Journal of Personality and Social Psychology
1980, Vol. 39, No. 6, 1037-1049

Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information

Craig A. Anderson, Mark R. Lepper, and Lee Ross
Stanford University

The perseverance of social theories was examined in two experiments within a debriefing paradigm. Subjects were initially given two case studies suggestive of either a positive or a negative relationship between risk taking and success as a firefighter. Some subjects were asked to provide a written explanation of the relationship; others were not. In addition, experimental subjects were thoroughly debriefed concerning the fictitious nature of the initial case studies. Subsequent assessments of subjects' personal beliefs about the relationship indicated that even when initially based on weak data, social theories can survive the total discrediting of that initial evidential base. Both correlational and experimental results suggested that such unwarranted theory perseverance may be mediated, in part, by the cognitive process of formulating causal scenarios or explanations. Normative issues and the cognitive processes underlying perseverance were examined in detail, and possible techniques for overcoming unwarranted theory perseverance were discussed.

All of us have tried to change a friend's view about some social, political or scientific issue—from the efficacy of capital punishment as a deterrent to crime to the validity of the IQ test—only to experience frustrating failure. We offer seemingly compelling evidence or thoroughly rebut opposing arguments but produce little if any change in our friend's beliefs. Indeed, we suspect that we ourselves frequently may be guilty of similar intransigence when our views come under attack. From such everyday observations, two obvious questions arise that form the focus of the present article: Are we, in fact, prone to persist in our beliefs and theories about the world to a degree that is normatively indefensible, and if so, why?

This research was supported in part by Research Grants MH-26736 from the National Institute of Mental Health and BNS-78-01211 from the National Science Foundation to the second and third authors, and by a National Science Foundation Fellowship to the first author. The authors wish to express their thanks to Mark Zanna for his insightful comments on an earlier draft of this article and to Dan Boeriu for his assistance in conducting Experiment 2.

Requests for reprints should be sent to Craig Anderson, who is now at the Department of Psychology, Rice University, Houston, Texas 77001.

The first, "normativeness," question inevitably proves to be a complex and subtle one, and we shall postpone much of our discussion of this question until we have presented details about the present procedures and results. For now let us simply note that any consideration of the proposition that our beliefs are less responsive to empirical or logical attacks than they "ought to be" requires that one be able to specify *how much* change in our beliefs would be warranted by any particular challenge to those beliefs. Everyday observations, however, rarely permit such specification. Generally, all one can say with certainty is that some change is warranted whenever the empirical or logical attacks seem to have merit.

Nevertheless one case does arise in which normative standards seem less flexible. We refer to circumstances in which *all* of the evidence that initially gave rise to a particular theory is thoroughly and completely discredited. When all of the evidence on which I have based a belief, or based a change in belief, is shown to my satisfaction to be totally without evidential value—if, for example, all of it is shown to be fictitious—then most

arbiters of normativeness would agree that my belief ought to revert to its original state.

To date, there is no experimental research concerning the perseverance of theories in the face of total evidential discrediting. Some pertinent evidence exists, however, for one class of less abstract beliefs. It appears that specific personal impressions in a given domain concerning one's own abilities or those of a peer may survive even the complete invalidation of the evidence on which the impressions initially were based. Ross, Lepper, and Hubbard (1975), for example, provided subjects with false feedback indicating their apparent success or failure at discriminating authentic suicide notes from inauthentic ones, a task purported to assess their social sensitivity and empathetic ability. For half of the subjects the probative value of this feedback was subsequently completely negated by a thorough "debriefing" procedure. Although subjects understood and accepted this "debriefing," their predictions of future task success and ratings of their own abilities continued to be heavily influenced by the discredited prior success or failure feedback. Similar perseverance effects were also apparent in the social impressions and predictions made by observers who had witnessed the subjects' original outcomes and subsequent debriefings. Other experiments, carried out in more applied contexts and with more naturalistic discrediting procedures, have shown that erroneous first impressions about other abilities such as "personal persuasiveness" (Jennings, Lepper, & Ross, Note 1) and "logical reasoning" (Lepper, Ross, & Lau, Note 2) can likewise survive the removal of their initial evidential basis.

A first major objective of the present experiments on theory perseverance, therefore, was to extend these previous findings regarding the unwarranted perseverance of initial beliefs. On the one hand, we sought to show that the perseverance phenomenon would apply beyond the limited domain of highly specific personal and social impressions. On the other, we sought to demonstrate that perseverance effects may occur even when subjects' theories are initially based on minimal, and indeed logically inadequate, evidence—when their beliefs are of exactly the tentative, hastily-formed, and ill-founded variety most likely to face subsequent logical or evidential discrediting in everyday experience. By focusing on tentative social theories, the studies to be reported attempted to examine the perseverance of beliefs that may occur even in the absence of strong emotional or behavioral commitments, or logically compelling prior evidence.

