31. Debiasing

Baruch Fischhoff

Once a behavioral phenomenon has been identified in some experimental context, it is appropriate to start questioning its robustness. A popular and often productive questioning strategy might be called destructive testing, after a kindred technique in engineering. A proposed design is subjected to conditions intended to push it to and beyond its limits of viability. Such controlled destruction can clarify where it is to be trusted and why it works when it does. Applied to a behavioral phenomenon, this philosophy would promote research attempting to circumscribe the conditions for its observation and the psychological processes that must be evoked or controlled in order to eliminate it. Where the phenomenon is a judgmental bias, destructive testing takes the form of debiasing efforts. Destructive testing shows where a design fails; when a bias fails, the result is improved judgment.

The study of heuristics and biases might itself be seen as the application of destructive testing to the earlier hypothesis that people are competent intuitive statisticians. Casual observation suggests that people's judgment is generally "good enough" to let them make it through life without getting into too much trouble. Early studies (Peterson & Beach, 1967) supported this belief, indicating that, to a first approximation, people might be described as veridical observers and normative judges. Subsequent studies, represented in this volume, tested the accuracy of this approximation by looking at the limits of people's apparent successes. Could better judgment have made them richer or healthier? Can the success they achieved be attributed to a lenient environment, which does not presume particularly knowledgeable behavior? Tragic mistakes provide important insight into the nature and quality of people's decision-

My thanks to Ruth Beyth-Marom, Don MacGregor, and Paul Slovic for their helpful comments on earlier drafts of this paper. This work was supported by the Office of Naval Research under Contract N00014-80-C-0150 to Perceptronics, Inc.

making processes; fortunately, they are rare enough that we have too small a data base to disentangle the factors that may have led people astray. Judgment research has used the destructive-testing strategy to generate biased judgments in moderately well-characterized situations. The theoretician hopes that a pattern of errors and successes will emerge that lends itself to few possible explanations. Thus, the study of biases clarifies the sources and limits of apparent wisdom, just as the study of debiasing clarifies the sources and limits of apparent folly. Both are essential to the study of judgment.

Although some judgment studies are primarily demonstrations that a particular bias can occur under some, perhaps contrived, conditions, many other studies have attempted to stack the deck against the observation of bias. Some of these are explicitly debiasing studies, conducted in the hope that procedures that prove effective in the laboratory will also improve performance in the field. Others had the more theoretical goal of clarifying the contexts that induce suboptimal judgments. The core of this chapter is a review of studies that can be construed as efforts to reduce two familiar biases, hindsight bias and overconfidence. It considers failures as well as successes in the belief that (a) failure helps clarify the virulence of a problem and the need for corrective or protective measures, and (b) the overall pattern of studies is the key to discovering the psychological dimensions that are important in characterizing real-life situations and anticipating the extent of biased performance in them.

The review attempts to be exhaustive, subject to the following three selection criteria:

1. Only studies published in sources with peer review are considered. Thus, responsibility for quality control is externalized.
2. Anecdotal evidence is (with a few exceptions) excluded. Although such reports are the primary source of information about some kinds of debiasing attempts (e.g., use of experts), they are subject to interpretive and selection biases that require special attention beyond the scope of this summary (see Chap. **23**).
3. Some empirical evidence is offered. Excluded are suggestions that have yet to be tested and theoretical arguments (e.g., about the ecological validity of experiments) that cannot be tested.

Prior to that review, a framework for debiasing efforts will be offered, characterizing possible approaches and the assumptions underlying them. Such a framework might reveal recurrent patterns when applied to a variety of judgmental biases.

Debiasing methods

When there is a problem, it is natural to look for a culprit. Debiasing procedures may be most clearly categorized according to their implicit

https://doi.org/10.1017/CBO9780511809477.032 Published online by Cambridge University Press

Table 1. *Debiasing methods according to underlying assumption*

Assumption	Strategies
Faulty tasks	
Unfair tasks	Raise stakes
	Clarify instructions/stimuli
	Discourage second-guessing
	Use better response modes
	Ask fewer questions
Misunderstood tasks	Demonstrate alternative goal
	Demonstrate semantic disagreement
	Demonstrate impossibility of task
	Demonstrate overlooked distinction
Faulty judges	
Perfectible individuals	Warn of problem
	Describe problem
	Provide personalized feedback
	Train extensively
Incorrigible individuals	Replace them
	Recalibrate their responses
	Plan on error
Mismatch between judges and task	
Restructuring	Make knowledge explicit
	Search for discrepant information
	Decompose problem
	Consider alternative situations
	Offer alternative formulations
Education	Rely on substantive experts
	Educate from childhood

allegation of culpability. The most important distinction is whether responsibility for biases is laid at the doorstep of the judge, the task, or some mismatch between the two. Do the biases represent artifacts of incompetent experimentation and dubious interpretation, clear-cut cases of judgmental fallibility, or the unfortunate result of judges having, but misapplying, the requisite cognitive skills? As summarized in Table 1, and described below, each of these categories can be broken down further according to what might be called the depth of the problem. How fundamental is the difficulty? Are technical or structural changes needed? Strategies for developing debiasing techniques are quite different for the different causal categories.

Faulty tasks

Unfair tasks. Experimentalists have standard questions that they pose to their own and others' work. Studies are published only if they instill

confidence (in reviewers and editors) that the more obvious artifacts have been eliminated. Since, however, it is impossible to control for everything and satisfy everyone in an initial study or series of studies, the identification of putative methodological artifacts is a first line of attack in attempting to discredit an effect. Among the claims that may be raised are: (a) Subjects did not care about the task – therefore one should raise the stakes accruing to good performance; (b) subjects were confused by the task – therefore use more careful instructions and more familiar stimuli; (c) subjects did not believe the experimenters' assertions about the nature of the task or perceived a payoff structure other than that intended by the experimenter – therefore assure them that their best guess at the right answer is all that is of interest and that they should respond as they see fit; (d) subjects were unable to express what they know – therefore use more familiar or pliable response modes; (e) subjects were asked too many questions and developed stereotypic response patterns to help them get through the task – therefore ask fewer questions (or define one's research interest as stereotypic responses).

