skeptics (30%) completed and returned the survey in time for the data to be included in the analyses (within 6 months of the original mailing).

Materials. The survey packets included the following materials: a cover letter that briefly outlined the purpose and nature of the survey, a hypothetical parapsychological research report (approximately 7 single-spaced pages), a set of evaluative and analytical Likert-type questions about the report, open-ended questions about the strengths and weaknesses of the report, questions about whether the outcomes of the study did and should influence judgments of quality, and a series of demographic questions. The packets also included a stamped addressed envelope in which to return the completed survey.

The research report was modeled after published parapsychological studies that employ "ganzfeld" methodology, a popular and well-known parapsychological technique (Honorton, 1985; Rao & Palmer, 1987).[5]

Procedure. Six versions of the reports were prepared: (1) high quality, positive results; (2) high quality, negative results; (3) low quality, positive results; (4) low quality, negative results; (5) high quality, no results; (6) low quality, no results. The parapsychologists received one of the six versions of the report, while the skeptics, fewer in number, received one of the first four versions only.

The low quality report was similar to the high quality report in content and length, although the latter included tighter controls. For example, two experimenters were used in the high quality study to insure that ESP "senders" and "receivers" were monitored at all times. Longer distances and closed doors also separated senders and receivers in the high quality version. High and low quality studies also differed in the size of the pool

[4] For convenience, the members and affiliates of the parapsychological and skeptical groups will henceforth be referred to as parapsychologists and skeptics, respectively.

[5] Ganzfeld stimulation involves placing a subject (or "percipient") in a condition of reduced sensory stimulation in preparation for receiving extra-sensory impressions. This is typically accomplished by placing the subject in a comfortable chair, covering his or her eyes with halved Ping-Pong balls beneath a uniform white light, and piping "white" or "pink" noise into his or her ears through headphones. A subject typically undergoes thirty minutes of ganzfeld stimulation, during which time he or she reports all images, impressions and feelings. At some point during the ganzfeld, a sending period takes place, in which an agent views a target (in a distant room) and attempts to convey impressions of the target to the subject through nonsensory means. At the conclusion of the sending period, the subject is removed from the ganzfeld, and asked to rank order several potential target pictures in terms of their correspondence with his or her ganzfeld imagery. For a general review, see Rao and Palmer (1987).

from which target objects were chosen (one hundred vs four), the manner in which targets were assigned (pseudo-random vs random), and the procedures used to judge the data (blind judges were used in the high quality study only).

Parapsychologists who received research reports that lacked results were in the neutral group. The neutral group's materials were different in several other respects. First, the cover letter referred to the empirical study not as a "hypothetical, parapsychological study" but as a "hypothetical parapsychological research proposal" to explain the missing results and discussion sections. Second, as in the laboratory experiment, neutral subjects were not asked questions that directed attention to the missing results section.

After reading the report, the scientists made six analytical and three evaluative judgments about it using seven point Likert-type scales. The evaluative questions were identical to those used in the laboratory experiment, while the analytical questions were content-specific.

The scientists were encouraged to discuss (in writing) strengths and weaknesses of the reports and to append comments to their evaluative and analytical judgments. It was predicted that the ratio of positive to negative comments would be higher among those evaluating reports that confirmed their beliefs about ESP than among those evaluating disconfirming reports.

Finally, respondents completed a demographics page and answered questions about whether the outcomes of the report did and should influence their quality judgments. As in the laboratory experiment, it was predicted that the scientists would deny the influence of expectations on their judgments and regard such influence to be improper.

Design

The design of the survey was simpler than that of the experiment because the scientists evaluated only a single study. Thus, agreement (agree, disagree), study quality (high, low) and group (believers, skeptics) were between-subjects variables. Notice that agreement depends upon the outcome of the study (positive or negative) and the respondent's group membership.

Return rates and demographics

The return rates and demographics are summarized in Table 1.