At the same time, we also sought to address a second major question, concerning the mechanisms that may underlie theory perseverance, through a direct examination of one cognitive process hypothesized to foster belief perseverance. This process involves the formulation of relevant causal scripts or explanations, and derives from people's propensity to seek or construct explanations to account for salient events or relationships among events that one has noted (cf. Kelley, 1967, 1973). Such causal accounts provide the perceiver with an important and efficient means of organizing and understanding the social world. Yet, because such accounts may become independent of the data that originally gave rise to them, they may contribute to the unwarranted persistence of initial impressions and theories as well. Once a causal account has been generated, it will continue to imply the likelihood of the "explained" state of affairs even after the original basis for believing in that state of affairs has been eliminated (Ross & Anderson, in press; Ross & Lepper, in press). For example, the scientist who has explained why early humans might have inhabitated a particular region will continue to believe such inhabitation is likely even after the fossil evidence that originally prompted the explanation has been thoroughly discredited. Consistent with this analysis, Ross, Lepper, Strack, & Steinmetz (1977) have shown that providing an explanation for some possible outcome in an individual's life increases the subjective likelihood of that outcome. Fischhoff's provocative investigations of hindsight phenomena (e.g., Fischhoff, 1975) similarly suggest the power of causal explanations to influence expressions of likelihood or even inevitability. The role of such explanation processes in the perseverance of social

theories, however, has not received previous study.

The two studies that follow, therefore, examine the operation of this explanation mechanism in the perseverance of theories based on inadequate initial evidence. In both experiments subjects were asked to explain an empirical relationship—that existing between success in the occupation of firefighter and preference for risk as measured by a paper and pencil test—just before learning that the "evidence" that initially had led them to believe in that relationship was entirely fictitious. In Experiment 1, the explanation task was introduced for all subjects; it was designed primarily to enhance the likelihood of belief perseverance, but it also permitted us to correlate the quality of an explanation with its impact on the subject's postdiscrediting beliefs. In Experiment 2 inclusion versus exclusion of the explanation task was deliberately manipulated, allowing us to contrast directly the magnitude of theory perseverance in the presence and absence of the interpolated explanation task. Our interest in these studies thus focused on two main questions: First, would subjects continue to hold a given theory about an empirical relationship between variables after the meager evidential basis for that theory had been invalidated? Second, would the process of providing an explanation for a given relationship increase subjects' tendency to persevere in their theories following this discrediting procedure?

Experiment 1

In our initial experiment, subjects were first led to believe that either a positive or negative relationship existed between a trainee's preference for risky versus conservative choices and his subsequent success as a firefighter, and were asked to provide a written explanation of this relationship. Subjects in "debriefing" conditions were then explicitly informed that the initial information they had been asked to consider was bogus and of absolutely no probative value. This debriefing was omitted for a group of "no debriefing" subjects. All subjects then completed several dependent measures assessing their beliefs concerning the true relationship between these two variables and the predictive power of the relationship. A baseline control group received no information about the relationship between the two variables but completed the various dependent measures. The final design was thus a 2×2 factorial (Positive vs. Negative Relationship $\times$ Debriefing vs. No Debriefing), with an added baseline control group.

Method

Subjects

Seventy Stanford undergraduates, participating in groups of two to eight, took part in the experiment, for which they received course credit. Since subjects in the experimental groups were to receive instructions that differed slightly from those in the baseline control group, these latter subjects participated in separate sessions. Within the experimental conditions, subject assignment was randomized in blocks of 12, and the experimenter remained blind to subjects' conditions.

Procedure

Experimental subjects were told that the experiment was concerned with how well people are able to discover and explain relationships between personal characteristics of people and their behavioral outcomes. They were informed that they would be asked to examine some data and to see if they could discover and subsequently explain underlying relationships between general traits and specific behaviors. Control subjects were told simply that they would be asked to make predictions concerning the abilities of persons as a function of their performance on a test of risk preference. After answering any questions about these general instructions, the experimenter gave subjects booklets containing the experimental materials.

Manipulation of initial theories. Booklets for subjects in the experimental conditions first stated that the subject's task was to examine the relationship between eventual success or failure as a firefighter and prior performance on a "Risky–Conservative Choice Test" (RCC test). Next, nondiagnostic background information (age, marital status, hobbies, etc.) for one successful and one unsuccessful firefighter was presented, along with each firefighter's responses to the five "most representative" risky–conservative choice problems. Subjects were instructed to examine this information about riskiness and ability, to attempt to discover the underlying relationship, and to provide a written explanation of any relationship they uncovered.

In all conditions the RCC test items were similar to items used previously in research on group-induced shifts to risk taking. Each item presented a dilemma and two possible behavioral alternatives, one risky and one conservative. Each response was a short paragraph, purportedly written by the firefighter, that gave his choice of action and explanation of the choice. For half the subjects, the purported responses were arranged to demonstrate a positive relationship between risky choices and later success as a firefighter; for the remaining subjects, these responses were arranged to demonstrate a negative relationship between risky choices and success as a firefighter.

To ascertain that subjects had indeed "discovered" different relationships in these two conditions, a manipulation check was included immediately before the explanation task. The results from this measure—a 101-point scale anchored at "Highly Positive Relationship" (50), "No Relationship" (0), and "Highly Negative Relationship" (−50)—indicated that subjects in the positive relationship conditions did "discover" a positive relationship, $M = 33.87$, $t(54) = 9.99$, $p < .0001$, whereas subjects in negative relationship conditions "discovered" a negative relationship, $M = -20.57$, $t(54) = -6.01$, $p < .0001$. After completing these measures, all subjects were asked to provide a one-page, written explanation of the relationship they had uncovered in the two case studies.