Coping with such problems is part of good scientific hygiene. However, such efforts usually have little theoretical content. Since its goal is producing a better experimental environment, the study of artifacts may not even be very informative about the universe of contexts to which observed results can be safely generalized. "Successful" artifact studies provide primarily negative information, casting doubt on whether an effect has been observed in "fair" conditions. Whether life is "fair" in the same sense, when it poses questions, is a separate issue.

Misunderstood tasks. Artifact studies carry an implicit aspersion of experimental malpractice. The original investigator should have known better or should have been more careful. Such allegations are less appropriate with a second kind of task deficiency: the failure of the investigator to understand respondents' phenomenology or conceptual universe. Reformulation of the task to clarify what subjects were really doing has been used by critics of the heuristics-and-biases approach as well as by its promulgators. Among the ways one might try to show the wisdom of apparently biased behavior are: (a) demonstrating some alternative goal that is achieved by sacrificing optimality in the task at hand (e.g., learning about the properties of a system by making diagnostic mistakes); (b) demonstrating that respondents share a definition of key terms different from that held or presumed by the experimenter; (c) demonstrating that the task could not be done unless respondents chose to make some additional assumptions that would have to concur fortuitously with those made by the experimenter; (d) demonstrating that subjects make a reasonable distinction to which the experimenter was insensitive.

To make a contribution, such reformulations should include empirical demonstrations, not just claims about "what subjects might have been

thinking." At their worst, such assertions can have a strong ad hoc flavor and defy falsification; indeed, contradictory versions may be used to explain away different biases. At their best, they can make strong theoretical statements about cognitive representations (Fischhoff, in press-a).

Faulty judges

Perfectible judges. If the task has been polished and the bias remains, the respondent must assume some responsibility. To eliminate an unwanted behavior, one might use an escalation design, with steps reflecting increasing pessimism about the ease of perfecting human performance: (a) warning about the possibility of bias without specifying its nature (this strategy differs from inspiring people to work harder by implying that the potential error is systematic and that respondents need instruction, not just a fair chance); (b) describing the direction (and perhaps extent) of the bias that is typically observed; (c) providing a dose of feedback, personalizing the implications of the warning; (d) offering an extended program of training with feedback, coaching, and whatever else it takes to afford the respondent cognitive mastery of the task.

Such steps fault the judge, not the task, by assuming that solutions will not emerge spontaneously or merely with careful question rephrasing. Although of great practical import, training exercises may have limited theoretical impact. The attempt to find something that works may create a grab bag of maneuvers whose effective elements are poorly defined. More systematic experimentation may then be needed to identify those elements. The ultimate goal is understanding how the artificial experience created by the training program differs from the natural experience that life offers. Why does one technique work to eliminate bias, while another does not?

Incorrigible judges. At some point, the would-be trainer may decide that success is impossible, or only attainable with procedures that coerce the subject to respond optimally. The "successes" that are obtained by essentially giving respondents the right answer or by creating unavoidable demand characteristics are bereft of both theoretical and practical interest. It is hardly news when people listen to what they are told; if they have to be told every time how to respond, who needs them?

Three options seem open in such situations: (a) replacing people with some superior answering device; (b) recalibrating fallible judgments to more appropriate values, assuming that the amount and direction of errors are predictable; (c) acknowledging the imprecision in people's judgments when planning actions based on them. The decision maker or decision analyst who has given up on people in any of these ways may still contribute to our understanding of judgment by assessing the size, preva-

lence, and resilience of such indelible biases. However, because improved judgment is not the intent of these corrective actions, they will be considered only cursorily here.

Mismatch between judge and task

Restructuring. Perhaps the most charitable, and psychological, viewpoint is to point no fingers and blame neither judge nor task. Instead, assume that the question is acceptably posed and that the judge has all requisite skills, but somehow these skills are not being used. In the spirit of human engineering, this approach argues that the proper unit of observation is the person-task system. Success lies in making them as compatible as possible. Just as a mechanically intact airplane needs good instrument design to become flyable, an honest (i.e., not misleading) judgment task may only become tractable when it has been restructured to a form that allows respondents to use their existing cognitive skills to best advantage.

Although such cognitive engineering tends to be task specific, a number of recurrent strategies emerge: (a) forcing respondents to express what they know explicitly rather than letting it remain "in the head"; (b) encouraging respondents to search for discrepant evidence, rather than collecting details corroborating a preferred answer; (c) offering ways to decompose an overwhelming problem to more tractable and familiar components; (d) suggesting that respondents consider the set of possible situations that they might have encountered in order to understand better the specific situation at hand; and (e) proposing alternative formulations of the presented problem (e.g., using different terms, concretizing, offering analogies).

Education. A variant on the people-task "systems" approach is to argue that people can do this task, but not these people. The alternatives are to use: (a) experts who, along with their substantive knowledge, have acquired some special capabilities in processing information under conditions of uncertainty; or (b) a new breed of individual, educated from some early age to think probabilistically. In a sense, this view holds that although people are not, in principle, incorrigible, most of those presently around are. Education differs from training (a previous category) in its focus on developing general capabilities rather than specific skills.

Hindsight bias: An example of debiasing efforts

A critical aspect of any responsible job is learning from experience. Once we know how something turned out, we try to understand why it happened and to evaluate how well we, or others, planned for it. Although such outcome knowledge is thought to confer the wisdom of

hindsight on our judgments, its advantages may be oversold. In hindsight, people consistently exaggerate what could have been anticipated in foresight. They not only tend to view what has happened as having been inevitable, but also to view it as having appeared "relatively inevitable" before it happened. People believe that others should have been able to anticipate events much better than was actually the case. They even misremember their own predictions so as to exaggerate in hindsight what they knew in foresight (Fischhoff, 1975). Although it is flattering to believe that we would have known all along what we could only know in hindsight, that belief hardly affords us a fair appraisal of the extent to which surprises and failures are inevitable. It is both unfair and self-defeating to castigate decision makers who have erred in fallible systems, without admitting to that fallibility and doing something to improve the system. By encouraging us to exaggerate the extent of our knowledge, this bias can make us overconfident in our predictive ability. Perception of a surprise-free past may portend a surpriseful future.