Nearly 40% of the parapsychologists and 30% of the skeptics completed and returned the survey within six months. These return rates are not significantly different, $z = 1.61$. The large time commitment required to read and evaluate the reports (approximately 1 h), lack of incentives, and

TABLE 1
DEMOGRAPHICS OF SURVEY RESPONDENTS

Item	Parapsychologists	Skeptics
Number surveyed	195	131
Response rate	38.5%	29.8%
Ph.D.	64%	81%
Discipline		
Social science	55%	30%
Physical science	29%	59%
Other	16%	11%
Methodological training	72%	55%
Age (median range)	50–59	40–49
Parapsychological interest (years)	24	19
Published articles in area (median)	18	5
ESP Belief		
Believer	71%	0%
Leaning toward belief	25%	0%
Leaning against belief	4%	14%
Disbeliever	0%	86%

the fact that 25–30% of those surveyed lived outside of North America may help explain the relatively low return rates.

Responses to questions about belief in ESP confirm that a strong, fundamental disagreement exists between parapsychologists and skeptics. Large majorities of both the parapsychological and skeptical groups described themselves as strong believers and disbelievers in ESP respectively.

Respondents from both groups had strong educational backgrounds. Large majorities held Ph.D.s or the equivalent, and most reported some formal methodological training. Although members of the groups had varied backgrounds, the parapsychologists tended to be social scientists, while the skeptics tended to be natural scientists.

On the whole, the parapsychologists were more seasoned than the skeptics. They were slightly older, had longer active interests in this area, and had published a larger number of relevant articles. The difference in number of articles published was significant, $t(109) = 3.04$, $p < .01$, although not surprising because there are no journals dedicated to publish skeptical parapsychological research.[6]

Results

Separate 2 × 2 × 2 MANOVAs were performed on the evaluative and

[6] *The Skeptical Inquirer*, a magazine published by the Committee for the Scientific Investigation of Claims of the Paranormal (CSICOP), publishes skeptical pieces, but these are not usually reports of experimental research.

analytical judgments. Group (parapsychologists, skeptics), study quality (high, low), and agreement (agree, disagree) were between-subjects variables. Despite unequal cell sizes, violations of the repeated measures MANOVA homogeneity of variance assumption were not detected; consequently the data were not transformed.

Evaluative judgments. As predicted, a significant agreement effect was found, $F(1,79) = 9.20, p < .005$. Studies that agreed with scientists' prior beliefs were given more favorable ratings than studies that disagreed. This pattern was observed in both groups (see Fig. 3). However, the group × agreement interaction was marginally significant, $F(1,79) = 2.84, p < .10$, suggesting that the agreement effect may be larger for the skeptics than the parapsychologists (skeptics: $M_{\text{agree}} = 5.1$, $M_{\text{disagree}} = 3.3$; parapsychologists: $M_{\text{agree}} = 4.5$, $M_{\text{disagree}} = 3.9$). Because few parapsychologists or skeptics reported weak prior beliefs about ESP, it is difficult to determine what effect, if any, prior belief strength had on their quality judgments.