Debriefing manipulations. Within the two relationship conditions, two thirds of the experimental subjects were assigned, at random, to the debriefing conditions.[1] These subjects received a detailed, written debriefing following the explanation task informing them that they had been randomly assigned the task of discovering and explaining either a positive or a negative relationship between risk preference and success as a firefighter. To insure that subjects did not still perceive the case data they had received to be representative of a true relationship, subjects were also explicitly informed that the experimenters had provided fictitious information consistent with a positive relationship or a negative relationship to subjects in different conditions, and that the experimenters did not know the nature or strength of the "true" relationship.

The last section of the debriefing materials explained to subjects that the prediction and estimation tasks to follow were for "control purposes," to see if their personal theories about the relationship in question had influenced their discovery of a relationship or the quality of their explanations. Subjects were urged to make all their judgments based on their personal beliefs and not the fictitious information initially presented. During final postexperimental debriefing sessions, all subjects indicated an awareness and understanding of these critical instructions.

Subjects in the no-debriefing conditions received no such information. After completing their explanations, these subjects proceeded directly to the dependent measures. Control subjects received booklets informing them that their task was to examine the information within, and they were also asked to make their predictions and estimates based on their personal beliefs. These subjects were given no information about the relative success of the two case study firefighters, nor were they asked to explain any relationship between risk preference and task success. In all other respects, all subjects received identical materials and completed identical measures.

Dependent measures. Several dependent measures were employed, each designed to assess subjects' beliefs concerning the true relationship between a preference for risky choices and ability as a firefighter. To minimize the relevance of social evaluation concerns, these measures were collected under conditions of anonymity.

The first of these measures asked subjects to judge directly the "criterion validity" of the risk preference scale as an index of firefighting ability. Subjects were asked to estimate the average percentage of risky choices for two groups of firefighters—those who had subsequently become highly successful and those who had been failures at the job. The perceived criterion validity of the RCC test was assessed by subtracting the expected percentage of risky responses among failure firefighters from the expected percentage of such choices among successful firefighters, yielding a difference score that could range from 100 (a maximally strong positive relationship) to −100 (a maximally strong negative relationship).

A second set of measures dealt with subjects' willingness to generalize, on the basis of their beliefs about the riskiness–success relationship, in making predictions about new cases and new items. For the former measure, subjects were presented with information on four new trainees, including both nondiagnostic background evidence (e.g., father's occupation, marital status, etc.) and the individual's response to one RCC Test item. Subjects then predicted each trainee's subsequent success, allowing a

[1] Two types of debriefing were used, both of which fully invalidated the "evidence" initially presented. In "ability debriefing" groups, subjects were informed that the relative abilities of the two sample firefighters had been manufactured but that the remainder of the information about them (i.e., their riskiness responses) had been authentic. In the "trait debriefing" groups, the riskiness information was discredited, but the ability information was allowed to stand. This variation was introduced to see whether subjects would be more responsive to discounting of one type of information than another. Preliminary analyses yielded no consistent effects of the two variants of the debriefing. Hence, the data were collapsed across this dimension in all tables, and in all further analyses equal contrast weights were assigned to the two debriefing conditions within the positive or negative relationship manipulations (see Winer, 1971).

Table 1
Mean Postexperimental Beliefs Concerning Relationship Between Risk Preference and Firefighter Success: Experiment 1

Dependent measure	Positive relationship		Negative relationship		Control
	No debrief	Debrief	No debrief	Debrief	
Perceived criterion validity[a]	51.8	36.0	−14.2	−1.3	25.5
Generalization to new cases[b]	1.4	1.1	−2.0	−1.3	.2
Generalization to new items[c]	41.8	31.7	−10.9	−5.5	19.4
n	10	20	10	20	10

Note. Positive scores indicate belief in a positive relationship; negative scores indicate belief in a negative relationship.
[a] Predicted percentage of risky responses on the risky–conservative choice test for superior minus unsuccessful firefighters. Range of possible scores is 100 to −100.
[b] (Number of success–risky + number of failure–conservative) − (number of success–conservative + number of failure–risky) predictions to four new cases. Range of possible scores is 4 to −4.
[c] Predicted percentage of risky responses on five new items for superior minus unsuccessful firefighters. Range of possible scores is 100 to −100.

test of the extent to which their predictions conformed to those that would follow from beliefs in a positive or a negative relationship between RCC scores and occupational success. For the latter measure, subjects were presented with five novel hypothetical-choice items in the same general format as those in the RCC test, and were asked to indicate the percentage of risky choices that superior and inferior firefighters would make on these items. A difference score served as an index of the subjects' willingness to generalize to new test items on the basis of their correlational theory. At the completion of the experiment, subjects were probed for suspicion and given a thorough explanation of the procedures and purposes of the study and of the processes that may mediate the unwarranted perseverance of initial beliefs.

