Journal of Personality and Social Psychology
1988, Vol. 54, No. 4, 569-579

0022-3514/88/$00.75

Outcome Bias in Decision Evaluation

Jonathan Baron
University of Pennsylvania

John C. Hershey
Department of Decision Sciences
University of Pennsylvania

In 5 studies, undergraduate subjects were given descriptions and outcomes of decisions made by others under conditions of uncertainty. Decisions concerned either medical matters or monetary gambles. Subjects rated the quality of thinking of the decisions, the competence of the decision maker, or their willingness to let the decision maker decide on their behalf. Subjects understood that they had all relevant information available to the decision maker. Subjects rated the thinking as better, rated the decision maker as more competent, or indicated greater willingness to yield the decision when the outcome was favorable than when it was unfavorable. In monetary gambles, subjects rated the thinking as better when the outcome of the option not chosen turned out poorly than when it turned out well. Although subjects who were asked felt that they should not consider outcomes in making these evaluations, they did so. This effect of outcome knowledge on evaluation may be explained partly in terms of its effect on the salience of arguments for each side of the choice. Implications for the theory of rationality and for practical situations are discussed.

> A fault condemned but seldom avoided is the evaluation of the intention of an act in terms of the act's outcome. An agent who acted as wisely as the foreseeable circumstances permitted is censured for the ill-effects which come to pass through chance or through malicious opposition or through unforeseeable circumstances. Men desire to be fortunate as much as they desire to be wise, but yet they fail to discriminate between fortune and wisdom or between misfortune and guilt. . . . We are ingenious in 'discovering' the defect of character we believe would account for a person's misfortune. (Arnauld, 1662/1964, p. 285)

> Since good decisions can lead to bad outcomes (and vice versa) decision makers cannot infallibly be graded by their results. (Brown, Kahr, & Peterson, 1974, p. 4)

> A good decision cannot guarantee a good outcome. All real decisions are made under uncertainty. A decision is therefore a bet, and evaluating it as good or not must depend on the stakes and the odds, not on the outcome. (Edwards, 1984, p. 7)

Evaluations of decisions are made in our personal lives, in organizations, in judging the performance of elected officials, and in certain legal disputes such as malpractice suits, liability cases, and regulatory decisions. Because evaluations are made after the fact, there is often information available to the judge that was not available to the decision maker, including information about the outcome of the decision. It has often been suggested that such information is used unfairly, that reasonable decisions are criticized by Monday-morning quarterbacks who think they might have decided otherwise, and that decision makers end up being punished for their bad luck (e.g., Arnauld, 1662/1964; Berlin, 1984; Nichols, 1985).

The distinction between a good decision and a good outcome is a basic one to all decision analysts. The quotation from Edwards (1984) cited earlier is labeled by the author as "a very familiar elementary point" (p. 7). In this paper, we explore how well the distinction between decisions and outcomes is recognized outside the decision-analysis profession.

Information that is available only after a decision is made is irrelevant to the quality of the decision. Such information plays no direct role in the advice we may give decision makers ex ante or in the lessons they may learn (Baron, 1985, chapter 1). The outcome of a decision, by itself, cannot be used to improve a decision unless the decision maker is clairvoyant.

Information about possible outcomes and their probabilities falls into three relevant classes: *actor information,* known only to the decision maker at the time the decision is made; *judge information,* known only to the judge at the time the decision is evaluated; and *joint information,* known both to the decision maker at the time of decision and to the judge at the time of evaluation. (In some cases, the decision maker and the judge will be the same person, at different times.) In the cases we consider, the judge has the outcome information and the actor does not.

Although outcome information plays no direct role in the evaluation of decisions, it may play an appropriate indirect role. In particular, it may affect a judge's beliefs about actor information. A judge who does not know the decision maker's probabilities may assume that the probability was higher for an outcome that occurred than for the same outcome had it not occurred. (Note, however, that outcome information tells us nothing about the *utilities* of a decision maker, even if we have no other information about them.) In the extreme, if we have no information except outcome, it is a reasonable prima facie hypothesis that bad outcomes (e.g., space-shuttle accidents) result from

This work was supported by grants from the National Institute of Mental Health (to Jonathan Baron, MH37241) and from the National Science Foundation (to Jonathan Baron and John C. Hershey, SES-8509807).