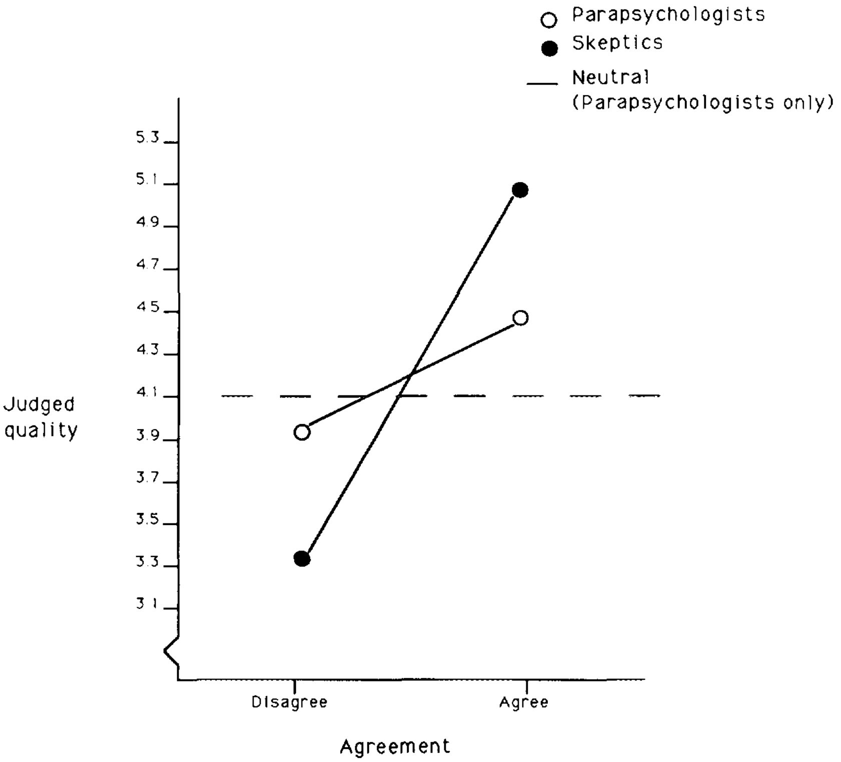

FIG. 3. Study 2 (Survey). Mean judgments of quality for the evaluative questions as a function of agreement and prior belief strength.

The evaluative ratings given by the parapsychological neutral group did not differ significantly from those given by parapsychologists in either the agree or disagree conditions. Finally, the ratings given to high quality studies were not significantly higher than those given to the low quality studies ($M_{\text{high qual}} = 4.7$, $M_{\text{low qual}} = 4.1$, $F(1,79) = 1.36$, n.s.).

Analytical judgments. An overall agreement effect was not found on the analytical judgments, $F(1,78) = 2.03$, n.s. However, separate planned contrasts on the two groups revealed a significant agreement effect among the skeptics, $t(34) = 2.13$, $p < .04$ ($M_{\text{agree}} = 4.7$, $M_{\text{disagree}} = 3.8$). A main effect for study quality was found on the analytical judgments ($M_{\text{high qual}} = 4.9$, $M_{\text{low qual}} = 4.1$, $F(1,78) = 4.95$, $p < .03$).

Open-ended results. Ninety-six of the 114 respondents (84.2%) made an average of 9.2 codable written comments about the research reports. Written comments were retyped, separated from scientists' numerical ratings, and given to two judges for coding. The judges coded all comments as positive, negative or neutral.

The judges coded the comments similarly, Judge 1 coded 12.0% of the comments as positive, 73.2% as negative and 14.8% as neutral; judge 2 coded 11.7% of the comments as positive, 69.7% as negative and 18.6% as neutral. Binomial tests of proportions did not reveal differences between the raters. Consequently, the means of the judges' ratings were used in the analyses below.

A 2 × 2 × 2 ANOVA was performed with group, study quality, and agreement as between-subjects variables. The dependent measure was an arcsin transformation of the proportion of negative comments to the total number of positive and negative comments (see Kirk, 1982, p. 82).

Contrary to predictions, there was no agreement effect on the open-ended questions; subjects in the disagree conditions did not make proportionally more negative comments about the studies than did subjects in the agree conditions, $F(1,58) < 1$. However, a significant main effect for study quality, $F(1,58) = 10.11$, $p < .01$, and a marginally significant group × study quality interaction, $F(1,58) = 3.40$, $p < .07$, were detected. The main effect indicates that a smaller proportion of negative comments were made about high quality studies (72%) than low quality studies (86%) overall. The interaction indicates that while a large proportion of negative comments were made by the parapsychologists for both high and low quality studies (81 and 86%, respectively), the skeptics were less critical of the high quality study (52%) than the low quality study (88%).

Self-perceptions and normative views. The results of the self-perception and normative questions were similar to those in the laboratory experiment. Sixty-five percent of the scientists in the experimental groups felt that their assessment of the methodological quality of the ganzfeld study *was not* influenced by the direction of the results it yielded; 21% felt that their judgments were influenced and 14% weren't sure. A large majority (85%) felt that a scientist's assessment of the methodological quality of a study *should not* depend upon its outcome; 13% said that such a dependency should exist, and 2% were not sure.

Parapsychologists and skeptics responded similarly to these questions, and there were no significant question framing effects. As in the laboratory experiment, no significant differences in the size of the agreement effects were found as a function of the scientists' responses to either the self-perception or normative questions.

Discussion

Agreement. The results of the experimental survey of practicing scientists provide further evidence for an agreement effect in scientific judgments of evidence quality. Scientists tended to judge studies that supported their beliefs about ESP to be more relevant, methodologically sound and clearly presented than otherwise identical studies that opposed their beliefs. However, the scientists did not make more negative comments about disconfirming studies, nor did they consistently regard the specific aspects of disconfirming studies (captured in the analytical questions) to be inferior to those of confirming studies.