Research on this bias has included investigations of most of the possible debiasing strategies included in the previous section. Few of these techniques have successfully reduced the hindsight bias; none has eliminated it. They are described below and summarized in Table 2.

Faulty tasks

Unfair tasks. In an initial experimental demonstration of hindsight bias (Fischhoff, 1975), subjects read paragraph-long descriptions of a historical event and assessed the probability that they would have assigned to each of its possible outcomes had they not been told what happened. Regardless of whether the reported outcome was true or false (i.e., whether it happened in reality), subjects believed that they would have assigned it a higher probability than was assigned by outcome-ignorant subjects. This study is listed among the debiasing attempts, since by concentrating on a few stories it answered the methodological criticism of "asking too many questions" that might be leveled against subsequent studies. Other studies that asked few questions without eliminating hindsight bias include Slovic and Fischhoff (1977), who had subjects analyze the likelihood of possible outcomes of several scientific experiments; Mitchell and Kalb (in press), who had nurses evaluate incidents taken from hospital settings; and Pennington, Rutter, McKenna, and Morley (1980), who had women assess their personal probability of receiving a positive result on a single pregnancy test (although the low power of this study renders its conclusion somewhat tentative).

Other attempts to demonstrate an artifactual source of hindsight bias that have been tried and failed include: substituting rating-scale judgments of "surprisingness" for probability assessments (Slovic & Fischhoff,

1977); using more homogeneous items to allow fuller evocation of one set of knowledge, rather than using general-knowledge questions scattered over a variety of content areas, none of which might be thought about very deeply (Fischhoff & Beyth, 1975); imploring subjects to work harder (Fischhoff, 1977b); trying to dispel doubts about the nature of the experiment (G. Wood, 1978); and using contemporary events that judges have considered in foresight prior to making their hindsight assessments (Fischhoff & Beyth, 1975).

Misunderstood tasks. One possible attraction of hindsight bias is that it may be quite flattering to represent oneself as having known all along what was going to happen. One pays a price for such undeserved self-flattery only if (a) one's foresight leads to an action that appears foolish in hindsight or (b) systematic exaggeration of what one *knew* leads to overconfidence in what one presently *knows*, possibly causing capricious actions or failure to seek needed information. Since these long-range consequences are not very relevant in the typical experiment, one might worry about subjects being tempted to paint themselves in a favorable light. Although most experiments have been posed as tests of subjects' ability to reconstruct a foresightful state of knowledge, rather than as tests of how extensive that knowledge was, temptations to exaggerate might still remain. If so, they would reflect a discrepancy between subjects' and experimenters' interpretations of the task. One manipulation designed to eliminate this possibility requires subjects first to answer questions and then to remember their own answers, with the acuity of their memory being at issue (Fischhoff, 1977b; Fischhoff & Beyth, 1975; Pennington et al., 1980; G. Wood, 1978). A second manipulation requires hindsight subjects to estimate the foresight responses of their peers, on the assumption that they have no reason to exaggerate what others knew (Fischhoff, 1975; G. Wood, 1978). Neither manipulation has proven successful. Subjects remembered themselves to have been more knowledgeable than was, in fact, the case. They were uncharitable second-guessers in the sense of exaggerating how much others would have (or should have) known in foresight.

Faulty judges

Learning to avoid the biases that arise from being a prisoner of one's present perspective constitutes a, or perhaps the, focus of historians' training (see Chap. **23**). There have, however, been no empirical studies of the success of these efforts. The emphasis that historians place on primary sources, with their fossilized records of the perceptions of the past, may reflect a feeling that the human mind is sufficiently incorrigible to require that sort of discipline by document. Although it used a vastly less rigorous procedure, the one experimental training study offers no reason for

optimism: Fischhoff (1977b) explicitly described the bias to subjects and asked them to avoid it in their judgments – to no avail.

Mismatch between judges and tasks

Restructuring. Three strategies have been adopted to restructure hindsight tasks, so as to make them more compatible with the cognitive skills and predispositions that judges bring to them. One such strategy separates subjects in time from the report of the event, in hopes of reducing its tendency to dominate their perceptual field (Fischhoff & Beyth, 1975; G. Wood, 1978); this strategy was not effective. With the second strategy, judges assess the likelihood of the reported event's recurring rather than the likelihood of its happening in the first place, in the hope that uncertainty would be more available in the forward-looking perspective (Mitchell & Kalb, in press; Slovic & Fischhoff, 1977); this, too, failed. The final strategy requires subjects to indicate how they could have explained the occurrence of the outcome that did *not* happen (Slovic & Fischhoff, 1977). Recruiting such negative evidence appreciably reduced the judged inevitability of the reported event. Such contradictory evidence was apparently available to subjects in memory or imagination but not accessible without a restructuring of the problem.

Education. There is little experimental evidence that hindsight bias is reduced by the sort of intense involvement with a topic that comes with a professional education. Detmer, Fryback, and Gassner (1978) found hindsight bias in the judgments of surgeons (both faculty and residents) appraising an episode involving a possible leaking abdominal aortic aneurism. Arkes, Wortmann, Saville, and Harkness (1981) demonstrated the bias with physicians considering clinical descriptions of a bartender with acute knee pain. Mitchell and Kalb (in press) found bias in nurses' appraisal of the outcome of acts performed by subordinates. If people judging events in their own lives are considered to be substantive experts, then the study by Pennington et al. (1980) of women judging the results of personal pregnancy tests might be considered a further example of bias in experts. In an even more limited sense of expertise, G. Wood (1978) found that with a task involving general-knowledge questions his most knowledgeable subjects were no less bias prone than less knowledgeable ones. The anecdotal evidence of experts falling prey to this bias is described briefly in Chapter 23 (this volume). It includes both casual observations and exhaustive studies, such as that of Wohlstetter (1962), who characterized the efforts of the highly motivated experts comprising the congressional investigatory committee following Pearl Harbor as 39 volumes of hindsight bias.