We thank Mark Spranca and several reviewers for many helpful suggestions.

Both authors are senior fellows of the Leonard Davis Institute for Health Economics.

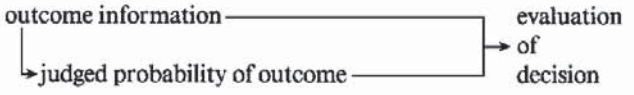
Correspondence concerning this article should be addressed to Jonathan Baron, Psychology Department, University of Pennsylvania, 3815 Walnut Street, Philadelphia, Pennsylvania 19104-6916.

badly made decisions. We do not usually set up commissions of inquiry to delve into policy decisions that turn out well.

Another appropriate indirect role of outcome information is that it allows decision makers to modify beliefs about probabilities in similar situations. If they know nothing about the proportion of red cards in a deck, they can learn something about that proportion by drawing cards from the deck. (However, if they know that the deck is an ordinary one, sampled with replacement, they learn nothing by drawing cards.) This effect of outcome information can operate only within a sequence of similar decisions, not in a single decision.

At issue here is whether there is an *outcome bias,* in which people take outcomes into account in a way that is irrelevant to the true quality of the decision. This sort of bias is not established by showing that people take outcomes into account. As we argued earlier, outcomes are relevant when they can inform us about actor information. One way to show an outcome bias is to give the judge all relevant information about outcome probabilities known to the decision maker, plus the outcome. That is, there is only joint information and judge information (the outcome), no actor information.

Information (relevant or irrelevant) may have two effects on evaluations: (a) an effect on the judged probability of outcomes, which, in turn, affects evaluation; and (b) a direct effect on the judged quality of the decision, as shown below:



For example, we may think a decision is bad if we believe that bad outcomes were highly probable, but outcome information may also affect our evaluation even if the probability of an outcome is known.

Fischhoff (1975) demonstrated the existence of a *hindsight bias,* an effect of outcome information on the judged probability of an outcome. Subjects were given scenarios and asked to provide probabilities for different outcomes. When subjects were told the outcome and asked what probability other subjects who did not know the outcome (or they themselves if they did not know it) would give, they gave higher probabilities than those given by actual other subjects not told the outcome (or told that some other outcome had occurred). Note that these demonstrations filled our condition of eliminating actor information (where the actors were the other subjects). Subjects were asked to judge the probability for someone who had exactly the same information they had (except for outcome), no more.

Although it seems likely that the hindsight bias would lead to biased evaluations of decision quality, this has not been shown, nor is it what we seek to show here. Rather, we seek a direct effect of outcome on evaluation of decisions, an effect that does not operate through an effect of outcome knowledge on a judge's assessed probabilities of outcomes. To this end, we held probability information constant by telling subjects that probabilities were known, or by otherwise limiting probability information. Of course, in real life, the outcome bias we seek could work together with the hindsight bias (as shown in the diagram) to distort evaluations of decisions even more than either bias alone.

Zakay (1984) showed that managers counted good outcomes as one of the criteria for evaluating decisions made by other managers. However, as we have argued, it is perfectly reasonable to do this when there are facts known only to the decision maker (actor information). At issue in this article is not whether people use outcome information but whether there are conditions under which they overuse it. Thus, we look for an effect of outcome information when the subject is told everything that is relevant. In this case, outcome should play no role in our evaluations of decisions, although we hypothesize that it will.

The outcome bias we seek may be related to Walster's (1966) finding that subjects judged a driver as more "responsible" for an accident when the damage was more severe. However, questions about responsibility might be understood as concerning the appropriate degree of punishment or blame rather than rationality or quality of decision making. As a general rule, it makes sense to punish actors more severely for more severe consequences; it is usually difficult to know what the actor knew, and severity of consequences is a clue as to the degree of negligence. Even when we know what the actor knew, use of this general rule may set clearer precedents for others (as in the utilitarian rationale for "punishing the innocent"). Walster apparently intended the question about responsibility to tap subjects' beliefs about the extent to which the driver could have prevented the accident by acting differently. Walster suggested that her results were due to subjects' desire to believe that events were controllable: If bad outcomes are caused by poor decisions or bad people, we can prevent them by correcting the decision making or by punishing the people. If subjects interpreted the question this way, they would be making an error, but not the same error we seek in this study.