Previous studies of nonscientist ESP believers and disbelievers indicate that believers think less critically (Alcock & Otis, 1980), are more dogmatic (Zusne & Jones, 1982), and show poorer recall of disconfirming evidence (Russell & Jones, 1980) than disbelievers. In the present study, however, parapsychologists did not show a greater propensity than skeptics to give biased assessments of the quality of belief-confirming and belief-disconfirming evidence. If anything, the agreement effect was larger for skeptics in both the evaluative and analytical domains. Although the present study did not directly address this issue, it does raise some doubts about whether previously reported differences between ESP believers and disbelievers extend to those with scientific training.

As in the laboratory experiment, a diminution of the agreement effect on the analytical questions was observed. This result may be normatively justifiable; the results of an experiment should generally have less bearing on very specific, "objective" analytical judgments than on more global, evaluative judgments.

Self-perceptions and normative views. Scientists' responses to the self-perception and normative questions were similar to those observed in the laboratory experiment. Most scientists mistakenly believed that their quality judgments were not influenced by the study's results, and most believed that such influence was undesirable. This latter result suggests that many scientists share a belief in at least some tenets of the classical normative model of scientific conduct. However, a few scientists felt that the evidence evaluation process is a probabilistic one in which even the results inform the evaluator about the likely quality of the process that produced them. Regardless of whether one accepts the normative views of the scientific majority or minority, these self reports extend Nisbett and Wilson's (1977) observation that people have limited access to cognitive processes to the domain of scientific judgments of evidence quality.

GENERAL DISCUSSION

The results of the laboratory experiment and the quasi-experimental survey of practicing scientists support the presence of an agreement effect in scientific judgment. Research reports that confirmed scientists' prior beliefs were judged to be of higher quality than those that did not. There is also some evidence from the laboratory experiment that the agreement effect is mediated by the strength with which prior beliefs are held such that stronger beliefs induce larger agreement effects. This effect appears to be more pronounced in scientists' general, evaluative judgments.

The agreement effect may be offered as a partial explanation for belief perseverance and polarization phenomena. Lord et al.'s (1979) capital punishment study showed that subjects changed their beliefs little following exposure to disconfirming evidence, but expressed more confidence in their beliefs following exposure to confirming evidence. This resulted in a polarization of beliefs in which subjects exposed to a mixture of confirming and disconfirming evidence became more certain of their initial views. If the quality judgments made by subjects in the Lord et al. study were influenced by an agreement effect, then belief perseverance and, indeed, belief polarization were likely consequences. If subjects who are presented with confirming and disconfirming evidence first determine that the former is of higher quality than the latter, then their revised beliefs will (and perhaps should) reflect their differential probative value by becoming more extreme in the direction of their prior beliefs.

The Normative Issue

On the one hand, these results are disturbing because they intimate that quality judgments are variable and person-specific. The "preestablished impersonal criteria" (Merton, 1942/1973, p. 270) required by the classical model for evaluating scientific research do not seem to be employed in practice. In the present study, even scientists who had a good deal of methodological training and research experience differentially perceived the quality of scientific research as a function of how well the data supported their beliefs and how strongly the beliefs were held.

But as discussed earlier, an influence of prior beliefs on judgments of quality may be justifiable within a Bayesian framework. This is because the results of a study provide some information about its likely quality.

But does this mean that the scientists in the present study did the right thing (albeit unwittingly) when they allowed their prior beliefs to influence their judgments of evidence quality? Not necessarily. There are several important reasons why the Bayesian approach may not be an appropriate model for judgments of this sort. First, it is difficult to determine the appropriate amounts of influence outcome information should have on quality judgments. Simulations in Koehler (1989) indicate that the optimal agreement effect varies as a function of several factors, including the probability that good and bad quality studies will yield true results, the probability that true results will agree with one's prior beliefs, and the prior probability that the study in question meets a good quality standard.

