Journal of Experimental Psychology: General
1998, Vol. 127, No. 2, 161–188

# Individual Differences in Rational Thought

Keith E. Stanovich
University of Toronto

Richard F. West
James Madison University

Much research in the last 2 decades has demonstrated that humans deviate from normative models of decision making and rational judgment. In 4 studies involving 954 participants, the authors explored the extent to which measures of cognitive ability and thinking dispositions can predict discrepancies from normative responding on a variety of tasks from the heuristics and biases literature including the selection task, belief bias in syllogistic reasoning, argument evaluation, base-rate use, covariation detection, hypothesis testing, outcome bias, if–only thinking, knowledge calibration, hindsight bias, and the false consensus paradigm. Significant relationships involving cognitive ability were interpreted as indicating algorithmic-level limitations on the computation of the normative response. Relationships with thinking dispositions were interpreted as indicating that styles of epistemic regulation can predict individual differences in performance on these tasks.

Although over two decades' worth of empirical studies have indicated that human responses often deviate from the response considered normative on many reasoning tasks, the reason for this discrepancy between the normative and descriptive is still the subject of intense controversy (Baron, 1994; Cosmides & Tooby, 1996; Gigerenzer, 1993, 1996; Kahneman & Tversky, 1996; Koehler, 1996; Piattelli-Palmarini, 1994; Stein, 1996). One aspect of performance that has been largely neglected by all parties in these disputes has been individual differences (see Galotti, Baron, & Sabini, 1986; Jepson, Krantz, & Nisbett, 1983; Roberts, 1993; and Yates, Lee, & Shinotsuka, 1996, for exceptions to the general neglect of individual differences in the literature). What has largely been ignored is that—although the average person in these experiments might well display an overconfidence effect, underuse base rates, choose *P* and *Q* in the selection task, commit the conjunction fallacy, etc.—on each of these tasks, some people give the standard normative response. For example, in knowledge calibration studies, although the mean performance level of the entire sample may be represented by a calibration curve that indicates overconfidence (Fischhoff, 1988), often, a few people do display near-perfect calibration. As another example, consider the problems that the Nisbett group (Fong, Krantz, & Nisbett, 1986; Lehman, Lempert, & Nisbett, 1988) have used to assess statistical thinking in everyday situations. Although the majority of people often ignore the more diagnostic but pallid statistical evidence, some actually do rely on the statistical evidence rather than the vivid case evidence. A few people even respond correctly on the notoriously difficult abstract selection task (Evans, Newstead, & Byrne, 1993; Newstead & Evans, 1995).

In this article, we argue that the nature of these individual differences and their patterns of covariance might have implications for explanations of why human behavior often departs from normative models. For example, theorists who argue that discrepancies between actual responses and those dictated by normative models are not indicative of human irrationality sometimes (e.g., L. J. Cohen, 1981) attribute the discrepancies to performance errors (see Stein, 1996, pp. 8–9). Borrowing the idea of a competence–performance distinction from linguists, these theorists view performance errors as the failure to apply a rule, strategy, or algorithm that is part of a person's competence because of a momentary and fairly random lapse in ancillary processes necessary to execute the strategy (lack of attention, temporary memory deactivation, distraction, etc.; see Stein, 1996). Stein (1996) explains the idea of a performance error by referring to a "mere mistake"—a more colloquial notion that involves

> a momentary lapse, a divergence from some typical behavior. This is in contrast to attributing a divergence from norm to reasoning in accordance with principles that diverge from the normative principles of reasoning. Behavior due to irrationality connotes a systematic divergence from the norm. It is this distinction between mere mistakes and systematic violations (between performance errors and competence errors) that is . . . implicitly assumed by friends of the rationality thesis when they deny that the reasoning experiments [demonstrate human irrationality]. (p. 8)

This notion of a performance error as a momentary attention, memory, or processing lapse that causes responses

Keith E. Stanovich, Department of Human Development and Applied Psychology, Ontario Institute for Studies in Education, University of Toronto, Toronto, Ontario, Canada; Richard F. West, Department of Psychology, James Madison University.

This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada and a James Madison University Program Faculty Assistance Grant.

We thank Penny Chiappe, Alexandra Gottardo, Walter Sa, Robin Sidhu, and Ron Stringer for their assistance in data coding. Jonathan Baron and Jonathan St. B. T. Evans are thanked for their comments on a draft of this article. Anne E. Cunningham is thanked for help in obtaining the experts' evaluation for the Argument Evaluation Test.

Correspondence concerning this article should be addressed to Keith E. Stanovich, Department of Human Development and Applied Psychology, Ontario Institute for Studies in Education, University of Toronto, 252 Bloor Street West, Toronto, Ontario, Canada M5S 1V6. Electronic mail may be sent to kstanovich@oise.utoronto.ca.

to appear nonnormative even when the model of competence is fully normative has implications for patterns of individual differences across reasoning tasks. For example, the strongest possible form of this view is that all discrepancies from normative responses are due to performance errors. This strong form has the implication that there should be virtually no correlations among performance on disparate reasoning tasks. If each departure from normative responding represents a momentary processing lapse due to distraction, carelessness, or temporary confusion, there is no reason to expect covariance in performance across various indexes of rational thinking. In contrast, positive manifold among disparate rational thinking tasks would call into question the notion that all variability in responding can be attributable to performance errors.

## IMPLICATIONS OF INDIVIDUAL DIFFERENCES FOR PRESCRIPTIVE MODELS: ALGORITHMIC-LEVEL LIMITATIONS

Patterns of individual differences might have implications that extend beyond testing the view that discrepancies between descriptive models and normative models arise entirely from performance errors. For example, patterns of individual differences have implications for prescriptive models of rationality. The contrast between descriptive, normative, and prescriptive models of human reasoning is cogently discussed in Baron (1985) and Bell, Raiffa, and Tversky (1988). Briefly, descriptive models are the goal of most work in empirical psychology. In contrast to descriptive models, which are concerned with observed behavior, normative models embody standards of cognitive activity—standards that, if met, serve to optimize the accuracy of beliefs and the efficacy of actions. Demonstrating that descriptive accounts of human behavior often diverged from these normative models was a main theme in the early literature on heuristics and biases (Arkes & Hammond, 1986; Evans, 1989; Kahneman, Slovic, & Tversky, 1982).

As interesting as such divergences are, there is a major difficulty in labeling them as instances of irrationality. Judgments about the rationality of actions and beliefs must take into account the resource-limited nature of the human cognitive apparatus (Cherniak, 1986; Goldman, 1978; Harman, 1995; Oaksford & Chater, 1993, 1995; Stich, 1990). Thus, prescriptive models are usually viewed as specifying how reasoning should proceed given the limitations of the human cognitive apparatus and the situational constraints (e.g., time pressure) with which the decision maker must deal (Baron, 1985; Bell et al., 1988). In cases in which the normative model is computable, it is also prescriptive (at least in situations with no time pressure). In a case in which the normative model is not computable by the human brain, then the standard for human performance becomes the computable strategy closest to the normative model—the strategy that maximizes the individual's goal satisfaction given the individual's cognitive limitations and environmental context (Baron, 1985).

The idea of computational limitations on optimal processing (which drives a wedge between the normative and prescriptive) is best discussed by first making some well-known distinctions between levels of analysis in cognitive theory (Anderson, 1990, 1991; Levelt, 1995; Marr, 1982; Newell, 1982, 1990; Oaksford & Chater, 1995; Pylyshyn, 1984; Sterelny, 1990). We follow here the taxonomy of Anderson (1990), who has drawn most heavily on the work of Marr (1982) and Newell (1982). Anderson (1990) has defined four levels of theorizing in cognitive science: a biological level that is inaccessible to cognitive theorizing, an implementation level designed to approximate the biological, an algorithmic level (an abstract specification of the computational processes necessary to carry out a task), and the rational level. The last level provides a specification of the goals of the system's computations (what the system is attempting to compute and why) and can be used to suggest constraints on the operation of the algorithmic level. According to Anderson (1990), the rational level specifies what are the "constraints on the behavior of the system in order for that behavior to be optimal" (p. 22). The description of this level of analysis proceeds from a "general principle of rationality" that assumes that: "the cognitive system operates at all times to optimize the adaptation of the behavior of the organism" (Anderson, 1990, p. 28). Thus, the rational level of analysis is concerned with the goals of the system, beliefs relevant to those goals, and the choice of action that is rational given the system's goals and beliefs (Anderson, 1990; Bratman, Israel, & Pollack, 1991; Dennett, 1987; Newell, 1982, 1990).