Similarly, studies of the effect of outcomes on children's moral judgments (e.g., Berg-Cross, 1975; Leon, 1982; Stokes & Leary, 1984; Surber, 1977) have used judgments of responsibility, deservingness of punishment, and badness, each of which could be appropriately affected by outcome. Also, in most cases no effort was made to provide the judge with all relevant information available to the actor.

Mitchell and Kalb (1981) also showed effects of outcome knowledge on judgments of both responsibility for outcomes and outcome probability. Subjects (nurses) read descriptions of poor performance by nurses (e.g., leaving a bed railing down) that either resulted in poor outcomes (e.g., the patient fell out of bed) or benign outcomes. In fact, outcome knowledge affected both probability judgments and responsibility judgments. Although the former effect might have been a hindsight bias, it might also have been an appropriate inference about actor information: Outcome information might have provided information about factors that affected outcome probability from the decision maker's viewpoint (e.g., whether the patient was alert and, if not, whether she slept fitfully). Mitchell and Kalb argued that the effect of outcome on probability did not explain the effect on responsibility judgment: The correlation between judged probability and judged responsibility, with outcome held constant, was nonsignificant across subjects. Of course, the problem still remains that the term *responsibility* need not refer only to quality of the decision.

In our experiments, instead of examining the correlation between outcome judgments and probability judgments, we fixed the outcome probabilities by telling the subjects what they were

from the decision maker's viewpoint. We also explicitly asked about the "quality of thinking." All decisions were expressed in the form of gambles. For example, an operation may lead to a cure or to death, with given probabilities. We gave the subjects probabilities of all possible outcomes and brief descriptions of each outcome. It is reasonable to assume that the quality of the decision would depend on the probabilities of the outcomes—which summarize all the information we have about uncertain states of the world that could affect the outcome—and the desirabilities or utilities of the outcomes. Although we did not provide all necessary information about desirabilities, the outcome provided no additional information on this score. In our studies an outcome bias existed when the evaluation of the decisions depended on their outcomes.

We expected to find an outcome bias because the generally useful heuristic of evaluating decisions according to their outcomes may be overgeneralized to situations in which it is inappropriate. It may be learned as a rigid rule, perhaps from seeing punishment meted out for bad outcomes resulting from reasonable decisions.

Of course, it can often be appropriate to use outcome information to evaluate decision quality, especially when actor information is substantial relative to judge information or joint information and when it is necessary to judge decisions by their outcomes (as fallible as this may be) simply because there is little other useful information. This is especially true when decision makers are motivated to deceive their evaluators about the nature of their own information.

Ordinarily, it is relatively harmless to overgeneralize the heuristic of evaluating decisions according to their outcomes. However, when severe punishments (as in malpractice suits) or consequential decisions (as in elections) are contingent on a judgment of poor decision making, insight into the possibility of overgeneralization may be warranted.

A second reason for outcome bias is that the outcome calls attention to those arguments that would make the decision good or bad. For example, when a patient dies on the operating table, this calls attention to the risk of death as an argument against the decision to perform surgery. When subjects attempt to reexamine the arguments to consider what they would have thought if they had not been told the outcome, the critical information remains salient. Fischhoff (1975) found an analogous mechanism to be operating in hindsight bias. When subjects were asked to rate the relevance to their judgment of each item in the scenario, the relevance of the items depended on the outcome subjects were given. Note that the salience of an argument based on risk or possible benefit may not be fully captured by a description of the subjective probability and utility of the outcome in question.

One type of argument for or against a decision concerns the difference between outcomes resulting from different decisions in otherwise identical states of the world. For example, a decision to buy a stock or not may compare one's feelings about buying or not buying if the stock goes up (rejoicing vs. regret), or if the stock goes down. Regret theory (Bell, 1982; Loomes & Sugden, 1982) explicitly takes such differences into account in explaining choice. Once the true state is revealed (e.g., the stock goes down), the judge may overweigh the regret associated with this state (the difference between buying and not buying in this case) when judging decision quality.