Results and Discussion

The results for each of these three measures are presented by condition in Table 1. As one might expect, the three measures proved to be highly intercorrelated (average $r = .73$). Thus, the data on each were transformed into Z scores and summed to provide a composite measure of subjects' beliefs concerning the true relationship between risk preferences and subsequent success as a firefighter. The data from these composite scores are presented in Figure 1, and it is on these data that our primary analyses were performed.[2]

We should first note, perhaps, that although the "data" to which subjects had been initially exposed were objectively quite weak (consisting of only two cases) and in a domain of little personal relevance, this initial ostensible evidence clearly exerted a strong effect on subjects' theories about the true relationship between the two variables. Thus, in the no-debriefing conditions, subjects exposed to a positive relationship saw risky responses as highly diagnostic of later success, whereas subjects exposed to an apparent negative relationship believed the opposite to be true, $F(1, 63) = 20.21$, $p < .0001$.

Given these clear effects of initial information on subjects' beliefs prior to debriefing, it is possible to examine the perseverance of these beliefs in the debriefing conditions, after subjects learned that the two case studies were fictitious. As is evident in Figure 1, the total discrediting of the evidence on which subjects' initial theories had been based had only a minimal impact on subjects' beliefs concerning the relationship between risk preference and firefighting ability. Within the debriefing conditions, subjects initially exposed to data indicative of a positive relationship continued to believe that a positive relationship existed, whereas subjects in the negative relationship condition continued to

[2] Separate analyses parallel to those to be reported below were also performed for each of the three component measures. In all cases, effects that proved statistically significant in the combined analyses were individually significant ($p < .05$) for each of the three component measures as well.

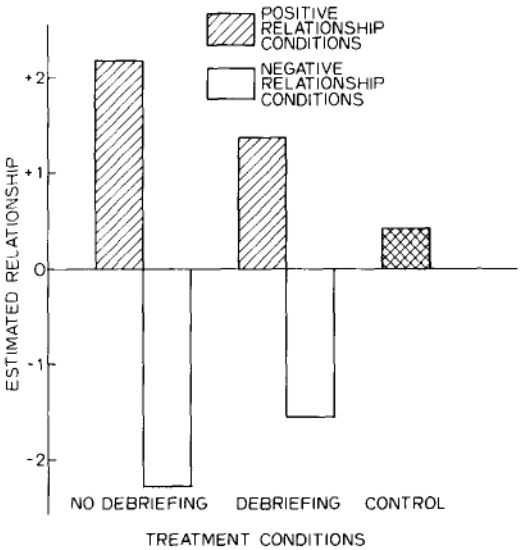

Figure 1. Mean composite indices (Z scores summed across the three measures) of subjects' personal estimates of the true relationship between risk preference and firefighter success, Experiment 1. (Positive scores indicate belief in a positive relationship; negative scores indicate belief in a negative relationship.)

believe in a negative relationship, $F(1, 63) = 17.43$, $p < .0001$. In fact, the slight decrease in the strength of subjects' beliefs following debriefing, as assessed by the Relationship × Debriefing interaction, was not statistically significant, $F(1, 63) = 1.59$.

These effects, moreover, were roughly symmetrical with respect to the responses of control subjects not initially exposed to the two case studies, who tended to believe that there would be a slight positive relationship between a preference for risky choices and later success as a firefighter. Consideration of the debriefed groups and the control group in a one-way analysis of variance yielded strong evidence of a monotonic relationship between initial information and subsequent beliefs, monotonic contrast $F(1, 45) = 16.13$, $p < .0005$, with a nonsignificant residual, $F(3, 45) = 1.24$.

In sum, the results strongly support the hypothesis that even after the initial evidential basis for their beliefs has been totally refuted, people fail to make appropriate revisions in those beliefs. That subjects' theories survived virtually intact is particularly impressive when one contrasts the minimal nature of the evidential base from which subjects' initial beliefs were derived (i.e., two "data points"), with the decisiveness of the discrediting to which that evidence was subjected. In everyday experience our intuitive theories and beliefs are sometimes based on just such inconclusive data, but challenges to such beliefs and the formative data for those beliefs are rarely as decisive as the discrediting procedures employed in this study.

If our speculations about underlying mechanisms are correct, however, the design of Experiment 1 also included one feature that is only occasionally present in everyday experience, but that should have served to augment any perseverance effects—the requirement that subjects provide an explicit explanation for the relationship they had observed. While Experiment 1 contains no direct evidence on this issue, some indirect support for this hypothesized perseverance-enhancing effect of explanation is provided by an internal analysis of the relationship between the nature of subjects' explanations and the persistence of their initial theories following debriefing.

Although an unintended source of variance in our procedure, the explanations subjects provided proved to be of two distinct types. Some subjects responded to the request as intended, offering some general causal account that explained the specific case studies they had read. These accounts typically focused on the risks inherent in fighting fires and the importance, depending on condition, of either a willingness to take necessary risks or the need to avoid foolhardy and impulsive action as a determinant of successful performance. Other subjects apparently construed the task differently. These subjects simply "explained" how the specific information contained in the two particular case studies they had read illustrated either a positive or a negative relationship—they basically restated the fact that the more successful candidate had selected considerably more risky, or more conservative, alternatives than his less successful counterpart.