However, even if all humans were optimally adapted to their environments at the rational level of analysis, there may still be computational limitations at the algorithmic level (e.g., Cherniak, 1986; Goldman, 1978; Oaksford & Chater, 1993, 1995; Overton, 1985, 1990). Individual differences in actual performance would therefore still be expected (despite no rational-level differences) because of differences at the algorithmic level. Thus, the magnitude of the correlation between performance on a reasoning task and cognitive capacity provides an empirical clue about the importance of algorithmic limitations in creating discrepancies between descriptive and normative models. A strong correlation suggests important algorithmic limitations that might make the normative response not prescriptive for those of lower cognitive capacity. In contrast, the absence of a correlation between the normative response and cognitive capacity suggests no computational limitation and thus no reason why the normative response should not be considered prescriptive.

Additionally, the direction of the correlation might also help to indicate whether the correct normative model is being used to evaluate performance. For example, theorists in the heuristics and biases literature who defend the standard normative models are sometimes criticized for explaining divergences between normative models and actual performance by claiming that limitations in computational capacity prevent the normative response. But critics who claim that the wrong normative model is being invoked have argued that there is "no support for the view that people would choose in accord with normative prescriptions if they were provided with increased capacity" (Lopes & Oden, 1991, p. 209).

One way to indirectly test this claim is to investigate how responses on reasoning tasks correlate with measures of cognitive capacity. If that correlation is positive, it would seem to justify the use of the normative model being employed to evaluate performance, whereas negative correlations might be indicating that an inappropriate normative model is being applied to the situation. This would seem to follow from the arguments of the optimization theorists who emphasize the adaptiveness of human cognition (Anderson, 1990, 1991; D. T. Campbell, 1987; Cooper, 1987; Cosmides & Tooby, 1994, 1996; Oaksford & Chater, 1993, 1994, 1995; Payne, Bettman, & Johnson, 1993; Schoemaker, 1991; Shanks, 1995). Assuming that the rational model for all humans in a given environment is optimized in an evolutionary sense, as previously mentioned, individual differences in actual performance would still be expected (despite no rational-level differences) because of differences at the algorithmic level. The responses of organisms with fewer algorithmic limitations would be assumed to be closer to the response that a rational analysis would reveal as optimal. For example, the optimal strategy might be computationally more complex, and only those with the requisite computational power might be able to compute it. Under standard assumptions about the adaptive allocation of cognitive resources (Anderson, 1991; Payne et al., 1993; Schoemaker, 1991), the additional computational complexity would only be worth dealing with if the strategy were indeed more efficacious. Alternatively, the optimal strategy might not be more computationally complex. It might simply be more efficient and more readily recognized as such by more intelligent organisms. Thus, negative correlations with the response considered normative might call into question the appropriateness of the normative model being applied. This follows because it is assumed that most adaptation theorists would wish to avoid the conclusion that humans with more computational power systematically choose the nonnormative response.

Finally there is the possibility that cognitive ability might not correlate with performance on a particular reasoning task. Thus, the discrepancy between descriptive and normative models of behavior in such a situation cannot be attributed to algorithmic-level limitations. If the gap also cannot be attributed to performance errors, then there are two other important possibilities. One is that humans are systematically computing a nonnormative rule (Baron, 1991b; Kahneman & Tversky, 1996; Shafir, 1994). The other is that some individuals are adopting an alternative task construal that warrants another normative model. The latter possibility follows from the by now widely recognized point that the evaluation of the normative appropriateness of a response to a particular task is always relative to a particular interpretation of the task (Adler, 1984; Berkeley & Humphreys, 1982; Broome, 1990; Gigerenzer, 1996; Schick, 1987, 1997; Tversky, 1975).

## IMPLICATIONS OF INDIVIDUAL DIFFERENCES FOR PRESCRIPTIVE MODELS: THINKING DISPOSITIONS AS MODERATORS OF COMPETENCE

An important tradition within developmental psychology views various organismic factors as moderators of basic cognitive competence (Neimark, 1981, 1985; Overton, 1985, 1990; Overton & Newman, 1982). In this study, we have examined whether we can identify such factors using as a framework the distinction between cognitive capacities and thinking dispositions. Baron (1985, 1988) has provided one of the most extensive discussions of the distinction. In his conceptualization, capacities refer to the types of cognitive processes studied by information-processing researchers seeking the underlying cognitive basis of performance on IQ tests. Perceptual speed, discrimination accuracy, working memory capacity, and the efficiency of the retrieval of information stored in long-term memory are examples of cognitive capacities that underlie traditional psychometric intelligence (Carpenter, Just, & Shell, 1990; Deary & Stough, 1996; Dougherty & Haith, 1997; Estes, 1982; Fry & Hale, 1996; Hunt, 1978, 1987; Vernon, 1991, 1993).

According to Baron's (1985) conception, cognitive capacities cannot be improved in the short-term by admonition or instruction. They are, nevertheless, affected by long-term practice. Thinking dispositions, in contrast, are better viewed as cognitive styles that are more malleable: "Although you cannot improve working memory by instruction, you can tell someone to spend more time on problems before she gives up, and if she is so inclined, she can do what you say" (Baron, 1985, p. 15). Rational thinking dispositions are those that relate to the adequacy of belief formation and decision making, things like

> the disposition to weigh new evidence against a favored belief heavily (or lightly), the disposition to spend a great deal of time (or very little) on a problem before giving up, or the disposition to weigh heavily the opinions of others in forming one's own. (Baron, 1985, p. 15)

Because thinking dispositions and cognitive capacity are thought to differ in their degree of malleability, it is important to determine the relative proportion of variance in rational thinking skills that can be explained by each. To the extent that thinking dispositions explain variance in a rational thinking skill independent of cognitive capacity, theorists such as Baron (1985, 1988, 1993b) would predict that the skill would be more teachable.

We further suggest here that individual differences in cognitive ability and thinking dispositions refer to variation in components of a cognitive system at different levels of analysis. Variation in cognitive ability refers to individual differences in the efficiency of processing at the algorithmic level. In contrast, thinking dispositions of the type studied in this investigation elucidate individual differences at the rational level. They index the individual's goals and epistemic values (Anderson, 1990, 1991; Kruglanski & Webster, 1996).

Thus, the magnitude of the correlation between performance on a reasoning task and thinking dispositions provides *an empirical clue about the importance of rational* level thinking styles and goals in hindering or facilitating normative responding. A strong correlation would suggest that the normative response might not currently be prescriptive for those possessing certain rational-level thinking styles. But here is where Baron's (1985) point about differences between cognitive capacities and thinking dispositions in relative malleability becomes especially important. If a cognitive style was the factor limiting normative responding, the prescriptive model for such an individual might be much more malleable (and potentially convergent with the normative model) than in the case of an individual with an algorithmic limitation.

The empirical search for individual difference factors that account for discrepancies between descriptive models of actual behavior and normative models can also be viewed within the competence-activation–utilization approach of developmental theorists (Overton, 1990; Overton & Newman, 1982). Overton and Newman (1982) argue that two distinct components are required for a complete psychological theory. One is a competence component that is an idealized model of the abstract knowledge possessed by an individual in a given domain. The activation–utilization component encompasses the psychological procedures and situational factors that determine the manifestation of the competence. Activation–utilization factors serve to explain at least some of the discrepancies between descriptive and normative models. These factors incorporate the potential algorithmic limitations discussed previously. However, Overton and Newman (1982; see also Neimark, 1981, 1985; Overton, 1990) are clear that cognitive styles are also activation–utilization factors that both hinder and facilitate the manifestation of competence (see Overton, Byrnes, & O'Brien, 1985, for an empirical example).