Summary

Although one of the lesser-studied judgmental problems, hindsight bias has produced enough research to allow some tentative general statements: It appears to be quite robust and widespread. Reducing it requires some understanding of and hypotheses about people's cognitive processes. One such hypothesis is that the manner in which people normally approach hindsight tasks does not use their knowledge or inferential skills to best advantage. Producing contrary evidence appeared to remedy that problem in part and to help them make better use of their own minds (Slovic & Fischhoff, 1977).

Before endorsing this solution, however, a number of empirical issues need to be addressed: (a) What additional steps are needed for the bias to be eliminated, not only reduced? (b) Will this procedure work with less clearly structured tasks? (c) Will practice in the procedure with a few exemplary tasks suffice to change behavior with other tasks, where no specific instruction is given? A debiasing procedure may be more trouble than it is worth if it increases people's faith in their judgmental abilities more than it improves the abilities themselves.

Overconfidence: Debiasing efforts

"Decision making under uncertainty" implies incomplete knowledge. As a result, one major component of making such decisions is appraising the quality of whatever knowledge is available. Although statistical methods may guide this appraisal, at some point or other judgment is needed to assess the confidence that can be placed in one's best guess at the state of the world. Because improper confidence assessment can lead to poor decisions, by inducing either undue or insufficient caution, a continuing focus of judgment research has been the identification of factors affecting confidence inappropriately. Receipt of outcome knowledge is one such factor, insofar as it leads people to exaggerate the completeness of their own knowledge. Although one suspects that outcome knowledge leaves people overconfident in their own knowledge, it is conceivable that people are subject to some sort of endemic underconfidence to which hindsight bias provides a useful counterbalance. Clarifying this possibility requires research evaluating the absolute validity of confidence judgments.

Because it is difficult to assess the absolute validity of any single confidence judgment, most research in this area has looked at the quality, or *calibration*, of sets of judgments, each representing the subjective probability that a statement of fact is correct (Chap. 22, this volume). For the perfectly calibrated individual, assessments of, say, .70 are associated with correct statements 70% of the time.

Overconfidence is by far the most commonly observed finding. A typical study might show probabilities of .75 to be associated with a "hit rate" of only 60% and expressions of certainty ($p = 1.00$) being correct only 85% of the time. When people assess how much they know about the values of numerical quantities (e.g., "I am .98 certain that the number of registered Republican voters in Lane County is between 12,000 and 30,000"), it is not uncommon to find true answers falling outside of their 98% confidence intervals 20% to 40% of the time. Such results are disturbing both to those who must rely on confidence assessments and to those accused (directly or indirectly) of exaggerating how much they know. The abundant research that has been produced to disprove, discredit, bolster, or bound the finding of overconfidence is characterized below from the perspective of debiasing efforts. This reanalysis of existing studies has been aided greatly by the availability of several comprehensive reviews of this literature, albeit conducted for somewhat different purposes. These include Henrion (1980), Hogarth (1975), Lichtenstein, Fischhoff, and Phillips (Chap. 22), and Wallsten and Budescu (1980). This reanalysis has been complicated by the fact that many of the studies cited also were conducted for somewhat different purposes. As a result, they do not always fall neatly into a single debiasing category. This mild mismatch may reflect limits on the present categorical scheme (for making unclear distinctions) or limits to the studies (for confounding debiasing manipulations).

Faulty tasks

Unfair tasks. The applied implications of overconfidence have spawned a large number of technical efforts at its eradication, almost all of which have proven unsuccessful. Many of these have involved response-mode manipulations, such as comparing probability and odds expressions of confidence (Ludke, Stauss, & Gustafson, 1977) or varying the confidence intervals assessed in creating subjective probability distributions (Selvidge, 1980). Freed of the necessity of generating and justifying their manipulations on the basis of some substantive theory, experimenters using such "engineering" approaches often show great ingenuity in the procedures they are willing to try. However, the absence of theory also makes it more difficult to know how to interpret or generalize their successes or failures. For example, Seaver, von Winterfeldt, and Edwards (1978) found less overconfidence when confidence intervals were elicited with a "fixed-value" method, in which the experimenter selected values and subjects assessed their likelihood, than with the "fixed-probability" method, in which the experimenter provides a probability and the respondent gives the associated value. This success may reflect some sort of greater compatibility between the fixed-value method and respondents'

psychological processes, or it may reflect the information about the true value conveyed by the experimenter's choice of fixed values. A similar result by Tversky and Kahneman (1974, **1**) is grounded on a hypothesized anchoring-and-adjustment heuristic, although it too may have informed fixed-value subjects.

In addition to the rather intense search for the right response mode for eliciting confidence, there have also been scattered attempts to eliminate the other threats to task fairness listed in the top section of Table 1. For example, the large number of responses elicited in many calibration studies so as to obtain statistically reliable individual results might be a matter of concern had not overconfidence been observed in studies with as few as 10 or even 1 question per subject (e.g., Hynes & Vanmarcke, 1976; Lichtenstein & Fischhoff, 1977). The brevity of the instructions used in some studies might be troublesome had not similar results been found with instructions that seem to be as long and detailed as subjects would tolerate (e.g., Chap. **21**; Lichtenstein & Fischhoff, 1980b). The exhaustiveness, even pedantry, of such instructions might also be seen as an antidote to any temptation for subjects to second-guess the investigator. Regarding the clarity of the stimuli used, no change in overconfidence has been observed when diverse sets of general-knowledge questions are replaced with homogeneous items (e.g., Fischhoff & Slovic, 1980; Oskamp, 1962) or with non-verbal "perceptual" items (e.g., Dawes, 1980; Lichtenstein & Fischhoff, 1980b).