Theoretically, the effects of these two sorts of explanations on belief perseverance should

differ, since only subjects in the former case have generated principles that should continue, even after debriefing, to imply the existence of the relationship they had initially observed. Only in this first case should explanations enhance theory perseverance. Consistent with this analysis, the data revealed a highly significant point-biserial correlation between the presence or absence of general explanatory principles in subjects' explanations and the degree of postdebriefing belief perseverance, $r = .54$, $p < .0005$.

Experiment 2

Experiment 1 demonstrated that a theory concerning the relationship between two variables—generated through exposure to a minimal data set—can survive even a complete refutation of the formative evidence on which the theory was initially based. Questions remain, however, concerning the role of the explanation processes in mediating these perseverance effects—an issue to which Experiment 2 was directed.

Specifically, Experiment 2 addressed two primary questions: Whether explicit explanation is a necessary precondition for the postdiscrediting perseverance of a theory or, if not, whether it nevertheless increases the magnitude of such perseverance. To address these issues, Experiment 2 compared six conditions. As in Experiment 1, all subjects were first presented with information on illustrative cases involving one highly successful and one clearly unsuccessful firefighter and were asked to discover the relationship between risk preference and firefighting ability—either positive or negative—contained in these case data. Within these conditions, one third of the subjects were next asked to write an explanation of the discovered relationships and were then debriefed concerning the fictitious nature of the case study materials, as in Experiment 1. No mention of explanation was made in the remaining four conditions. In these no explanation conditions, half of the subjects were debriefed concerning the fictitious nature of the experimental materials; the remainder were not debriefed. The effects of exposure to initial evidence illustrating either a positive or a negative relationship were thus examined under three variations of debriefing and explanation: debriefing–explanation, debriefing–no explanation, and no debriefing–no explanation. Finally, as before, all subjects completed a set of dependent measures assessing their beliefs concerning the nature of the actual relationship between the two critical variables.

Method

Subjects in Experiment 2 were 62 Stanford University undergraduates, who received credit toward an introductory psychology class requirement or $2.00 for their participation in the study. Data from two subjects who failed to complete the test materials were not used. Procedures, booklets, and instructions were basically identical to those of Experiment 1, with only those changes necessary to accommodate the new design of this second study. Thus, all subjects were initially presented with the same case study materials used in Experiment 1, suggesting either a positive relationship between risk preference and subsequent success or a negative relationship between these variables. Orthogonally, subjects in two sets of debriefing conditions were fully debriefed concerning the fictitious nature of these putative initial data.[3] Prior to debriefing, however, half of these subjects were asked to provide an explicit explanation of the relationship they had uncovered in the case studies; half were not given this explanation task. For purposes of comparison, subjects in the no-debriefing conditions were exposed to the initial data but were neither debriefed concerning its fictitious character nor asked to explain the relationship they had uncovered. Following these procedures, as before, three measures of subjects' personal beliefs concerning the true relationship between these variables—the perceived criterion validity of the RCC Test and generalization to new case studies and new test items—were anonymously assessed. Finally, all subjects were probed for suspicion and given a complete explanation of the procedures and purposes of the study and the processes that may mediate the unwarranted perseverance of initial beliefs.

Results

The results from these three measures of subjects' subsequent beliefs concerning the

[3] To provide an even more complete discrediting of the data on which subjects' initial beliefs rested, subjects in Experiment 2 were informed that the two case studies were entirely fictitious, that is, that *both* the ability ratings and the risk preference information had been manufactured by the experimenters and had been randomly assigned to them.

Table 2
Mean Postexperimental Beliefs Concerning Relationship Between Risk Preference and Firefighter Success: Experiment 2

Dependent measure	Positive relationship			Negative relationship		
	No debrief–no explain	Debrief–explain	Debrief–no explain	No debrief–no explain	Debrief–explain	Debrief–no explain
Perceived criterion validity[a]	54.0	39.0	39.5	−57.4	−35.3	−14.0
Generalization to new cases[b]	2.2	1.0	0.2	−3.0	−2.8	−.6
Generalization to new items[c]	42.1	37.1	26.1	−38.5	−30.5	−9.9
n	10	10	10	10	10	10

Note. Positive scores indicate belief in a positive relationship; negative scores indicate belief in a negative relationship.
[a] Subjects' predicted percentage of risky responses on the risky–conservative choice test for superior minus unsuccessful firefighters. Range of possible scores is 100 to −100.
[b] (Number of success–risky + number of failure–conservative) − (number of success–conservative + number of failure–risky) predictions to four new cases. Range of possible scores is 4 to −4.
[c] Subjects' predicted percentage of risky responses on five new items for superior minus unsuccessful firefighters. Range of possible scores is 100 to −100.

relationship between risk preference and subsequent job success are presented in Table 2. As in Experiment 1, these measures proved to be highly intercorrelated (average $r = .74$). Hence composite Z scores summing across three measures were again calculated, and analyses were conducted on these composite scores, illustrated in Figure 2.[4]

From these data, it is clear that exposure to the two initial case studies again exerted a powerful effect on subjects' later beliefs; no-debriefing subjects exposed to a positive relationship saw risky responses as indicative of future success, whereas those exposed to a negative relationship believed the opposite to be true, $F(1, 54) = 125.76$, $p < .0001$. These initial differences in beliefs between the positive and negative relationship conditions persisted even after subjects had been thoroughly debriefed in both the debriefing–explanation, $F(1, 54) = 69.74$, $p < .0001$, and the debriefing–no-explanation, $F(1, 54) = 19.24$, $p < .0001$, conditions. At the same time, it is also apparent that subjects' beliefs were not wholly unaffected by the debriefing manipulation. Within the no-explanation conditions, debriefed subjects endorsed less extreme theories than did nondebriefed subjects, as indicated by a significant Relationship × Debriefing interaction within these four cells, $F(1, 54) = 23.31$, $p < .0001$.