## THE ANALYTIC STRATEGY

In the following series of studies, we have explored the implications of individual differences in rational thought for explanations of the gap between descriptive and normative models of reasoning and decision making. We have examined whether individual differences in a variety of deductive and inductive reasoning tasks display patterns of association among themselves and with measures of cognitive ability and rational thinking dispositions.

Associations among the reasoning tasks themselves have been interpreted as falsifying the strong view that all discrepancies between actual responses and normative responses can be attributed to nonsystematic performance errors. Correlations among disparate tasks would, in contrast, at least suggest the operation of systematic factors such as limitations at the algorithmic or rational levels that prevent normative responding. Correlations with measures of cognitive capacity would suggest the presence of computational limitations that render the normative response not prescriptive for those of lower cognitive capacity. Analogously, the magnitude of the correlation between performance on a reasoning task and thinking dispositions provides *an empirical clue about the importance of rational* level thinking styles and goals in hindering or facilitating normative responding. The direction of the correlation with these factors may provide clues as to which response model is being optimized if it is accepted as a corollary of optimization theory that extra cognitive capacity is always used to execute more efficacious strategies.

## EXPERIMENT 1: TASK SELECTION

We present in this study the most comprehensive analysis to date of individual differences on tasks from the reasoning and from the heuristics and biases literature. Deductive tasks from the reasoning literature and inductive reasoning tasks from the heuristics and biases literature have traditionally been investigated within separate research programs (Evans, 1989; but see Evans, Over, & Manktelow, 1993; Johnson-Laird & Shafir, 1993; Shafir, 1994). Tasks from both literatures were examined in the experiments reported here. For the remainder of this discussion, tasks from both of these literatures are collectively termed *rational thinking,* or *reasoning,* tasks. In Experiment 1 we employed three rational thinking tasks from the research literature and one that we devised.

The first of the three tasks that we chose from the existing research literature was a syllogistic reasoning task. In order to highlight the logical aspect of the task and to emphasize the type of decontextualization believed to be critical for rational thought, we employed conclusions that contradicted world knowledge when the syllogism was valid and that were consistent with world knowledge when the syllogism was invalid (see Evans, Barston, & Pollard, 1983; Evans, Over, et al., 1993; Markovits & Bouffard-Bouchard, 1992).

The research literatures on deductive reasoning and on statistical reasoning have developed largely in separation, although some integrative efforts have recently been made (see Evans, Over, et al., 1993; Legrenzi, Girotto, & Johnson-Laird, 1993). Because studies of reasoning with contrary-to-fact syllogisms derive from the former, our next task was drawn from the statistical reasoning literature. The type of statistical reasoning problem we examined was one inspired by the work of Nisbett and Ross (1980) on the tendency for human judgment to be overly influenced by vivid but unrepresentative personal and case evidence and to be underinfluenced by more representative and diagnostic, but pallid, statistical evidence.

The third task that we chose from the literature was Wason's (1966) selection task (sometimes termed the *four-card task*). Variants of the task have been the subject of intense investigation (see Beattie & Baron, 1988; Evans, Newstead, et al., 1993; Liberman & Klar, 1996; Newstead & Evans, 1995). Unlike the case of the contrary-to-fact syllogisms task and statistical reasoning task, whereby a substantial number of people give the normative response, the gap between the descriptive and normative for the abstract selection task is unusually large. Less than 10% of the participants give the normative response on the abstract version (Evans, Newstead, et al., 1993). This has led some to question whether the *P* and not-*Q* choice should in fact be considered the normative response (Fetzer, 1990; Finocchiaro, 1980; Lowe, 1993; Nickerson, 1996; Oaksford & Chater, 1994; Wetherick, 1995).

The fourth task in Experiment 1 was one that we devised and that was constructed in an attempt to capture an aspect of rational thought emphasized by many theorists: the ability to evaluate the quality of an argument independent of one's feelings and personal biases about the proposition at issue. With this task—the argument evaluation test (AET)—we employed an analytic technique for developing separate indexes of a person's reliance on the quality of an argument and on their own personal beliefs about the issue in dispute (Stanovich & West, 1997).

Our methodology involved assessing, on a separate instrument, the participant's degree of agreement with a series of propositions. On an argument evaluation measure, administered at a later time, the person evaluated arguments related to the same propositions. The arguments had an operationally determined quality. Our analytic strategy was to regress the participant's evaluation of the argument simultaneously on the objective measure of argument quality and on the individual's prior opinion. The standardized beta weight for argument quality then becomes an index of the participant's reliance on the quality of arguments independent of their opinions on the issues in question. The task assesses argument evaluation skills of the type studied in the informal reasoning literature (Baron, 1991a, 1995; Klaczynski & Gordon, 1996; Kuhn, 1991, 1993; Perkins, Farady, & Bushey, 1991).

Cognitive capacity was measured in these experiments by administering well-known cognitive ability and academic aptitude tasks. All are known to load highly on psychometric g (Carpenter et al., 1990; Carroll, 1993; Matarazzo, 1972) and such measures have been linked to neurophysiological and information-processing indicators of efficient cognitive computation (Caryl, 1994; Deary, 1995; Deary & Stough, 1996; Detterman, 1994; Fry & Hale, 1996; Hunt, 1987; Stankov & Dunn, 1993; Vernon, 1991, 1993; Vernon & Mori, 1992).

With regard to thinking dispositions, we focused on those most relevant to epistemic rationality—processes leading to more accurate belief formation and to more consistent belief networks (Audi, 1993a, 1993b; Foley, 1987, 1988, 1991; Goldman, 1986; Harman, 1995; Kitcher, 1993; Nozick, 1993; Stanovich, 1994, in press; Stanovich & West, 1997; Thagard, 1992). We attempted to tap the following dimensions: epistemological absolutism, willingness to switch perspectives, willingness to decontextualize, and the tendency to consider alternative opinions and evidence. Baron (1985, 1988, 1993b) has viewed such cognitive styles as tapping a dimension he calls actively open-minded thinking (see also, Kruglanski & Webster, 1996; Schommer, 1993, 1994).

# Method

## *Participants*

The participants were 197 undergraduate students (56 men and 141 women) recruited through an introductory psychology participant pool at a medium-sized state university. Their mean age was 18.9 years (*SD* = 1.3).

## *Rational Thinking Tasks*

### *Syllogistic Reasoning Task*

Students evaluated eight syllogisms whereby logic was in conflict with believability. The eight problems were taken from the work of Markovits and Nantel (1989). Four of the problems had conclusions that followed logically but were unbelievable (e.g., All mammals walk. Whales are mammals. Conclusion: Whales walk), and four problems had conclusions that did not follow logically but were believable. Participants were given instructions strongly emphasizing that they should evaluate logical validity and not the believability of the conclusion. The score on this task was the raw number correct out of eight. The mean score was 4.4 (*SD* = 2.2). Only 18 of the 195 students completing this task answered all eight items correctly.

### *Selection Task*

Since its introduction by Wason (1966), the selection task has been investigated in a variety of different forms (see Evans, Newstead, et al., 1993, for a review). We employed five items, all of which were composed of real-life but somewhat arbitrary content. The five rules were: Papers with an 'A' on one side have the comment 'excellent' on the other side; If 'Baltimore' is on one side of the ticket, then 'plane' is on the other side of the ticket; If it's a 'USA' postcard, then a '20¢' stamp is on the other side of the postcard; Whenever the menu has 'fish' on one side, 'wine' is on the other side; Every coin with 'Madison' on one side has 'library' on the other side.

All five problems were accompanied by a graphic choice displaying four alternatives that represented the choices *P*, not-*P*, *Q*, and not-*Q* for that particular problem. Students were asked to decide which alternatives they would need to turn over in order to find out whether the rule was true or false.