Another type of argument is that a bad outcome may be avoided by considering choices other than those considered so far, or by gathering more information about probabilities (Toda, 1984). Such arguments are equally true regardless of whether the outcome is good or bad (Baron, 1985), but a bad outcome may make them more salient. In many of our examples, there is no possibility of additional choices or information.

A third reason for outcome bias is that people may regard luck as a property of individuals. That is, people may act as if they believe that some people's decisions are influenced by unforeseeable outcomes. Such a belief might have been operating in the experiments of Langer (1975), who found that people were less willing to sell their lottery tickets when they had chosen the ticket number themselves than when the numbers had been chosen for them. Langer interpreted this finding (and others like it) in terms of a confusion between chance and skill, but the skill involved might have been the sort of clairvoyance described earlier. (The results of Lerner & Matthews, 1967, may be similarly explained.) Our experiments did not test this explanation directly, but we mention it here for completeness.

Experiment 1

Method

Materials and procedure. Subjects were given a questionnaire with a list of 15 medical decisions. They were asked to evaluate each decision on the following 7-point scale:

3 = clearly correct, and the opposite decision would be inexcusable;

2 = correct, all things considered;

1 = correct, but the opposite would be reasonable too;

0 = the decision and its opposite are equally good;

−1 = incorrect, but not unreasonable;

−2 = incorrect, all things considered;

−3 = incorrect and inexcusable.

The subjects were encouraged to use intermediate numbers if they wished and to explain answers that would not be obvious. They were reminded "to evaluate the decision itself, the quality of thinking that went into it."

The 15 cases are listed in Table 1. Case 1 read as follows:

> A 55-year-old man had a heart condition. He had to stop working because of chest pain. He enjoyed his work and did not want to stop. His pain also interfered with other things, such as travel and recreation. A type of bypass operation would relieve his pain and increase his life expectancy from age 65 to age 70. However, 8% of the people who have this operation die from the operation itself.[1] His physician decided to go ahead with the operation. The operation succeeded. Evaluate the physician's decision to go ahead with the operation.

Case 2 was the same except that the operation failed and the man died. Cases 3 and 4 paralleled Cases 1 and 2, respectively, except that

[1] The 8% figure was chosen on the basis of pilot data to make the decision appear difficult to the subjects.

Table 1
Conditions and Mean Ratings for Experiment 1

Case	Choice	Decision maker	Outcome	*M*	*SD*
1	Heart surgery	Physician	Success	0.85	1.62
2	Heart surgery	Physician	Failure	−0.05	1.77
3	Heart surgery	Patient	Success	1.00	1.05
4	Heart surgery	Patient	Failure	0.75	1.26
5	Liver surgery	Physician	Success	0.45	1.75
6	Liver surgery	Physician	Failure	−0.30	1.79
7	Liver surgery	Patient	Success	1.05	1.02
8	Liver surgery	Patient	Failure	0.35	1.24
9	Test, positive, treat	Physician	Success	1.40	1.83
10	Test, negative, treat	Physician	Success	1.15	1.75
11	Test, negative, treat	Physician	Failure	1.20	1.83
12	Test 1, Disease A	Physician	Success	−0.07	1.57
13	Test 1, Disease A	Physician	Failure	−1.30	0.71
14	Test 1, Disease B	Physician	Success	−0.22	1.69
15	Test 1, Disease B	Physician	Failure	−1.35	1.28

the man made the decision rather than the physician and the man's decision was the one that was evaluated. Cases 5 through 8 paralleled Cases 1 through 4, except that a liver ailment rather than a heart ailment was described.

Cases 9 through 11 involved a testing situation of the type studied by Baron, Beattie, and Hershey (in press). A test was described that had such poor accuracy that the best action, on normative grounds, would have been to treat the patient (for a foot infection with using an antibiotic) regardless of the test result. In Case 9, which was included for a purpose not addressed in this article, the test was positive and the disease was treated and cured. In Cases 10 and 11, the test was negative but the disease was treated anyway; it was cured in Case 10 but not in Case 11. Subjects were asked to evaluate whether the physician was correct in ordering the worthless test. A comparison of Cases 10 and 11, which differed in success versus failure, could also be used to look for an outcome bias.