It would be reassuring to believe that overconfidence disappears when the stakes are raised and judges perform "for real" (i.e., not just for experiments). Unfortunately, however, the research strategies that might be used to study this hypothesis tend to encounter interpretive difficulties. Monitoring the confidence expressions of experts performing their customary tasks is one obvious approach. It is frustrated by the possibility that the experts' expressions are being evaluated on criteria that conflict with calibration; that is, there may be rewards for deliberately exuding undue confidence or for sounding overly cautious. For example, when physicians overestimate the likelihood of a malady (e.g., Christensen-Szalanski & Bushyhead, 1981; Lusted, 1977), it may be because they are out of touch with how much they know or because of malpractice worries, greed for the financial rewards that additional testing may bring, or other concerns irrelevant to the present purposes. Because of these complications, studies with experts are listed in the section devoted to them at the bottom of Table 2, rather than as attempts to raise the stakes.

A second strategy for raising the stakes is to append confidence assessments to inherently important tasks for which those assessments have no action implications. Sieber (1974) did so by soliciting students' confidence in their own test answers. The result was (the now-familiar) overconfidence, perhaps because calibration is insensitive to the stakes involved, perhaps because this method was not effective in raising them. The

Table 2. *Debiasing experience*

Strategies	Studies examining hindsight bias	Studies examining overconfidence
Faulty tasks		
Unfair tasks		
Raise stakes	4	1,30
Clarify instructions/stimuli	6	3,10,13,14,21
Discourage second guessing	11	13,21
Use better response modes	9	13,14,20,22,23,32,34,35?, **36**,40?
Ask fewer questions	3,7,8,9	16
Misunderstood tasks		
Demonstrate alternative goal	3,4,6,8,9	14
Demonstrate semantic disagreement	—	3,14,19,30?
Demonstrate impossibility of task	—	13
Demonstrate overlooked distinction	—	15?
Faulty judges		
Perfectible individuals		
Warn of problem	—	13
Describe problem	4	**3**
Provide personalized feedback	—	**21**
Train extensively	5?	**1,2,4,17,21,26,27,31,34**
Incorrigible individuals		
Replace them	—	—
Recalibrate their responses	—	2,5,24
Plan on error	—	—
Mismatch between judges and task		
Restructuring		
Make knowledge explicit	—	18
Search for discrepant information	**9**	**18**
Decompose problem	6,11	—
Consider alternative situations	—	—
Offer alternative formulations	7,9	35?
Education		
Rely on substantive experts	1,2,7,8,10,11	**11**,16,20,24,29,33,38,39/ **8,9,23,28,31,32**[a]
Educate from childhood	—	6,7

Notes: Key to studies follows notes. Manipulations that have proven at least partially successful appear in boldface. Those that have yet to be subjected to empirical test or for which the evidence is unclear are marked by a question mark.
[a]Entries before the slash are studies using experts who have not had calibration training; entries after the slash are studies using variable difficulty levels.

Key to studies

Hindsight

1. Arkes, Wortmann, Saville, & Harkness (1981)
2. Detmer, Fryback, & Gassner (1978)
3. Fischhoff (1975)
4. Fischhoff (1977b)
5. Fischhoff (1980)
6. Fischhoff & Beyth (1975)

Table 2. *(cont.)*

7. Mitchell & Kalb (in press)
8. Pennington, Rutter, McKenna, & Morley (1980)
9. Slovic & Fischhoff (1977)
10. Wohlstetter (1962)
11. G. Wood (1978)

Overconfidence

1. Adams & Adams (1958)
2. Adams & Adams (1961)
3. Alpert & Raiffa (1969, **21**)
4. Armelius (1979)
5. Becker & Greenberg (1978)
6. Beyth-Marom & Dekel (in press)
7. Cavanaugh & Borkowski (1980)
8. Clarke (1960)
9. Cocozza & Steadman (1978)
10. Dawes (1980)
11. Dowie (1976)
12. Ferrell & McGoey (1980)
13. Fischhoff & Slovic (1980)
14. Fischhoff, Slovic, & Lichtenstein (1977)
15. Howell & Burnett (1978)
16. Hynes & Vanmarcke (1976)
17. King, Zechmeister, & Shaughnessy (in press)
18. Koriat, Lichtenstein, & Fischhoff (1980)
19. Larson & Reenan (1979)
20. Lichtenstein & Fischhoff (1977)
21. Lichtenstein & Fischhoff (1980b)
22. Lichtenstein, Fischhoff, & Phillips (Chap. 22)
23. Ludke, Stauss, & Gustafson (1977)
24. Moore (1977)
25. Morris (1974)
26. Murphy & Winkler (1974)
27. Murphy & Winkler (1977a)
28. Nickerson & McGoldrick (1965)
29. Oskamp (1962)
30. Phillips & Wright (1977)
31. Pickhardt & Wallace (1974)
32. Pitz (1974)
33. Root (1962)
34. Schaefer & Borcherding (1973)
35. Seaver, von Winterfeldt, & Edwards (1978)
36. Selvidge (1980)
37. Sieber (1974)
38. Staël'von Holstein (1971a)
39. Staël von Holstein (1972)
40. Tversky & Kahneman (1974)

theoretically perfect strategy for manipulating stakes is to reward subjects with proper scoring rules, which penalize unfrank expressions of uncertainty. Such rules are, however, quite asymmetric, in the sense that they penalize overconfidence much more than underconfidence. As a result, subjects who understand the gist of those rules but who are uninterested in their particulars, might interpret scoring rules as roundabout instructions never to express great confidence. In that case, people might just mechanically reduce their confidence without improving understanding. All in all, perhaps the best way to get subjects to work hard is by exercising the experimentalists' standard techniques for increasing a task's intrinsic motivation and subjects' involvement in it.

Misunderstood tasks. However carefully one describes a task to respondents, some doubts may linger as to whether they really understood it and accepted its intended reward structure. A standard maneuver for checking whether a manipulation has "worked" is to see if participants will stand by the responses that they already have made when those responses are used in a new task with the reward structure intended for the old task.