Overall performance on the selection task was extremely low, as in other studies using arbitrary or abstract content (Beattie & Baron, 1988; Evans, Newstead, et al., 1993; Klaczynski & Laipple, 1993). Only 23 of the 191 students gave a correct response (*P* and not-*Q*) on at least one of the five problems. Only 3 of 191 individuals answered all five items correctly. The pattern of incorrect responses displayed by our participants replicated that found in other studies of performance on the selection task (see Evans, Newstead, et al., 1993; Jackson & Griggs, 1988; Newstead & Evans, 1995). In a meta-analysis of selection task studies, Oaksford and Chater (1994) found that the probabilities of choosing the *P*, *Q*, not-*Q*, and not-*P* cards were .89, .62, .25, and .16, respectively. The probabilities in our experiment (.88, .66, .22, and .15) were highly convergent. Because the mean number of correct responses across the five trials was so low (0.27) and so highly skewed, we sought a more continuous index of performance. Our choice was to sum—across all five trials—the number of correct responses (*P*, not *Q*) and the number of incorrect choices (not *P*, *Q*). We then formed a score for each student by subtracting the number of incorrect responses from the number of correct responses (this index is the summed version of the logic index employed by Pollard & Evans, 1987; other scoring methods produced highly convergent results). Scores on this index ranged from a high of 10 to a low of −5, and the mean score was 1.47 (*SD* = 2.7).

## *Statistical Reasoning: Inductive Preferences Task*

These problems were like those used in the work of Fong et al. (1986) and Jepson et al. (1983) but were adapted into a multiple-choice format. The problems were structured so that the participant had to make an inductive inference in a simulation of a real-life decision. The information relevant to the decision was conflicting and of two different types. One type of evidence was statistical: either probabilistic or aggregate base-rate information that favored one of the bipolar decisions. The other evidence was a concrete case or personal experience that pointed in the opposite direction. The classic Volvo-versus-Saab item (see p. 285 of Fong et al., 1986) provides an example and was included in our battery. In this problem, a couple are deciding to buy one of two otherwise equal cars. Consumer surveys, statistics on repair records, and polls of experts favor the Volvo over the Saab. However, a friend reports experiencing a severe mechanical problem with the Volvo he owns. Preference for the Volvo indicates a tendency to rely on the large sample of positive information in spite of salient negative personal testimony. Five additional problems of this type were employed: the college choice, admissions, and class choice problems adapted from Jepson et al. (1983), and the curriculum choice, and marriage and baseball performance problems adapted from Fong et al. (1986). The problems were all scored in the direction giving higher scores to the choice of the aggregate information and lower scores to the single case, or individuating, evidence. These six problems all involved causal aggregate information, analogous to the causal base rates discussed by Ajzen (1977) and Bar-Hillel (1980, 1990), that is, base rates that had a causal relationship to the criterion behavior (under the alternative conceptualization of Bar-Hillel, 1980, the problems had relevant base rates). A seventh problem that was examined was the batting average problem adapted from Lehman et al. (1988). Performance on each of these seven inductive preference items was standardized, and the seven z scores were summed to form a composite score for statistical reasoning.

## *Argument Evaluation Test*

The argument evaluation test (AET) consisted of 23 items. The instructions introduced the participants to a fictitious individual, Dale, whose arguments they were to evaluate. Each item began with Dale stating an opinion about a social issue (e.g., "The welfare system should be drastically cut back in size."). The individual is then asked to indicate the extent to which they agree with the opinion on a 4-point scale: *strongly disagree* (1), *disagree* (2), *agree* (3), *strongly agree* (4). Dale then gives a justification for the opinion (in this case, for example, "The welfare system should be drastically reduced in size because welfare recipients take advantage of the system and buy expensive foods with their food stamps."). A critic then presents an argument to counter this justification (e.g., "Ninety-five percent of welfare recipients use their food stamps to obtain the bare essentials for their families."). The participant is told to assume that the counterargument is factually correct. Finally, Dale attempts to rebut the counterargument (e.g., "Many people who are on welfare are lazy and don't want to work for a living."). Again assuming that the argument is factually correct, the student is told to evaluate the strength of Dale's rebuttal to the critic's argument. The instructions remind the individual that he or she is to focus on the quality of Dale's rebuttal and to ignore the issue of whether or not he or she agreed with Dale's original opinion. The participant then evaluates the strength of the rebuttal on a 4-point scale: *very weak* (1), *weak* (2), *strong* (3), *very strong* (4). The remaining 22 items are structured analogously. With one exception, they all concerned real social and political issues (e.g., gun control, taxes, university governance, automobile speed limits). Several examples of items on the AET are provided in Stanovich and West (1997). AET items varied greatly in the degree to which participants endorsed the original opinion—from a low of 1.88 for the item "It is more dangerous to travel by air than by car" to a high of 3.79 for the item "Seat belts should always be worn when traveling in a car." Likewise, the mean evaluation of the rebuttal ranged from 1.96 to 3.57.

The analysis of performance on the AET required that the individuals' judgments of the strength of the rebuttals be compared to some objective standard. We employed a summary measure of eight expert judges of the arguments as the normative index. Specifically, three full-time faculty members of the Department of Philosophy at the University of Washington; three full-time faculty members of the Department of Philosophy at the University of California, Berkeley; and the two authors judged the strength of the rebuttals. The median correlation between the judgments of the eight experts was .74. Although the problems were devised by the two authors, the median correlations between their judgments and those of the external experts were reasonably high (.78 and .73, respectively) and roughly equal to the median correlation among the judgments of the six external experts themselves (.73). Thus, for the purposes of the regression analyses described later, the median of the eight experts' judgments of each of the 23-item rebuttals served as the objective standard and was employed as the objective index of argument quality for each item. As an example, on the previous item, the median of the experts' ratings of the rebuttal was 1.50 (between weak and very weak). The mean rating given the item by the participants was 2.02 (weak).

One indication of the validity of the experts' ratings is that the experts were vastly more consistent among themselves in their evaluation of the items than were the participants. Because the median correlation among the eight experts' judgments was .74, a parallel analysis of consistency among the participants was conducted in order to provide a comparison. Twenty-four groups of 8 participants were formed and the correlations among the 8 individuals in each group calculated. The median correlation for each of the 24 groups was then determined. The highest median across all of the 24 groups was .42, substantially below the experts' value of .74. The mean of the median correlation in the 24 groups was .28, markedly below the degree of consistency in the experts' judgments.

Individual differences in participants' relative reliance on objective argument quality and prior belief were examined by running separate regression analyses on each student's responses. That is, a separate multiple regression equation was constructed for each student. The student's evaluations of argument quality served as the criterion variable in each of 194 separate regression analyses. Each individual's 23 evaluation scores were regressed simultaneously on both the 23 argument quality scores (the experts' ratings) and the 23 prior belief scores (the participant's original agreement with the target proposition). For each individual, these analyses resulted in two standardized beta weights: one for argument quality and one for prior belief. The former beta weight—an indication of the degree of reliance on argument quality independent of prior belief—is used as the primary indicator of the ability to evaluate arguments independent of beliefs.[1]

The mean multiple correlation across the 194 separate regressions in which the participant's evaluations was regressed on objective argument quality scores and his or her prior belief scores was .509 ($SD$ = .153). Across the 194 regressions, the mean β weight for argument quality was .340 ($SD$ = .202). These latter values varied widely—from a low of −.214 to a high of .835. Only 7 of 194 participants had negative β weights. Across the 194 regressions, the mean β weight for prior belief was .243 ($SD$ = 235). These values also varied widely—from a low of −.384 to a high of .809. Only 33 of 194 participants had negative β weights.

## *Cognitive Ability Measures*

### *Scholastic Aptitude Test Scores*

Because Scholastic Aptitude Test (SAT) scores were not available to us because of university restrictions, students were asked to indicate their verbal and mathematical SAT scores on a demographics sheet. The mean reported verbal SAT score of the 184 students who filled in this part of the questionnaire was 521 ($SD$ = 70), the mean reported mathematical SAT score was 572 ($SD$ = 84), and mean total SAT score was 1093 ($SD$ = 118). These reported scores are reasonably close to the averages for this institution, which are 520, 587, and 1107, respectively (Straughn & Straughn, 1995; all SAT scores were from administrations of the test prior to its recent rescaling).

Participants indicated their degree of confidence in their memory of their scores on a 5-point scale (*high, moderately high, somewhat high, low, very low*). Of the sample, 76.6% indicated that their degree of confidence was high or moderately high and only 1.6% of the sample indicated that their confidence was very low. Finally, 163 of the 184 students granted permission for the experimenters to look up their SAT scores. The correlations to be reported were virtually unchanged when students less than moderately confident of their scores or students who did not give permission to look up their scores were excluded. This was true for all subsequent experiments as well. Thus, reported SAT scores from the entire sample are employed in the analyses. Considerable empirical evidence on the validity of the reported SAT scores is presented in Experiment 2 and in Stanovich, West, and Harrison (1995).