A second difficulty with the Bayesian approach stems from the observation that scientists, like laypeople, tend to cling tenaciously to prior beliefs (Mahoney & DeMonbreun, 1978; Mitroff, 1983; Oskamp, 1965). If their strong, prior beliefs are erroneous, Bayesian scientists may regard evidence that opposes their beliefs to be weak and give little serious thought to revising those beliefs.

Finally, it is interesting to note that agreement effects were found even among those who denied that these beliefs influenced their quality judgments. If scientists were told that their judgments *ought* to be influenced by their prior beliefs, the danger of overcorrection could be heightened. This danger is particularly acute among scientists who hold prior beliefs that are more extreme than the existing evidence warrants (cf. Anderson & Kellam, 1992). For these reasons, a Bayesian approach to scientific judgments of evidence quality may not be appropriate.

Directions For Future Research

If one assumes that agreement effects of the sort found here are excessive, the results of at least two studies provide some reason for optimism. Lord, Lepper, & Preston (1984) showed that when subjects are instructed to consider the possibility that evidence for and against the efficacy of capital punishment supported the opposite conclusion, subjects gave significantly less attitude-congruent evaluations of the evidence. Koriat et al. (1980) showed that the overconfidence phenomenon was reduced in subjects who were asked to provide reasons against their favored answers. Arkes, Faust, Guilmette, and Hart (1988) employed the same procedure to counteract the hindsight bias. Similarly, it may be possible to reduce agreement effects among scientists by asking them to simulate different results for the studies they evaluate, or to identify reasons why the results of the studies might have turned out differently.

Many other issues related to scientific judgments of evidence quality deserve further investigation. Under what conditions are agreement effects most likely to occur? Are there circumstances in which the effect will not occur, or will occur, but in the opposite direction? Can scientists be taught to reduce confidence in their beliefs and to make quality judgments that reflect normatively appropriate amounts of prior belief influence?

Not all of the interesting and relevant issues lend themselves to empirical study. At the meta-level, for example, it may be suggested that those who investigate or read about agreement effects make judgments about their robustness and normative status, in part, as a function of a priori beliefs about such effects. Thus, those who are inclined to believe that agreement effects influence scientific judgments of evidence quality may be relatively more likely to regard the present studies as relevant, methodologically appropriate, etc., than those who believe that the classical model of scientific conduct is descriptively accurate. Depending on where one comes down on the normative issue, and how strong these effects are, such behavior may or may not require explicit correction.

Finally, the investigation of evidence quality judgments need not be limited to scientists. For example, the practical consequences associated with evidence quality judgments in politics, medicine, and law are also important. Moreover, the normative status of agreement effects becomes increasingly dubious in these and other domains where goals other than truth-seeking are involved. In the courtroom, for instance, there is at least as much concern with the fairness of the trial process as there is with the truthfulness of the verdicts rendered (Koehler & Shaviro, 1990). Certainly, the spirit of justice would be violated if jurors were encouraged to assess the quality of an attorney's arguments, in part, by how closely the attorney's conclusions coincide with their own prior beliefs about the defendant's guilt or innocence. Further investigation of the normative and descriptive aspects of agreement effects in these and other domains can provide useful extensions of the present findings.

APPENDIX A

Proof: $P(H) = P(A|T)$

The following proof shows that in the unbiased case, $P(H) = P(A|T)$. Assume an unbiased study such that

$$P(T|H) = P(T|-H) = P(T). \tag{1}$$

$$\text{By Bayes's Theorem: } P(T|H) = \frac{P(H|T)\,P(T)}{P(H)}. \tag{2}$$

$$\text{Thus: } \frac{P(H|T)\,P(T)}{P(H)} = P(T) \tag{3}$$

$$P(H|T)\,P(T) = P(H)\,P(T) \tag{4}$$
$$P(H|T) = P(H). \tag{5}$$

Four states of Agreement (A) may be defined such that

a. H&T implies A.

b. H& − T implies − A.

c. − H&T implies − A.

d. − H& − T implies A.

Thus:

$$P(H\&T) = P(T\&A) = P(H\&A) \tag{6}$$
$$P(H\&-T) = P(-T\&-A) = P(H\&-A) \tag{7}$$
$$P(-H\&T) = P(T\&-A) = P(-H\&-A) \tag{8}$$
$$P(-H\&-T) = P(-T\&A) = P(-H\&A). \tag{9}$$

Hence:

$$P(A|T) = \frac{P(A\&T)}{P(T)} = \frac{P(H\&T)}{P(T)} = P(H|T). \tag{10}$$

Because Eq. (5) shows $P(H|T) = P(H)$, $P(A|T) = P(H)$.