To test our further hypothesis that theory perseverance would be enhanced when subjects had been specifically induced to provide an explanation of the theory required a consideration of the interaction between the explanation and relationship manipulations within the four debriefing conditions of this study. The relevant interaction term, as predicted, showed that the process of explaining the relationship observed in the initial case studies significantly enhanced the perseverance

[4] As in Experiment 1, separate analyses parallel to those performed on the composite belief measure were again performed for each of the three component measures separately. All effects to be reported below proved individually significant ($p < .05$) for each of the three measures, with two exceptions: For the perceived criterion validity measure, the Explanation × Relationship interaction term was not individually significant, $F(1, 54) = 1.71$, and for the generalization to new cases measure, the post-debriefing perseverance effect within the no-explanation conditions was not significant ($F < 1$). As is evident from Table 2, however, even in these cases the means fall in the predicted direction.

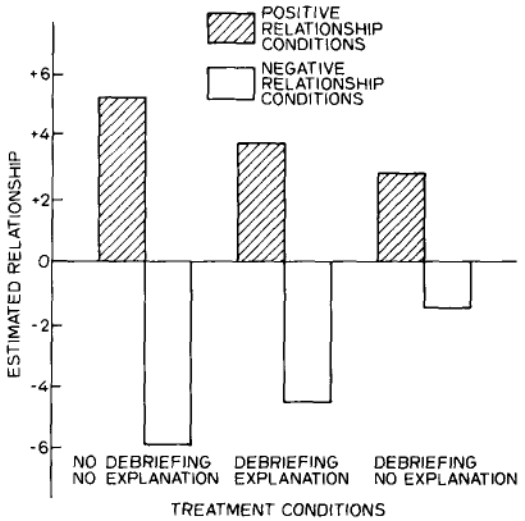

Figure 2. Mean composite indices (*Z* scores summed across the three measures) of subjects' personal estimates of the true relationship between risk preference and firefighter success, Experiment 2. (Positive scores indicate belief in a positive relationship; negative scores indicate belief in a negative relationship.)

of subjects' beliefs in the relationship they had discovered, $F(1, 54) = 7.86$, $p < .01$.[5]

General Discussion

The results of these two experiments provide support for three general conclusions. First, they offer further evidence for the basic hypothesis that people often cling to their beliefs to a considerably greater extent than is logically or normatively warranted. The experiments thus extend previous research on the perseverance of specific self-assessments and interpersonal judgments to the more general domain of theories concerning relationships among social variables. Second, these studies also extend prior research by suggesting that initial beliefs may persevere in the face of a subsequent invalidation of the evidence on which they are based, even when this initial evidence is itself as weak and inconclusive as a single pair of dubiously representative cases. Finally, the current results provide support for the hypothesis that belief perseverance effects may be mediated, in part, by the generation of causal explanations or scenarios that continue to imply the correctness of one's initial beliefs even in the later absence of any directly relevant evidence.

Before considering the implications of these findings in greater detail, however, it is important to consider two potential alternative explanations of these results. A first possibility is that subjects in the debriefing conditions simply failed to understand or believe the discrediting information provided during debriefing. Some presumptive evidence against this argument, of course, is provided by the fact that the debriefing procedure *did* seem to affect subjects' subsequent beliefs. Debriefed subjects consistently reported less strong beliefs than subjects who were not debriefed, an effect that proved individually significant in the second experiment. Perhaps more persuasive, however, were the results of postexperimental discussions with subjects in which they indicated that they were neither confused about nor suspicious of the debriefing manipulation: In all cases, during this final debriefing period, subjects seemed able to describe accurately the content of the discrediting procedure, and in no case did subjects voice any suspicion that this information may have been untrue. Instead, they simply felt that the relationship they had examined—whether positive or negative—appeared to be the correct one and that the discrediting of the evidential value of the initial cases was largely irrelevant to their personal beliefs concerning the "true" relationship existing between these variables. Interestingly, in their attempts to aid us as experimenters, several subjects expressed concern over the implausibility of the relationship given to

[5] The difference between the designs of Experiments 1 and 2 meant that only half as many subjects in the latter experiment were called on to provide explanations prior to debriefing. Nevertheless, internal analysis examining the relationship between the presence of some generalized explanatory scenario or principle in subjects' explanations and the perseverance of subjects' beliefs following debriefing was again performed. Although in the same direction as in the first experiment, this correlation was not significant, $r = .21$ (nor was it significantly different from the comparable correlation obtained in Experiment 1). The weighted average of these correlations (McNemar, 1962), we should note, remains highly significant, $r_{AV} = .44$, $p < .0005$.

subjects in the opposite relationship condition. As one subject in a positive relationship condition commented, "I don't think your experiment will work, since it will be impossible to convince anybody that the true relationship is negative." In short, we are led to discount any alternative account of the present results that rests on an alleged failure of subjects to comprehend or accept the discrediting information provided.