Fischhoff, Slovic, and Lichtenstein (1977) adopted this strategy in asking people if they would be willing to accept a gamble based on confidence assessments they had just made. This gamble favored them if those assessments were frank or tended to underrate their confidence, but penalized them if, for whatever reason, they had exaggerated how much they knew. Deliberate exaggeration might, for example, serve the alternative goal of acting more knowledgeable than is actually the case. These subjects were quite eager to accept the gamble, despite being as overconfident as subjects observed elsewhere.

Another basis for claiming that subjects have understood the task differently from the way intended by the experimenter comes from the observation that "degrees of certainty are often used in everyday speech (as are references to temperature), but they are seldom expressed numerically, nor is the opportunity to validate them often available. . . . People's inability to assess appropriately a probability of .80 may be no more surprising than the difficulty they might have in estimating brightness in candles or temperature in degrees Fahrenheit" (Fischhoff et al., 1977, p. 553). One response to this possibility is restricting attention to the extremes of the probability scale in the belief that "being 100% certain that a statement is true is readily understood by most people and its appropriateness is readily evaluated" (Fischhoff et al., 1977, p. 553). A second response is providing verbal labels for numerical probabilities in order to make them more readily comprehensible (e.g., Chap. **21**; Larson & Reenan, 1979). Neither manipulation has proven demonstrably effective. A deeper notion of semantic disagreement between experimenter and respondent may be found in claims that "uncertainty" itself may have a variety of interpretations, not all of which are meaningful to all individuals (Howell & Burnett, 1978; Phillips & Wright, 1977). Empirical debiasing efforts based on these concepts might prove fruitful.

Some of the most extreme overconfidence has been observed with tasks regarding which respondents have no knowledge whatsoever. Although experimenters typically attempt to give no hints as to how confident subjects should be, there still might be an implicit presumption that "the experimenter wouldn't give me a task that's impossible." If subjects had such expectations, having an appropriate level of confidence would then become impossible. Fischhoff and Slovic (1980) tested this possibility with a series of tasks whose content (e.g., diagnosing ulcers, forecasting the prices of obscure stocks) and instructions were designed to make them seem as impossible as they actually were. However, overconfidence was only reduced (and then but partially) when subjects were cautioned that "it may well be impossible to make this sort of discrimination. Try to do the best you can. But if, in the extreme you feel totally uncertain about [your answers], do not hesitate to respond with .5 [indicating a guess] for every one of them" (p. 752). Any stronger instructions might be suspected of having demand characteristics of their own.

Faulty judges

Perfectible individuals. With a modest change in interpretive assumptions, the last-mentioned study in the previous section might become the first-mentioned member of the present one. Assuring subjects that they could admit that every response was just a guess might be seen as a way to dispel any residual misunderstandings about the task or as a step toward correcting subjects who understand the task but not themselves. It carries an implicit warning that failure to admit to guessing may be a problem. This warning is made explicit in Alpert and Raiffa's (Chap. **21**) instruction to subjects to "spread the tails" of their subjective probability distributions in order to avoid overconfidence. Whether the partial success of these manipulations reflects increased understanding or sensitivity to orders is unclear. Such ambiguity may explain the paucity of studies adopting these approaches.

These worries about demand characteristics disappear with deliberate training studies, where "experimenter effects" are the order of the day. As indicated by Table 2, a variety of training efforts have been undertaken with an admirable success rate – although one might worry that journals' lack of enthusiasm for negative results studies may have reduced the visibility of failures. Trainers' willingness to do whatever it takes to get an effect has tended to make training efforts rather complex manipulations whose effective elements are somewhat obscure. Some of the more necessary conditions for learning seem to be: receiving feedback on large samples of responses, being told about one's own performance (and not just about common problems), and having the opportunity to discuss the relationship between one's subjective feelings of uncertainty and the numerical probability responses. To their own surprise, Lichtenstein and Fischhoff (1980b) found that one round of training with intensive, personalized feedback was as effective as a long series of trials. It is unclear to what extent these various successes represent training, in the narrow sense of mastering a particular task (e.g., learning the distribution of responses the experimenter requires), or the acquisition of more general skills.

Incorrigible individuals. Impatience with training studies or skepticism about their generality has led a number of investigators to take fallible confidence assessments as inevitable and concentrate on helping decision makers to cope with them. Some suggest replacing individuals with groups of experts whose assessments are combined by direct interaction or a mechanical aggregation scheme (e.g., Becker & Greenberg, 1978; Morris, 1974); others call for liberal use of sensitivity analysis whenever confidence assessments arise in a decision analysis (e.g., Jennergren & Keeney, in press); still others propose to recalibrate assessments, using a correction factor that indicates how confident assessors should be as a function of

how confident they are (Lichtenstein & Fischhoff, 1977). For example, the prevalence of overconfidence might suggest that when someone proclaims certainty, one might read it as a .85 chance of their being correct. Unfortunately for this strategy, when people are miscalibrated their degree of overconfidence depends upon the difficulty of the particular task facing them (Lichtenstein & Fischhoff, 1977). As a result, the needed amount of recalibration can be determined only if one knows the difficulty of the task at hand and can observe respondents' (over)confidence in a task of similar difficulty or at least surmise the relationship between observed and anticipated overconfidence (Ferrell & McGoey, 1980).

Mismatch between judges and task

Restructuring. The study of calibration, like some other topics in judgment, has remained relatively isolated from the mainstream of research in cognition, drawing more methodology than ideas from the psychological literature. Whether this lack of contact reflects the insularity of judgment researchers or the inadequate representations of confidence in current models of cognitive processes, it has likely hindered the development of methods to reduce overconfidence. Process models should both suggest more powerful manipulations and indicate why engineering approaches do or do not work (and how far their effects might generalize). Current research in eyewitness testimony, feeling of knowing, and metamemory might eventually provide points of contact (e.g., Gruneberg, Morris, & Sykes, 1978).