APPENDIX B

Bayesian Analyses of Agreement Effects

(a) *Agreement main effect.* A Bayesian analysis may be employed to determine when, if ever, $P(G|A) > P(G|-A)$.

Recall the probabilities defined in the text:

$P_1 = P(A|T) = P(H); .5 < P_1 < 1$

$P_2 = 0 < P(T|G); P_2 < 1$

$P_3 = 0 < P(T|-G); P_3 < 1$

$P_4 = 0 < P(G); P_4 < 1$

Four technical conditions are required:

1. Because every result is either true or not true, and every study is either good or not good:

$$P(-T|G) = 1 - P(T|G) = 1 - P_2 \tag{1A}$$
$$P(-T|\text{-}G) = 1 - P(T|-G) = 1 - P_3 \tag{1B}$$
$$P(-G) = 1 - P(G) = 1 - P_4. \tag{1C}$$

2. In the unbiased case analyzed here,

$$P(A|-T) = 1 - P(A|T) = 1 - P_1. \tag{2}$$

3. The probability that true results agree with one's prior beliefs does not depend on the quality of the study that produced those true results:

$$P(A|T) = P(A|T\&G) = P(A|T\&-G). \tag{3A}$$

Similarly,

$$P(A \mid -T) = P(A \mid -T\&G) = P(A \mid -T\&-G) \tag{3B}$$

4. Finally, because:

$$P(X \mid Y) > P(X \mid -Y) \text{ iff } P(Y \mid X) > P(Y \mid -X) \text{ for any X and Y} \tag{4A}$$
$$P(G \mid A) > P(G \mid -A) \text{ iff } P(A \mid G) > P(A \mid -G). \tag{4B}$$

Because of condition 4, our original inquiry may be restated as: When, if ever, is $P(A \mid G) > P(A \mid -G)$?

$$P(A \mid G) = \frac{P(A\&G)}{P(G)} \tag{5}$$
$$= \frac{P(A\&G\&T) + P(A\&G\&-T)}{P(G)} \tag{6}$$
$$= \frac{P(A \mid G\&T)\ P(G\&T)}{P(G)} + \frac{P(A \mid G\&-T)\ P(G\&-T)}{P(G)} \tag{7}$$

By Eqs. (3A) and (3B),

$$P(A \mid G) = P(A \mid T)\ P(T \mid G) + P(A \mid -T)\ P(-T \mid G). \tag{8}$$

By Eqs. (1A) and (2),

$$P(A \mid G) = P_1P_2 + (1 - P_1)(1 - P_2) \tag{9}$$
$$= 2P_1P_2 - P_1 - P_2 + 1. \tag{10}$$

Similarly,

$$P(A \mid -G) = P(A \mid T)\ P(T \mid -G) + P(A \mid -T)\ P(-T \mid -G) \tag{11}$$
$$= P_1P_3 + (1 - P_1)(1 - P_3) \tag{12}$$
$$= 2P_1P_3 - P_1 - P_3 + 1. \tag{13}$$

So, $P(A \mid G) > P(A \mid -G)$ when

$$2P_1P_2 - P_1 - P_2 + 1 > 2P_1P_3 - P_1 - P_3 + 1 \tag{14}$$
$$P_2\,(2P_1 - 1) > P_3\,(2P_1 - 1). \tag{15}$$

The expression in Eq. (15) is true if $P_1 > .5$ and $P_2 > P_3$.

Because these two conditions ($P_1 > .5$ and $P_2 > P_3$) were previously assumed, $P(A \mid G)$ is *always* greater than $P(A \mid -G)$, hence $P(G \mid A)$ is *always* greater than $P(G \mid -A)$.