A second alternative explanation involves the possibility that the check on our manipulation of subjects' initial theories in these studies, because it required subjects to state explicitly their beliefs prior to debriefing, may have introduced a potential artifact whereby subjects subsequently reported beliefs consistent with their prior statements, via classic dissonance or self-perception processes. This appears improbable for two reasons. First, subjects' responses to the manipulation check were obtained under precisely the type of conditions—high justification, low choice, and little consequence—that should theoretically have minimized any experience of dissonance or conviction that their responses reflected their enduring beliefs. In fact, subjects at this point were asked only to describe the relationship that they had detected in the two case studies at hand, not their beliefs about the relationship that might exist between the two variables in general. Second, available data suggest that the present effects do not depend on the presence of such a manipulation check. If this proposed alternative were correct, a condition in which subjects were not asked to state their beliefs prior to debriefing should not show a perseverance effect. In a previous study that made use of the same experimental materials and basic procedures as the present experiments (Anderson & Ross, Note 3), however, significant perseverance effects, paralleling those reported here, were obtained in the absence of such manipulation checks.

In light of the apparent clarity of our results, then, let us consider further the implications of belief perseverance effects. We should make clear that a tendency to persevere in one's previous beliefs is not always detrimental, or even illogical. Indeed, in cases where our existing theories have already received significant support from different sources, it might be more irrational *not* to view subsequent challenges to those beliefs, or the evidence on which they were initially based, with an appropriately jaundiced eye (cf. Lord, Ross, & Lepper, 1979; Nisbett & Ross, 1980; Ross & Lepper, in press). To be buffeted about by every random piece of disconfirming data or every challenge to the evidential basis for one's beliefs, whether in the course of scientific inquiry or in our daily lives, will frequently prove less adaptive than a tendency to persist in theories that have proven effective over time. What is particularly striking in the present study, therefore, is the demonstration of perseverance biases in a situation where subjects' theories were initially grounded in the most minimal of data sets—only two case studies. That subjects will persevere in beliefs with such weak empirical grounding in the face of a complete refutation of the formative evidence for those beliefs seems eloquent testimony to the pervasiveness of our propensity to resist changing our attitudes or beliefs.

Even under these relatively extreme circumstances, however, it is not clear that every particular instance of belief perseverance should be viewed as unreasonable or counternormative. For example, exposure to even a demonstrably inadequate data set might lead one to appreciate the role of potential causal mechanisms that might have produced such a data set. Or it might lead one to recall or recognize additional evidence that is not subsequently undermined and that had heretofore been given insufficient weight in one's formulation of beliefs. Under such circumstances persistent changes in belief in the direction suggested by ultimately discredited data may be quite appropriate. In any given case, therefore, the rationality of such changes will depend on the status of the individual's own prior beliefs and of the evidence and reasoning that originally underlay those beliefs.

These caveats about overly quick and facile charges of "counternormativeness," however, should not obscure an essential feature of the present findings. Specifically, we demonstrated

that different subjects—and, by inference, the same subject, if placed in different experimental conditions—can be led, with equal ease through exposure to one data set or the other, to adopt and persevere in beliefs that are conceptually opposite to each other. The actual relationship between the two variables considered cannot be simultaneously both more positive and more negative than our subjects initially believed it to be. Our argument, then, is not that the mechanisms underlying belief perseverance are inherently irrational or inevitably dysfunctional. Rather, we are suggesting that these mechanisms, like many other processes underlying human inference (cf. Nisbett & Ross, 1980; Ross & Anderson, in press; Ross & Lepper, in press), may lead in certain contexts to a normatively unwarranted judgment, belief, and behavior.

Perhaps most importantly, the present findings provide some direct evidence concerning the processes postulated to underlie belief perseverance. Our research demonstrated that belief perseverance is enhanced when subjects are explicitly induced to explain the evidence they have been shown. Clearly, it is a central postulate of current attribution models that people often engage in informal causal analyses in attempting to make sense of their social worlds (Heider, 1958; Kelley, 1967, 1973). The likelihood of such attribution processes presumably varies with the novelty, complexity, and personal relevance of the events or outcomes one is attempting to understand (as well as with the direct instructional manipulations of the sort employed in the present studies). Such explanations may differ dramatically in their logical and formal properties, and may range from the postulation of a set of antecedent conditions necessary or sufficient to produce a given effect or outcome to a more simple imagination of some relatively concrete scenario in which the event or outcome follows some previous state or condition. In the present studies, for example, subjects made frequent use of one or another of two basic scenarios to explain either a positive or a negative relationship between risk preference and firefighting ability. In the positive relationship conditions, subjects' explanations often included some contrast between the successful firefighter braving enormous personal risks to save the occupants of a burning building and his unsuccessful counterpart, playing it safe, who stands by helplessly as lives are lost. In the negative relationship conditions, by contrast, subjects' explanations often made reference to an alternative scenario in which the successful firefighter carefully weights the relevant risks before taking appropriate and decisive action, while the unsuccessful trainee plunges headlong into danger, risking both his own and others' lives by foolhardy actions.