### *The Raven Matrices*

Participants completed 18 problems from the Raven Advanced Progressive Matrices (Set II, Raven, 1962, hereinafter referred to as the Raven matrices), a task tapping general problem-solving skills and commonly viewed as a good measure of analytic intelligence (Carpenter et al., 1990). The students were given 15 min to complete the 18 items on the test. By eliminating 12 of the easiest problems, in which performance in a college sample is near ceiling, and 6 of the most difficult problems in which performance is nearly floored (Carpenter et al., 1990; Raven, Court, & Raven, 1977), we tried to achieve a cut-time version of the test that would still have adequate reliability. A previous investigation using a 16-item version of the Raven Standard Progressive Matrices achieved reliabilities over .75 in samples of children (Cahan & Cohen, 1989). The split-half reliability of our 18-item measure (.69, Spearman-Brown corrected) was similar. The mean score on the test was 9.6 ($SD$ = 3.0).

### *Reading Comprehension*

Participants completed the Comprehension subscale of the Nelson-Denny Reading Test (Form F; Brown, Bennett, & Hanna, 1981; hereinafter referred to as the Nelson-Denny Comprehension measure). In order to cut the administration time from 20 min to 14 min, the long initial passage of Form F and one other passage were omitted, along with their 12 questions. Students thus completed six of the eight passages and answered the 24 questions associated with those six passages. The split-half reliability of this shortened version of the test (.70, Spearman-Brown corrected) was only slightly lower than the alternate-form reliability of .77 reported in the test manual (Brown et al., 1981). The mean score on the comprehension subtest was 19.8 ($SD$ = 2.8).

### *Cognitive Ability Composite*

In several of the analyses reported later a composite cognitive ability score that combined performance on the ability measures was employed. To form this index, scores on the Raven matrices (1962), scores on the Nelson-Denny Comprehension measure (Brown et al., 1981), and the SAT Total scores were standardized and summed.

## *Thinking Dispositions Questionnaire*

Participants completed a questionnaire consisting of intermixed items from a number of subscales. The response format for each item in the questionnaire was *strongly agree* (4), *slightly agree* (3), *slightly disagree* (2), and *strongly disagree* (1). The subscales were as follows:

### *Actively Open-Minded Thinking Scale*

Items on the Actively Open-Minded Thinking (AOT) subscale were devised by the authors. The design of the items was influenced by a variety of sources from the critical thinking literature (e.g., Ennis, 1987; Facione, 1992; Nickerson, 1987; Norris & Ennis, 1989; Perkins, Jay, & Tishman, 1993; Zechmeister & Johnson, 1992) but most specifically by the work of Baron (1985, 1988, 1993b), who has emphasized the concept of actively open-minded thinking through the cultivation of reflectiveness rather than impulsivity, the seeking and processing of information that disconfirms one's belief (as opposed to confirmation bias in evidence seeking), and the willingness to change one's beliefs in the face of contradictory evidence. There were 10 items on the AOT scale that tapped the disposition toward reflectivity (e.g., "If I think longer about a problem I will be more likely to solve it."), willingness to consider evidence contradictory to beliefs (e.g., "People should always take into consideration evidence that goes against their beliefs."), and tolerance for ambiguity combined with a willingness to postpone closure ("There is nothing wrong with being undecided about many issues," "Changing your mind is a sign of weakness,"—the latter reverse scored).

### *Counterfactual Thinking Scale*

A two-item subscale designed to tap counterfactual thinking was devised by the authors. The two scale items were: "My beliefs

[1] Virtually identical results were obtained by using as a variable the beta weight for argument quality minus the beta weight for item agreement. Therefore, the simpler index—the beta weight for argument quality—was employed in the subsequent analyses.

would not have been very different if I had been raised by a different set of parents" and "Even if my environment (family, neighborhood, schools) had been different, I probably would have the same religious views." Both items were reversed scored so that higher scores indicate counterfactual thinking.

### *Absolutism Scale*

This subscale was adapted from the Scale of Intellectual Development (SID) developed by Erwin (1981, 1983). The SID represents an attempt to develop a multiple-choice scale to measure Perry's (1970) hypothesized stages of intellectual development in young adulthood (see also Kramer, Kahlbaugh, & Goldston, 1992; Schommer, 1993). We chose nine items designed to tap Perry's early stages, which are characterized by an absolutist orientation. This orientation is captured by items such as "It is better to simply believe in a religion than to be confused by doubts about it" and "Right and wrong never change."

### *Dogmatism*

The Dogmatism subscale consisted of three items taken from a short-form field version (Troldahl & Powell, 1965) of Rokeach's (1960) Dogmatism scale (e.g., "Of all the different philosophies which exist in the world there is probably only one which is correct.").

### *Paranormal Beliefs*

The Paranormal Beliefs subscale was composed of six items. Two items were concerned with belief in astrology ("It is advisable to consult your horoscope daily"; "Astrology can be useful in making personality judgments") and were adapted from the Paranormal Belief scale validated by Jones, Russell, and Nickel (1977). The four remaining items concerned the belief in the concept of luck (e.g., "The number 13 is unlucky.") and were similar to items on the Superstition subscale of a paranormal beliefs questionnaire developed by Tobacyk and Milford (1983).

### *Social Desirability Response Bias*

Five items reflecting social desirability response bias (Furnham, 1986; Paulhus & Reid, 1991) were taken from Erwin's (1981, 1983) SID (e.g., "I always put forth my best effort," "I never mislead people."). These items are similar to other five-item social desirability instruments in the literature (see Hays, Hayashi, & Stewart, 1989). None of the relationships to be discussed was mediated by social desirability effects, so this measure is not discussed further.

### *Thinking Dispositions Composite*

A thinking dispositions composite score (TDC) was formed by summing the scores on the AOT and Counterfactual Thinking scales and then subtracting the sum of the scores on the Absolutism, Dogmatism, and Paranormal scales (forming the composite from standardized scores of the five variables resulted in virtually identical correlational results). Thus high scores on the TDC indicate open-mindedness, cognitive flexibility, and a skeptical attitude; whereas low scores indicate cognitive rigidity and lack of skepticism.

## *Procedure*

Participants completed the tasks during a single 2- to 2.5-hr session. They were tested in small groups of 3–8 individuals.

## Results

### *Relationships Among Rational Thinking Tasks*

The correlations among the four rational thinking tasks are displayed in Table 1. Five of the six correlations were significant at the .001 level. The syllogistic reasoning task displayed significant correlations with each of the other three tasks, as did the selection task. The only correlation that did not attain significance was that between performance on the AET (β weight for argument quality) and statistical reasoning. Although the highest correlation obtained was that between syllogistic reasoning and selection task performance (.363), correlations almost as strong were obtained between tasks deriving from the deductive reasoning literature (syllogistic reasoning, selection task) and inductive reasoning literature (statistical reasoning).

### *Relationships With Cognitive Ability and Thinking Dispositions*

As indicated in Table 1, the cognitive ability composite variable was significantly correlated with performance on all four rational thinking tasks. The correlation with syllogistic reasoning was the highest (.496), and the other three correlations were roughly equal in magnitude (.298 to .334).

The last line of Table 1 indicates, as with the cognitive ability composite, that the thinking dispositions composite score was correlated with each of the four rational thinking tasks. Although in each case the correlation was smaller than

Table 1
*Intercorrelations Among the Primary Variables in Experiment 1*

| Variable | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. Syllogisms | — | | | | |
| 2. Selection task | .363**** | — | | | |
| 3. Statistical reasoning | .334**** | .258**** | — | | |
| 4. AET | .340**** | .310**** | .117 | — | |
| 5. Cognitive ability comp. | .496**** | .328**** | .298**** | .334**** | — |
| 6. TDC | .277**** | .228*** | .201*** | .226*** | .337**** |

*Note.* AET = beta weight for argument quality in the argument evaluation test; Cognitive ability comp. = cognitive ability composite score; TDC = thinking dispositions composite score.
***$p < .01$, two-tailed. ****$p < .001$, two-tailed.

that involving the cognitive ability composite, each of the four correlations with the thinking dispositions composite was statistically significant. This finding suggests that there may be differences in cognitive style at the rational level that are at least in part accounting for discrepancies between descriptive and normative models.