Cases 12 through 15 concerned a choice between two tests in order to decide which of two diseases to treat (as studied by Baron & Hershey, in press). The two diseases, *A* and *B*, were considered equally likely. Test 1 indicated Disease A correctly in 92% of patients with *A* and Disease B correctly in 80% of patients with *B*. Test 2 indicated Disease A correctly in 86% of patients with *A* and Disease B correctly in 98% of patients with *B*. If *A* was treated (by surgery), the treatment was always successful, but if *B* was treated, the treatment was successful one third of the time. (Normatively, the two tests were equally good, because errors in detecting *A* were three times as costly as errors in detecting *B*). The physician always chose Test 1. In Cases 12 and 13, the test indicated *A;* in Cases 14 and 15, it indicated *B*. In Cases 12 and 14, the operation succeeded; in Cases 13 and 15, it failed. Subjects were asked to evaluate the physician's decision to perform Test 1.

The cases were presented in a within-subjects design. Cases to be compared were separated in the sequence as widely as possible. (The sequence used was 2, 5, 13, 10, 3, 8, 15, 9, 1, 6, 12, 11, 4, 7, and 14.) Note that a within-subjects design makes it easier to distinguish small effects from random error but at the cost of reducing the magnitude of effects because subjects may remember responses they gave to similar cases.

Subjects. Subjects were 20 undergraduates at the University of Pennsylvania, obtained through a sign placed on a prominent campus walkway and paid by the hour. Ten subjects did the cases in the order given; 10 did them in reverse order.

Results

In our analysis, we defined an outcome bias as the mean rating assigned to cases with positive outcomes minus the mean rating for cases with negative outcomes. Mean ratings of all cases are shown in Table 1. Overall, there was an outcome bias. Cases in which the outcome was success (Cases 1, 3, 5, 7, 10, 12, and 14) were rated higher than matched cases in which the outcome was failure (Cases 2, 4, 6, 8, 11, 13, and 15): mean effect = 0.70, $t(19) = 4.04$, $p < .001$, one-tailed. For the two orders, respectively, $ts(9) = 3.10$ and 2.51, both $ps < .025$. In 44.3% of the 140 pairs of cases that differed only in success or failure, higher ratings were given to the case with success; in 9.3% higher ratings were given to the case with failure, and in 46.4% equal ratings were given to the two cases. (Many subjects said that they remembered their responses to previous cases and repeated them regardless of the outcome.) For each of the 7 pairs of comparable cases (e.g., 1 vs. 2), more subjects favored the success case than the failure case, except for Cases 10 and 11, in which the numbers were equal.

Subjects might have thought that physicians were more responsible for bad outcomes, or they might have believed that the physician had information that the patient did not have (despite our instructions to the contrary). However, the outcome bias was also found for just those cases (Cases 3 and 7 vs. 4 and 8) in which the patient made the decision rather than the physician: $M = 0.48$, $t(19) = 2.59$, $p < .01$. In 17 of the 40 pairs, the success case was rated higher; in 4 cases the failure case was rated higher.[2] This issue is addressed further in Experiment 4.

The last 8 subjects run were asked after the experiment whether they thought they should have taken outcome into account in evaluating the decisions. All but 1 subject said they should not, and 1 was unsure. The outcome bias was significant for the 7 subjects who said they should not, $t(6) = 3.26$, $p < .01$; for the cases in which the patient made the decision, $t(6) = 2.50$, $p < .025$. Of these 8 subjects, 2 (including the one who was unsure) volunteered that they thought they had taken outcome into account even though they should not have, and 4 said they had not taken outcome into account. The outcome bias shown by the latter 4 subjects was 0.43, 0.29, 1.43, and 0.71, respectively. It would appear that most subjects accept the irrelevance of outcomes to judgments of rationality, they show an outcome bias even though they think they should not, and some show an outcome bias even though they think they do not. Further evidence on subjects' normative beliefs was obtained in Experiment 4.

Experiment 2

In Experiment 2, subjects were asked to rate the importance of several factors in a decision. This allowed us to test the effect

[2] However, the outcome bias appeared to be greater when the physician made the decision ($M = .80$), $t(19) = 3.56$, than when the patient made the decision; for the difference in effects, $t(19) = 2.04$, $p = .05$, two tailed.