One possible direction for helping people use their existing cognitive skills in a way more compatible with the demands of confidence assessment may be seen in Koriat, Lichtenstein, and Fischhoff (1980), where overconfidence was reduced by having respondents list reasons why their preferred answer might be wrong. Listing reasons why one might be right or giving one reason for and one reason against one's chosen answer had no effect, indicating that the critical element is not just working harder or being explicit, but addressing one's memory differently from what is customary in confidence assessment tasks. Without the specific prompting to "consider why you might be wrong," people seem to be insufficiently critical or even intent on justifying their initial answer. Perhaps analogously, Markman (1979) found that 9- and 12-year-olds detected inconsistencies in textual material only when told to look for them.

Although it is advanced on practical rather than psychological grounds, Seaver et al.'s (1978) fixed-value technique might be seen as another way of restructuring respondents' approach to the task. Organizing one's knowledge around a set of values presumed to be incorrect may lead to a more complete appraisal of what one knows than the "traditional" fixed-

probability method, in which attention may be focused on the respondents' best guess at the correct answer.

Education. Does overconfidence disappear as an indirect result of the substantive education that experts receive in their specialty? As mentioned earlier, the obvious way to explore this question, looking at the confidence expressions accompanying the performance of real tasks, is complicated by the possibility that real pressures restrict experts' candor. For example, one might find evidence of overconfidence in professions that make confident judgments with no demonstrated validity (e.g., predictions of stock price movements [Dreman, 1979; Slovic, 1972c], psychiatric diagnoses of dangerousness [Cocozza & Steadman, 1978]). Of course, if such "experts" are consulted (and paid) as a function of the confidence they inspire, they may be tempted to misrepresent how much they know.

Undoubtedly, the greatest efforts to ensure candor have been with weather forecasters, whose training often explicitly rewards them for good calibration. Their performance is superb (e.g., Murphy & Winkler, 1974, 1977a). Whether this success is due to calibration training or a by-product of their general professional education is unclear. A review of other studies with experts who have not had calibration training suggests that such training, and not just substantive education, is the effective element. Experiments that used problems drawn from their respective areas of expertise but isolated from real-world pressures have found overconfidence with psychology graduate students (Lichtenstein & Fischhoff, 1977), bankers (Staël von Holstein, 1972), clinical psychologists (Oskamp, 1962), executives (Moore, 1977), civil engineers (Hynes & Vanmarcke, 1976), and untrained professional weather forecasters (Root, 1962; Staël von Holstein, 1971a).

Dowie (1976) has found good calibration among the newspaper predictions of horse-racing columnists. Although these experts receive neither an explicit payoff function nor formal feedback, one might guess that they supply their own, monitoring their performance from day to day and rewarding themselves for good calibration. The idea that we should be trained from childhood for this kind of self-monitoring may be found in recent proposals to make judgment a part of the school curriculum (e.g., Beyth-Marom & Dekel, in press; Cavanaugh & Borkowski, 1980). The promise of these proposals remains to be tested.

Finally, there is a rather narrow form of expertise that has proven to be the most potent (and least interesting) method of reducing overconfidence. One reflection of people's insensitivity to how much they know is the fact that their mean confidence changes relatively slowly in response to changes in the difficulty of the tasks they face (Lichtenstein & Fischhoff, 1977). Typical pairs of proportions of correct answers and mean

confidence are: .51, .65; .62, .74; .80, .78; and .92, .86. As accuracy ranges over .41, confidence changes only .23. The calibration curves corresponding to these summary statistics are in some senses about equally bad (or flat); however, their degree of overconfidence varies considerably. Whereas the first two of these pairs represent overconfidence, the third shows appropriate overall confidence and the fourth underconfidence. These examples are taken from Lichtenstein and Fischhoff (1977), but the same pattern has been revealed by Clarke (1960), Nickerson and McGoldrick (1965), Pickhardt and Wallace (1974), and Pitz (1974), among others. Indeed, any comparison of overconfidence across conditions must take into account the difficulty of the tasks used. In this light, the preponderance of overconfidence in the literature reflects, in part, the (perhaps natural) tendency not to present people with very easy questions.

Summary

Confidence assessments have been extracted from a variety of people in a variety of ways, almost always showing considerable insensitivity to the extent of their knowledge. Although the door need not be closed on methodological manipulations, they have so far proven relatively ineffective and their results difficult to generalize. What they have done is to show that overconfidence is relatively resistant to many forms of tinkering (other than changes in difficulty level). Greater reliance on psychological theory would seem to be the key to producing more powerful and predictable manipulations. The effectiveness of calibration training suggests that a careful analysis of what unique experiences are provided by that training but not by professional education could both guide debiasing and enrich psychological theory.

Discussion

Assuming that the studies reviewed here have been characterized accurately and that they exhaust (or at least fairly represent) the universe of relevant studies, their aggregate message would seem to be fairly reassuring to the cognitive psychologist. Both biases have proven moderately robust, resisting attempts to interpret them as artifacts and eliminate them by "mechanical" manipulations, such as making subjects work harder. Effective debiasing usually has involved changing the psychological nature of the task (and subjects' approach to it). In such cases, at least some of the credit must go to psychological theory. For example, a hypothesis about how people retrieve memory information prior to assessing confidence guided Koriat et al.'s (1980) manipulation of that retrieval process.

Even "throw everything at the subject" training programs have been based on well-tested and generally-applicable principles of learning.

Several conceptual caveats should accompany this summary (in addition to the methodological ones with which it opened). One is that the distinction between artifactual and psychological manipulations may be less clear than has been suggested here. For example, exhorting people to work harder would be an artifact manipulation when rooted in a claim that more casual instructions do not elicit "real behavior." However, if the investigator could advance substantive hypotheses about how different instructions affect judgmental processes, the artifact would become a main effect with separate predictions for real-world behavior in situations with and without explicit exhortations.

The second conceptual caveat is that questioning the reality of biases can reflect a limited and unproductive perspective on psychological research. To continue the example of the preceding paragraph, life has both casual and work-hard situations; neither one is inherently more "real" than the other. By like token, the relative validity of casual and work-hard laboratory experiments depends upon the real-world situations to which their results are to be extrapolated. Each has its place. Understanding the laboratory-world match requires good judgment in characterizing both contexts. For example, work-hard situations are not necessarily synonymous with important situations. People may not work hard on an important problem unless they realize both the centrality of a judgment to the problem's outcome and the potential fallibility of that judgment.