Of course, cognitive ability and thinking dispositions explain partially overlapping variance in the rational thinking tasks (the correlation between the two composites was .337). We thus examined the extent to which joint limitations at the algorithmic and rational levels can explain variation from normative responding by conducting a multiple regression analysis using a composite measure of performance on the rational thinking tasks. The scores on the four reasoning tasks—syllogisms, selection task, statistical reasoning, AET—were standardized and summed to yield a composite score. The cognitive ability composite and TDC attained a multiple $R$ with this criterion variable of .559, $F(2, 194) = 44.12$, $p < .001$). Thus, a substantial amount of variance (31.3%) on the rational thinking tasks is jointly explained by these two predictors. The cognitive ability composite was a significant unique predictor, partial correlation = .473, unique variance explained = .198, $F(1, 194) = 55.82$, $p < .001$, as was the TDC, partial correlation = .200, unique variance explained = .029, $F(1, 194) = 8.09$, $p < .01$.

## Discussion

Each of the rational thinking tasks investigated in Experiment 1 displayed individual differences that tended to be reliably correlated with the individual differences displayed on other reasoning tasks. The significant relationships between most of the rational thinking tasks suggests that departures from normative responding on each of them are due to systematic limitations in processing and not to nonsystematic performance errors. The positive manifold displayed by the tasks suggests that random performance errors cannot be the sole reason that responses deviate from normative responding because there were systematic response tendencies across very different reasoning tasks.

Cognitive capacities, indexed by the cognitive ability composite, were significantly correlated with performance on all four rational thinking tasks. This finding indicates that, for each of the tasks, discrepancies between participants' responses and the response considered normative can to some extent be explained by algorithmic-level computational limitations. Interestingly, this was true both in the case of a deductive reasoning task in which there is little dispute concerning what is the normative response (syllogisms) and in the case of the inductive reasoning tasks in which models of normative responding are less secure.

Performance on all four tasks was likewise significantly related to thinking dispositions. Jointly, cognitive ability and thinking dispositions accounted for a moderate amount of variance in overall rational thinking performance, although cognitive ability was a stronger unique predictor. It appears that to a considerable extent, discrepancies between actual performance and normative models can be accounted for by variation in capacity limitations at the algorithmic level and cognitive style differences at the rational level. However, the moderate joint associations with cognitive capacity and thinking dispositions also leaves ample room for some proportion of the discrepancies to be due to differential task construal (Gigerenzer, 1996; Schick, 1987, 1997) or to tendencies to systematically compute according to a nonnormative rule.

# EXPERIMENT 2

In Experiment 2 we attempted a replication of some of the basic results of Experiment 1 with a larger sample size. Additionally, we examined performance on several other tasks commonly used to assess rational thinking ability (a covariation judgment task and a hypothesis testing task) and heuristic processing tendencies (if–only thinking, outcome bias).

## Method

### *Participants*

The participants were 546 undergraduate students (143 men and 403 women) recruited through two introductory psychology participant pools. Their mean age was 19.0 years ($SD$ = 2.3). We followed the procedure described in Experiment 1 for ascertaining SAT scores and received estimated SAT scores from 529 participants. Of the 529 students, 484 gave permission to have their scores verified, 76.7% of the sample indicated that their degree of confidence was high or moderately high, and only 1.7% of the sample indicated that their confidence was very low. The mean reported verbal SAT score of the sample was 527 ($SD$ = 70), and the mean reported mathematical SAT score was 582 ($SD$ = 80). The mean reported total SAT score for the 529 students was 1108 ($SD$ = 118). As an additional objective test of the accuracy of the SAT estimates, a brief vocabulary measure was administered to the participants (because vocabulary is the strongest specific correlate of general intelligence, see Matarazzo, 1972). This vocabulary recognition test (described in Stanovich et al., 1995) displayed a correlation of .47 with the SAT total score. This .47 correlation is quite similar to the .51 correlation between the vocabulary checklist and verified total SAT scores in a previous investigation (West & Stanovich, 1991). A further indication of the validity of the SAT estimates is that the vocabulary checklist displayed a higher correlation with the verbal SAT estimates (.53) than with the quantitative estimates (.22), and the difference in dependent correlations (see J. Cohen & Cohen, 1983, pp. 56–57) was highly significant ($p < .001$).

In Experiment 2, the demographics form filled out by the students included questions on their educational history in mathematics and statistics courses. We constructed a 0–4-point scale that assessed the student's mathematics–statistics course background. Students received 1 point if they had taken a statistics course in college (131 students), 1 point if they had taken a statistics course in high school (64 students), 1 point if they had taken a mathematics course in college (469 students), and 1 point if they had had 4 years of high school mathematics (467 students). The mean score on the scale was 2.07 ($SD$ = .75).

### *Tasks*

Three tasks were administered and scored in a manner identical to that employed in Experiment 1: the argument evaluation test, the syllogisms task, and the seven statistical reasoning problems.

The new tasks and revised tasks were the following:

## *Methodological Thinking Tasks*

Two tasks were chosen that heavily emphasized methodological thinking (see Lehman et al., 1988). The tasks were as follows:

*Covariation judgment.* Judging event interrelationship is a critical component of much thinking in the everyday world (Is gun control associated with decreased murder rates?) and has been the subject of much investigation (see Allan, 1993; Alloy & Tabachnick, 1984; Cheng & Novick, 1992). We employed a paradigm whereby people are presented with covariation information that is accommodated by the format of a 2 × 2 contingency table (see Wasserman, Dorner, & Kao, 1990). The simulated problem for the participants was to determine whether a particular drug improved the condition of psoriasis. The contingency information concerned the number of rats who had or had not been given the drug and the number who had improved or not improved. The four cells of the contingency table were reported to the participants in summary form and they were told to indicate whether the drug had an effect on psoriasis on a scale ranging from −10 (*worsens the psoriasis*), through 0 (*no effect*), to +10 (*helps psoriasis*). The four cells of the standard 2 × 2 contingency table were represented by the four data points given to the participant: the number of rats who received the drug and improved (termed Cell A in the literature), the number of rats who received the drug and did not improve (termed Cell B), the number of rats who did not receive the drug and who improved (Cell C), and the number of rats who did not receive the drug and who did not improve (Cell D). The psoriasis problem was taken from the work of Wasserman et al. (1990), and the actual values used in the 25 problems were taken from Table 2 of that article.

The normatively appropriate strategy in the task is to use the conditional probability rule, whereby the probability of improvement with the drug is compared to the probability of improvement without the drug (see Allan, 1980; Kao & Wasserman, 1993; Shanks, 1995). Numerically, the rule amounts to calculating the $\Delta p$ statistic of conditional probability, $[A/(A + B)])] - [C/(C + D)]$ (see Allan, 1980). Thus, participants' judgments of the degree of contingency in each of the 25 problems were correlated with the $\Delta p$ value for each problem and were used as the index of normatively appropriate processing. The mean correlation for individual participants was .744 ($SD$ = .169) and the median correlation was .783. These individual analyses of participants are highly convergent with those that Wasserman et al. (1990) conducted on the same problems (e.g., our median correlation with $\Delta p$ was .783 compared with their .816). Six of the 543 participants made judgments that failed to correlate significantly with either $\Delta p$ or with any of the cells and were eliminated from the analyses.

*Hypothesis testing.* This task was modeled on the work of Tschirgi (1980). The participants were given eight problems consisting of vignettes in which a story character observed an outcome and had to test a hypothesis about the importance of one of three variables to an outcome. The individual is asked to choose one of three ways to test this hypothesis the next time the situation occurs. The three alternatives correspond to three alternative hypothesis testing strategies: changing all the variables (CA), hold one thing at a time constant (HOLDONE), and vary one thing at a time (VARYONE). In this simplified situation, both the HOLDONE and VARYONE strategies are equally normative (see Baron, 1985, pp. 142–144). However, the CA strategy is clearly nonnormative, thus it is used as the index of performance on this task. The more CA choices the more nonnormative the performance.