(b) *Prior belief strength* × *agreement interaction.* A related normative question is whether prior belief strength, P_1, should influence the magnitude of this agreement effect. Consider:

$$P(G|A) = \frac{P(A|G)}{P(A)} P(G). \tag{16}$$

A measure of the impact of learning that the results of a study agree with prior beliefs is given by the ratio of the posterior to the prior:

$$\frac{P(G|A)}{P(G)} = \frac{P(A|G)}{P(A)} . \tag{17}$$

Note:

$$\begin{aligned} &P(A) = \\ &P(A|G)\,P(G) + P(A|-G)\,P(-G) = && (18)\\ &\qquad [P_1P_2 + (1 - P_1)(1 - P_2)]\,P_4 + [P_1P_3 \\ &\qquad\quad + (1 - P_1)(1 - P_3)](1 - P_4) = && (19)\\ &P_1(2P_2P_4 + 2P_3 - 2P_3P_4 - 1) + (-P_2P_4 - P_3 + P_3P_4 + 1). && (20) \end{aligned}$$

By Eqs. (10) and (20),

$$\frac{P(A|G)}{P(A)} = \frac{P_1(2P_2 - 1) + (1 - P_2)}{P_1(2P_2P_4 + 2P_3 - 2P_3P_4 - 1) + (-P_2P_4 - P_3 + P_3P_4 + 1)} \tag{21}$$

Let $K1 = 2P_2 - 1$

$$\begin{aligned} K2 &= 1 - P_2 \\ K3 &= 2P_2P_4 + 2P_3 - 2P_3P_4 - 1 \\ K4 &= -P_2P_4 - P_3 + P_3P_4 + 1. \end{aligned}$$

This gives:

$$\frac{P(G|A)}{P(G)} = \frac{P(A|G)}{P(A)} = \frac{P_1K1 + K2}{P_1K3 + K4} . \tag{22}$$

The differential of Eq. (22) with respect to P_1 is:

$$\frac{K1(P_1K3 + K4) - K3(P_1K1 + K2)}{(P_1K3 + K4)^2} . \tag{23}$$

The expression in Eq. (23) is > 0 iff:

$$K1(P_1K3 + K4) - K3(P_1K1 + K2) > 0 \tag{24}$$

$$K1K4 > K2K3. \tag{25}$$

By substitution Eq. (25) becomes

$$(2P_2 - 1)(-P_2P_4 - P_3 + P_3P_4 + 1) > (1 - P_2)(2P_2P_4 + 2P_3 - 2P_3P_4 - 1) \tag{26}$$

$$P_2(1 - P_4) > P_3(1 - P_4) \tag{27}$$
$$P_2 > P_3. \tag{28}$$

In words, the ratio given in Eq. (22) increases as P_1 increases if and only if $P_2 > P_3$. Now recall that it was previously assumed that $P_2 > P_3$ (because good studies presumably are more likely to yield true results than are bad studies). Therefore, as prior beliefs become increasingly strong, the ratio of $P(G|A)/P(G)$—hence the agreement effect—should get larger.

REFERENCES

Alcock, J. E. (1987). Parapsychology: Science of the anomalous or search for the soul. *Behavioral and Brain Sciences*, **10**, 553–565.

Alcock, J. E., & Otis, L. P. (1980). Critical thinking and belief in the paranormal. *Psychological Reports*, **46**, 479–482.

Anderson, C. A., & Kellam, K. L. (1992). Belief perseverance, biased assimilation, and covariation detection: The effects of hypothetical social theories and new data. *Personality and Social Psychology Bulletin*, **18**, 555–565.

Arkes, H. R., Faust, D., Guilmette, T. J., & Hart, K. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology*, **73**, 305–307.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, **54**, 569–579.

Batson, C. D. (1975). Rational processing or rationalization?: The effect of disconfirming information on stated religious belief. *Journal of Personality and Social Psychology*, **32**, 176–184.

Bauslaugh, G. (1981). Science, intuition and ESP. In K. Frazier (Ed.), *Paranormal Borderlands of Science* (pp. 24–31). New York: Prometheus Books.

Child, I. L. (1985). Psychology and anomalous observations: The question of ESP in dreams. *American Psychologist*, **40**, 1219–1230.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal Representations of Human Judgment*. New York: Wiley.

Einhorn, H. J., & Koelb, C. (1982). A Psychometric study of literary-critical judgment. *Modern Language Studies*, **12**, 59–82.

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When Prophecy Fails*. Minneapolis: University of Minnesota Press.