Using debiasing studies to discover the boundary conditions for observing biases leads to the third conceptual caveat. In this review, the summary tables and discussion implicitly afforded equal weight to the various studies, qualified perhaps by some notion of each study's definitiveness (as determined by competence, extensiveness, etc.). Such tallying of statistically significant and non-significant results is a dubious procedure on methodological grounds alone (e.g., Hedges & Olkin, 1980). It becomes conceptually questionable when one doubts that the universe of possible studies is being sampled adequately. In such cases, those data that are collected constitute conceptually dependent observations and need not be given equal weight. Any summary of how people behave needs a careful specification of the subuniverse of behavioral situations from which studies are being sampled. For example, some critics have charged that early studies of judgmental heuristics were "looking for trouble," in the sense of searching (grasping) for situations in which people would behave in an errant fashion. If this claim is true, then each demonstration of biased behavior need not be interpreted as a strike against people's overall judgmental ability; its relevance is limited to the kind of situations being studied (or overstudied) in those experiments. By focusing on the boundary conditions for assessing biases, more recent studies are subject to

Table 3. *A universe of discourse for biases and debiasing efforts*

1. *The underlying processes about which inferences are required are probabilistic.* That is, judgments are made under conditions of uncertainty, with biases arising from the confrontation between a deterministic mind and a probabilistic environment.
2. *Problems arise in the integration rather than discovery of evidence.* Although stimuli are complete and unambiguous as possible, they tell little about how the task might be structured. The subjects' task is interpreting and using those pieces of information that are provided
3. *The biases are non-substantive.* The operation of a cognitive process should be similar in any content area with a given informational structure. This eliminates "errors" due to misinformation and "misconceptions" due to deliberate deception.
4. *Some normative theory is available characterizing appropriate judgment.* This criterion rules out problems from the realm of preference (e.g., inconsistent attitudes), where no one response can be identified as optimal.
5. *No computational aids are offered or allowed (beyond pencil and paper).* This focus on intuitive judgment excludes such aids as dedicated hand calculators, statistical consultants, and interactive computers.
6. *No obvious inducements for suboptimal behavior are apparent.* That is, biases are cognitive, not motivational in nature. The "point" of bias research is, of course, that where people have no good reason to act suboptimally, errors suggest that they just do not know any better.

their own sampling bias, which needs to be considered in generalizing their results.

Further questions

Whether similar patterns will emerge with other biases requires analogous literature reviews. Table 3 offers a characterization of the domain of biases within which recurrent patterns might be sought, distinguishing the contents of this volume from other biases that have troubled psychologists.

A lingering metaquestion facing those reviews is, How good are people? Are they cognitive cripples or cognoscenti? Providing a single answer requires an answer to imponderable questions about the nature of life and the overall similarity of human experience to laboratory conditions. An elusive summary from the present review is that people's reservoir of judgmental skills is both half empty and half full. People are skilled enough to get through life, unskilled enough to make predictable and consequential mistakes; they are clever enough to devise broadly and easily applicable heuristics that often serve them in good stead, unsophisticated enough not to realize the limits to those heuristics. A more specific appraisal of people's ability can be given only in the context of a particular judgment task.

Such blanket statements (or evasions) about "people" reflect a common feature of most judgmental research – lack of interest in individual differences. Although this preference for group effects may be just a matter of taste, it might be justified theoretically by arguing that the main effects in judgment studies are so large and inadequately explored that individual differences can wait. The rather meager insight provided by studying groups with known characteristics provides some empirical support for this claim. Particularly striking was the lack of differences in experimental studies of the most consequential of known groups, experts making judgments in their fields of expertise. The anecdotal and case-study evidence collected by Dawes (1976), Eddy (Chap. **18**, this volume), Fischer (1970), and others also indicates that extensive training and high stakes are no guarantees of judgmental prowess. Nonetheless, further research is needed, both because of the firmness with which many believe that experts are better and the applied importance of using expert judgment to best advantage.

For the immediate practical goal of best deploying experts so as to avoid bias, it is sufficient to know whether they are better than lay people or at least better aware of their own judgmental limitations. For the eventual practical goal of debiasing all judges, it is important to know how the experts got where they did or why they got no further. The following is a list of conditions that are generally conducive to learning. For each, one can see ways in which experts might be at a particular advantage or disadvantage, depending upon the circumstances:

1. Abundant practice with a set of reasonably homogeneous tasks. Experts should have such experience. They may use it to hone their judgmental skills or they may develop situation-specific habitual solutions, freeing themselves from the need to analyze (and think).
2. Clear-cut criterion events. Although experts are often required to make their judgments quite explicit, the objects of those judgments are often components of such complex (natural, social, or biological) systems that it is hard to evaluate the judges' level of understanding. Off-target judgments may be due to unanticipated contingencies, whereas on-target judgments may have been right for the wrong reason.
3. Task-specific reinforcement. Experts are, in principle, paid for performance. However, even when the wisdom of their judgments can be discerned, they may be rewarded on other grounds (e.g., did they bring good news? did they disrupt plans? did things turn out for the best?).
4. Explicit admission of the need for learning. Entering an apprenticeship program that confers expertise is surely a sign of modesty.

> Nonetheless, at every stage of that process and the professional life that follows it, certain advantages accrue to those who put on a good show and exude competence.

These are purely operant principles of learning, manipulating behavior without presuming any knowledge of underlying cognitive processes. Clarifying and exploiting those cognitive processes is obviously a major theoretical and practical task for debiasing research, especially when one considers that such manipulations seem to have a somewhat better track record than more mechanical efforts. Although the study of biases and debiasing has spanned a fair portion of the long path from basic research to field applications, it has yet to touch bases adequately at either end. It appears now that reaching one end will require reaching the other as well. Good practice will require better theory about how the mind works. Good theory will require better practice, clarifying and grappling with the conditions in which the mind actually works.