The problems were similar to those used by Tschirgi (1980) and Farris and Revlin (1989) or were actual revisions of problems used in those two studies. Four of the eight problems in Experiment 2 had positive outcomes, and four had negative outcomes. The mean number of VARYONE, HOLDONE, and CA choices was 4.19, 3.10, and 0.71, respectively.

## *Heuristic Thinking Tasks*

*Outcome bias.* Our measure of outcome bias derived from a problem investigated by Baron and Hershey (1988):

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 successful 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 as a result of the operation itself.* His physician decided to go ahead with the operation. The operation succeeded. Evaluate the physician's decision to go ahead with the operation.

Participants responded on a 7-point scale ranging from 1 (*incorrect, a very bad decision*) to 7 (*clearly correct, an excellent decision*). This question appeared early in our test battery. Later in the battery, participants evaluated a different decision to perform surgery that was designed to be objectively better than the first (2% chance of death rather than 8%; 10-year increase in life expectancy versus 5-year, etc.) even though it had an unfortunate negative outcome (death of the patient). Participants reported on the same scale as before. An outcome bias was demonstrated when these two items were compared: 152 participants rated the positive outcome decision better than the negative outcome decision, 308 rated the two decisions equally, and 85 participants rated the negative outcome decision as the better decision. The measure of outcome bias that we employed was the scale value of the decision on the positive outcome case (from 1 to 7) minus the scale value of the negative outcome decision. The higher the value, the more the outcome bias. The mean outcome bias score was .272 ($SD$ = 1.2).

*If–only thinking.* If–only thinking refers to the tendency for people to have differential responses to outcomes based on the differences in counterfactual alternative outcomes that might have occurred (Denes-Raj & Epstein, 1994; Epstein, Lipson, Holstein, & Huh, 1992; Miller, Turnbull, & McFarland, 1990). Some participants are more upset at a negative outcome when it is easier to imagine a positive outcome occurring. Two vignettes were taken from the work of Epstein et al. (1992). The first of these had been adapted from Kahneman and Tversky (1982) and read as follows:

> Mr. Paul, who has an average income, owned shares in Company A. During the past year he switched to stock in Company B. He has just learned that the stock in A has skyrocketed, and he would now be $10,000 ahead if he had kept his stock in Company A. Mr. George, who also has an average income, owns shares in Company B. During the past year he considered switching stock to Company A, but decided against it. He has just learned that the stock in A skyrocketed, and he would now be $10,000 ahead if he had made the switch. Putting emotions about the events aside, who do you think acted more foolishly in bringing about the unfortunate outcome that occurred, Mr. Paul or Mr. George?
>
> a. Mr. George acted much more foolishly
> b. Mr. George acted slightly more foolishly
> c. Mr. George and Mr. Paul acted equally foolish
> d. Mr. Paul acted slightly more foolishly
> e. Mr. Paul acted much more foolishly

The item was scored in a 0/1 manner. Participants' responses indicated if–only thinking if they chose alternative D or E. The second vignette was the car damage problem, also taken from Epstein et al. (1992; this item had been adapted from a study described in Miller et al., 1990). Participants thus received a score of 2 if they gave the if–only response on both items, 1 if they gave the if–only response on one item, and zero if they did not give an if–only response on either item. The mean score was .82 (*SD* = 0.7). One hundred and nine participants displayed if–only thinking on both problems, 229 participants displayed if–only thinking on one problem, and 207 participants did not display if–only thinking on either problem.

### *Thinking Dispositions Questionnaire*

The thinking dispositions questionnaire employed in this study was a variant of the instrument used in Experiment 1. Participants completed the 10-item Actively Open-Minded Thinking (AOT) scale, 2-item Counterfactual Thinking scale, 9-item Absolutism scale, 6-item Paranormal Beliefs subscale, and 5-item Social Desirability scale that were employed in Experiment 1. Two items drawn from the Dogmatic Thinking scale used by Paulhus and Reid (1991) were added to the Dogmatism scale. A more complete measure of socially desirable responding was included in Experiment 2 (the 40-item Balanced Inventory of Desirable Responding, Form 40A; see Paulhus, 1991), but again socially desirable responding did not mediate any of the relationships. The response scale of the thinking dispositions questionnaire was changed from Experiment 1. In Experiment 2, the following 6-point scale was used: 1 = *disagree strongly*; 2 = *disagree moderately*; 3 = *disagree slightly*; 4 = *agree slightly*; 5 = *agree moderately*; 6 = *agree strongly*.

A Thinking Dispositions Composite scale (TDC) was formed by summing the scores on the Counterfactual and Actively Open-Minded Thinking subscales and subtracting the sum of the scores on the Absolutism, Dogmatism, and Paranormal Belief subscales. Thus, high scores on the TDC indicate cognitive flexibility and a skeptical open-mindedness, whereas low scores indicate cognitive rigidity and dogmatic credulity.

## Results

### *Relationships Among Rational Thinking Tasks*

The correlations among the five rational thinking tasks are displayed in Table 2. All 10 of the correlations between pairs of tasks were significant at the .001 level. The correlations among the AET, syllogistic reasoning, and statistical reasoning tasks were of roughly the same magnitude as in Experiment 1. However, in contrast to Experiment 1, performance on the AET and statistical reasoning task displayed a significant correlation (.299). The correlations involving the covariation judgment task and hypothesis testing task (number of CA responses) were somewhat lower than those involving the other three tasks, but all were significant nevertheless. In the case of each of the five tasks, the direction of the correlation was that those responding more normatively on one task also tended to respond in accord with the normative model of correct responding on another. The correlations involving the hypothesis testing task are negative because the more CA responses given the more nonnormative the response tendency on that task.

Table 2
*Intercorrelations Among the Rational Thinking Tasks in Experiment 2*

| Variable | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Argument evaluation task | — | | | |
| 2. Syllogisms | .293 | — | | |
| 3. Statistical reasoning | .299 | .292 | — | |
| 4. Covariation: Δp | .161 | .182 | .190 | — |
| 5. Hypothesis testing (CA) | −.198 | −.252 | −.147 | −.204 |

*Note.* Hypothesis testing (CA) = number of change all responses on the hypothesis-testing task. Correlations higher than .145 in absolute value are significant at the .001 level (two-tailed).

Table 3
*Correlations Between the Reasoning Tasks and SAT Total Score, Thinking Dispositions Composite Score, and Mathematics–Statistics Background*

| Reasoning tasks | SAT total | TDC | Math background |
|---|---|---|---|
| Argument evaluation task | .371**** | .296**** | .137*** |
| Syllogisms | .410**** | .329**** | .091 |
| Statistical reasoning | .376**** | .263**** | .075 |
| Covariation: Δ*p* | .239**** | .176**** | .088 |
| Hypothesis testing (CA) | −.223**** | −.167**** | −.045 |
| Outcome bias | −.172**** | −.175**** | −.071 |
| If–only thinking | −.208**** | −.205**** | −.062 |
| RT1 composite | .530**** | .413**** | .145*** |
| RT2 composite | .383**** | .324**** | .125*** |
| RT composite, all tasks | .547**** | .442**** | .162**** |

*Note.* SAT = Scholastic Aptitude Test; TDC = thinking dispositions composite score; Hypothesis testing (CA) = number of change all responses on the hypothesis testing task; RT1 composite = standard score composite of performance on argument evaluation task, syllogisms, and statistical reasoning; RT2 composite = standard score composite of performance on covariation judgment, hypothesis testing task, if–only thinking, and outcome bias; RT composite, all tasks = rational thinking composite score of performance on all seven tasks.
****p* < .01, two-tailed. *****p* < .001, two-tailed.

### *Relationships With Cognitive Ability and Thinking Dispositions*

As indicated in Table 3, SAT total scores were significantly correlated ($p < .001$) with performance on all five rational thinking tasks. Also listed in Table 3 are the correlations with the indexes of performance on the two heuristic processing tasks: outcome bias and if–only thinking. As with the hypothesis testing task, the correlations involving these two tasks are negative because higher scores indicate more nonnormative responding. The range of correlations with the five reasoning tasks (.223 to .410) were similar to the correlations with cognitive ability in Experiment 1 (.298 to .496), even though in Experiment 2 only SAT scores were used instead of a composite. As in Experiment 1, the largest correlation (.410) was with the syllogistic reasoning task.