That such explanatory scripts or more formal and abstract explanations—whether generated spontaneously by subjects or produced in response to direct experimental instructions—should increase belief perseverance seems an obvious consequence of two properties of these explanatory accounts. First, such explanations are, by definition, selectively constructed to fit the evidence or outcome observed. Second, once created, such explanations become largely autonomous of the initial data that led to their postulation. Hence, they may remain available and continue to imply the existence of particular relationships or outcomes even if the data on which they were initially based subsequently prove to be completely devoid of evidential value. Evidence of a positive relationship between the presence of some general explanatory principle or script in subjects' explanations and the magnitude of subsequent belief perseverance effects, moreover, provides some direct evidence for such processes. Subjects who devoted equal time and effort to the explanation task but who focused solely on the specifics of the two subsequently discredited case studies did not seem to show increased perseverance.

There is, of course, one further assumption implicit in the foregoing account of our findings—that subjects, having explained the apparent existence of a particular relationship, are not led by our debriefing procedure to generate any corresponding counterexplanation to account for an opposite pattern of data that might have been observed (Ross & Lepper, in press; Slovic & Fischhoff, 1977).

Such an assumption is consistent with the proposition that people are likely to engage in the cognitive effort required to generate an explanation only when provided with some salient or unexpected event or outcome (Fischhoff, 1977; Kanouse, 1971; Lepper, Zanna, & Abelson, 1970). It suggests, in addition, an interesting potential antidote for unwarranted belief perseverance in the face of later challenges to the evidence on which our beliefs were based. Would such perseverance effects be eliminated or attenuated, for example, if subjects could be led, after debriefing, to consider explicitly the explanations that might be offered to support a contention in opposition to their initial beliefs? Alternatively, could subjects be "innoculated" against perseverance effects if they had been asked, at the outset of the study, to list all of the possible reasons they could imagine that might have produced either a positive or a negative relationship between the two variables being studied (cf. Slovic & Fischhoff, 1977)? In view of the pervasive significance of our social theories in decision-making contexts (cf. Abelson, 1976; Janis & Mann, 1977), and the demonstrable adverse effects of unwarranted belief perseverance in applied and clinical contexts (Allport, 1954; Chapman & Chapman, 1969; Janis & Mann, 1977; Lepper, Ross, & Lau, Note 2), the effectiveness of such debiasing techniques clearly merits further investigation.

Reference Notes

1. Jennings, D. L., Lepper, M. R., & Ross, L. *Persistence of impressions of personal persuasiveness: Perseverance of erroneous self-assessments outside the debriefing paradigm.* Unpublished manuscript, Stanford University, 1980.
2. Lepper, M. R., Ross, L., & Lau, R. *Persistence of inaccurate and discredited personal impressions: A field demonstration of attributional perseverance.* Unpublished manuscript, Stanford University, 1980.
3. Anderson, C. A., & Ross, L. *The survival of theories in the absence of evidence.* Paper presented at the Meeting of the Western Psychological Association, San Francisco, 1978.

References

Abelson, R. P. Script processing in attitude formation and decision-making. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior.* Hillsdale, N.J.: Erlbaum, 1976.

Allport, G. W. *The nature of prejudice.* Reading, Mass.: Addison-Wesley, 1954.

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

Fischhoff, B. Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance,* 1975, *1,* 288–299.

Fischhoff, B. Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance,* 1977, *3,* 349–358.

Heider, F. *The psychology of interpersonal relations.* New York: Wiley, 1958.

Janis, I. L., & Mann, L. *Decision making: A psychological analysis of conflict, choice, and commitment.* New York: Free Press, 1977.

Kanouse, D. E. Language, labeling, and attribution. In E. E. Jones, et al. (Eds.), *Attribution: Perceiving the causes of behavior.* Morristown, N.J.: General Learning Press, 1971.

Kelley, H. H. Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (Vol. 15). Lincoln: University of Nebraska Press, 1967.

Kelley, H. H. The processes of causal attribution. *American Psychologist,* 1973, *28,* 107–128.

Lepper, M. R., Zanna, M. P., & Abelson, R. P. Cognitive irreversibility in a dissonance-reduction situation. *Journal of Personality and Social Psychology,* 1970, *16,* 191–198.

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

McNemar, Q. *Psychological statistics.* New York: Wiley, 1962.

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

Ross, L., & Anderson, C. Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press, in press.

Ross, L., & Lepper, M. R. The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder & D. Fiske (Eds.), *New directions for methodology of behavioral sciences: Fallible judgment in behavioral research.* San Francisco: Jossey-Bass, in press.

Ross, L., Lepper, M. R., & Hubbard, M. Perseverance in self-perception and social perception:

Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology,* 1975, *32,* 880–892.

Ross, L., Lepper, M. R., Strack, F., & Steinmetz, J Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology,* 1977, *35,* 817–829.

Slovic, P., & Fischhoff, B. On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance,* 1977, *3,* 544–551.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1971.

Received August 20, 1979 ■