Geller, E. S., & Pitz, G. F. (1968). Confidence and decision speed in the revision of opinion. *Organizational Behavior and Human Performance*, **3**, 190–201.

Hansel, C. E. M. (1980). *ESP and Parapsychology: A Critical Re-evaluation*. New York: Prometheus Books.

Honorton, C. (1985). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology*, **49**, 51–91.

Kelly, E. F. (1979). Reply to Persi Diaconis. *Zetetic Scholar*, **5**, 20–28.

Kirk, R. E. (1982). *Experimental Design: Procedures for the Behavioral Sciences*. Monterey, CA: Brooks/Cole.

Kleck, R. E., & Wheaton, J. (1967). Dogmatism and responses to opinion-consistent and opinion-inconsistent information. *Journal of Personality and Social Psychology*, **5**, 249–252.

Koehler, J. J. (1989). *Judgments of evidence quality among scientists as a function of prior beliefs and commitments*. Unpublished Doctoral Dissertation.

Koehler, J. J., & Shaviro, D. (1990). Veridical verdicts: Increasing verdict accuracy through the use of overtly probabilistic evidence and methods. *Cornell Law Review*, **75**, 247–279.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, 107–118.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The Development of Scientific Thinking Skills*. San Diego: Academic Press.

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

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, **37**, 2098–2109.

Mahoney, M. J. (1976). *Scientist as Subject: The Psychological Imperative*. Cambridge, MA: Ballinger.

Mahoney, M. J. (1977). Publication Prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, **1**, 161–175.

Mahoney, M. (1979). Psychology of the scientist: An evaluative review. *Social Studies of Science*, **9**, 349–375.

Mahoney, M. J., & DeMonbreun, B. G. (1978). Psychology of the scientist: An analysis of problem-solving bias. *Cognitive Therapy and Research*, **1**, 229–238.

Marks, D. F. (1986). Investigating the paranormal. *Nature*, **320**, 119–124.

Merton, R. K. (1942/1973). The normative structure of science. In R. K. Merton & N. W. Storer (Eds.), *The Sociology of Science: Theoretical and Empirical Investigations* (pp. 267–278). Chicago: University of Chicago Press.

Mitroff, I. I. (1983). *The Subjective Side of Science*. Seaside, CA: Intersystems.

Mulkay, M., & Gilbert, N. (1982). Accounting for error: How scientists construct their social world when they account for correct and incorrect belief. *Sociology*, **16**, 165–183.

Nisbett, R. E., & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review*, **84**, 231–259.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, **29**, 261–265.

Pitz, G. F. (1969). An inertia (resistance to change) in the revision of opinion. *Canadian Journal of Psychology*, **23**, 24–33.

Pitz, G. F., Downing, L., & Reinhold, H. (1967). Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology*, **21**, 381–393.

Rao, K. R. (1979). On the scientific credibility of ESP. *Perceptual and Motor Skills*, **49**, 415–429.

Rao, K. R., & Palmer, J. (1987). The anomaly called psi: Recent research and criticisms. *Behavioral and Brain Sciences*, **10**, 539–551.

Rockwell, T., Rockwell, R., & Rockwell, W. T. (1978). Irrational rationalists: A critique of *The Humanist's* crusade against parapsychology. *Journal of the American Society for Psychical Research*, **72**, 23–34.

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.

Ross, M., McFarland, C., & Fletcher, G. J. O. (1981). The effect of attitude on recall of personal history. *Journal of Personality and Social Psychology*, **40**, 627–634.

Russell, D., & Jones, W. H. (1980). When superstition fails: Reactions to disconfirmations of paranormal beliefs. *Personality and Social Psychology Bulletin,* **6,** 83–88.

Scheffler, I. (1967). *Science and Subjectivity.* Indianapolis: Bobbs-Merrill.

Snyder, M., & Uranowitz, S. W. (1978). Reconstructing the past: Some cognitive consequences of person perception. *Journal of Personality and Social Psychology,* **36,** 941–950.

Zusne, L., & Jones, W. H. (1982). *Anomalistic Psychology: A Study of Extraordinary Phenomena of Behavior and Experience.* Hillsdale, NJ: Erlbaum.

Received: March 6, 1991