The next column of Table 3 indicates that the thinking dispositions composite score was also correlated with each of the seven experimental tasks ($p < .001$). As in Experiment 1, the correlations with AET, syllogistic reasoning, and statistical reasoning were somewhat lower than the correlations involving cognitive ability. However, for the remaining four tasks, the correlations involving cognitive ability and thinking dispositions were of roughly similar magnitude.

The final column indicates whether there were correlations between any of the tasks and the extent of the participants' mathematics–statistics background as indicated by the mathematics–statistics background variable. Only one of the seven correlations (with AET performance) was statistically significant, and that correlation (.137) was substantially lower than the correlation with SAT scores (.371) or thinking dispositions (.296). Thus, differences in mathematics–statistics background are not mediating the relationships with cognitive ability or thinking dispositions.

We examined the extent to which joint limitations at the algorithmic and rational levels can explain variation from normative responding by conducting a multiple regression analysis using composite measures of performance on the rational thinking tasks. The first composite involved the three tasks that were carried over from Experiment 1: the AET, statistical reasoning, and syllogistic reasoning tasks. The scores on each of these three tasks were standardized and summed to yield a composite score. SAT Total scores and the TDC attained a multiple *R* with this criterion variable of .600, $F(2, 526) = 148.15$, $p < .001$. Thus, a substantial amount of variance (36%) on these rational thinking tasks is jointly explained by these two predictors. SAT total was a significant unique predictor, partial correlation = .478, unique variance explained = .190, $F(1, 526) = 156.17$, $p < .001$, as was the TDC, partial correlation = .332, unique variance explained = .079, $F(1, 526) = 65.03$, $p < .001$.

A second rational thinking composite was formed by summing the standard scores of the remaining four tasks: covariation judgment, hypothesis testing, if–only thinking, outcome bias (the latter three scores reflected so that higher scores represent more normatively correct reasoning). SAT total scores and the TDC attained a multiple *R* with this rational thinking composite of .447, $F(2, 526) = 65.53$, $p < .001$. SAT total was a significant unique predictor, partial correlation = .325, unique variance explained = .094, $F(1, 526) = 61.94$, $p < .001$, as was the TDC, partial correlation = .249, unique variance explained = .053, $F(1, 526) = 34.88$, $p < .001$. Thus, as with the tasks investigated in Experiment 1, performance on the new rational thinking measures was independently predicted by cognitive ability and thinking dispositions.

Finally, both of the rational thinking composites were combined into a composite variable reflecting performance on all seven tasks. SAT total scores and the TDC attained a multiple *R* with this criterion variable of .627, $F(2, 526) = 170.56$, $p < .001$. SAT total was a significant unique predictor, partial correlation = .496, unique variance explained = .198, $F(1, 526) = 171.88$, $p < .001$, as was the TDC, partial correlation = .366, unique variance explained = .094, $F(1, 526) = 81.51$, $p < .001$. The zero-order correlations involving all three of the rational thinking composite variables are presented at the bottom of Table 3.

## Discussion

The new reasoning tasks introduced in Experiment 2 displayed relationships similar to those carried over from Experiment 1. The positive manifold displayed by the tasks suggests that systematic factors are affecting performance and that discrepancies from normative responses are not simply random performance errors. The finding that the cognitive ability composite was significantly correlated with performance on all seven rational thinking tasks indicates that for each of the tasks, discrepancies between descriptive and normative models of performance can to some extent be explained by algorithmic limitations.

Performance on all seven tasks was likewise significantly related to thinking dispositions. It is important to note that the measure of thinking dispositions proved able to predict individual differences on the rational thinking tasks even after cognitive ability had been partialed out. These cognitive styles appear not to be so strictly determined by algorithmic limitations that they cannot serve as independent predictors of the tendency to reason normatively (Brodzinsky, 1985; Globerson, 1989). This is what we would expect if they are indeed best conceptualized at a different level of cognitive analysis whereby they have at least partial explanatory independence.

On an individual task basis, however, some of the correlations were of a quite modest magnitude. This may indicate that the gap between the descriptive and normative for these tasks results additionally from a systematic tendency to compute the nonnormative response or from some proportion of participants adopting nonstandard task construals. However, it should also be emphasized that many of the relationships involving the rational thinking tasks might be underestimated because of modest reliability. Because of the logistical constraints of multivariate investigations involving so many different tasks, scores on some of these measures were based on a very few number of trials. The statistical reasoning composite score was based on only 7 items, syllogism task performance was based on only 8 items, and the regression weight on the AET for each participant was estimated from only 23 data points. The outcome bias score was based on only a single comparison, and the if–only thinking score was based on only 2 items. Schmidt and Hunter (1996) have recently cautioned laboratory investigators that "if a subject is observed in an exactly repeated situation, the correlation between replicated responses is rarely any higher than .25. That is, for unrehearsed single responses, it is unwise to assume a test-retest reliability higher than .25" (p. 203). The very modest correlations displayed by a task such as the outcome bias measure must be interpreted with this admonition in mind.

Schmidt and Hunter (1996) have demonstrated the well-known fact that aggregation can remedy the attenuation caused by the low reliability of single items. Certainly this phenomenon was present in this investigation. Despite the modest correlations displayed by some individual tasks, cognitive ability and thinking dispositions jointly accounted for a considerable amount of variance (39.3%) in the standard score composite of performance on all seven reasoning tasks. It appears that to a considerable extent, discrepancies between actual performance and normative models can be accounted for by variation in capacity limitations at the algorithmic level and cognitive style differences at the rational level—at least with respect to the tasks investigated in this experiment. In the next two experiments we examine individual differences in situations in which the interpretation of the gap between the descriptive and the normative is much more contentious.

## EXPERIMENTS 3A AND 3B: NONCAUSAL BASE RATES

Regarding most of the tasks investigated in Experiments 1 and 2, there is at least broad agreement that the deviations between descriptive and normative models are indeed deviations. That is, there is broad agreement on the nature of the normative response, although a few of the tasks are still the subject of some contention (e.g., Oaksford & Chater, 1994). For example, these statistical reasoning problems are less controversial because they involved causal aggregate information, analogous to the causal base rates discussed by Ajzen (1977) and Bar-Hillel (1980, 1990)—that is, base rates that had a causal relationship to the criterion behavior. In contrast, noncausal base rates—those bearing no obvious causal relationship to the criterion behavior—have been the subject of over a decade's worth of contentious dispute (Bar-Hillel, 1990; Birnbaum, 1983; L. J. Cohen, 1979, 1981, 1982, 1986; Cosmides & Tooby, 1996; Gigerenzer, 1991, 1993, 1996; Gigerenzer & Hoffrage, 1995; Kahneman & Tversky, 1996; Koehler, 1996; Kyburg, 1983; Levi, 1983; Macchi, 1995).

In Experiments 3A and 3B we examined individual differences in responding on two noncausal base-rate problems that are notorious for provoking philosophical dispute. The participants in Experiment 3A were 184 students who completed the performance battery described in Experiment 1 and the participants in Experiment 3B were 201 students who completed the performance battery described in Experiment 4. The first noncausal base-rate problem was the well-known cab problem (see Bar-Hillel, 1980; Lyon & Slovic, 1976; Tversky & Kahneman, 1982):

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city in which the accident occurred. You are given the following facts: 85 percent of the cabs in the city are Green and 15 percent are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each of the two colors 80 percent of the time. What is the probability that the cab involved in the accident was Blue?

Participants chose from six alternatives (*less than 10%, 10–30%, 30–50%, 50–70%, 70–90%, over 90%*).

For purposes of analysis, responses to this question were scored in terms of whether participants relied on the indicant information, relied on the base rate, or amalgamated the base rate and indicant in a manner approximating Bayes's rule (which yields .41 as the posterior probability of the cab being blue). Operationally, posterior probabilities greater than 70% were scored as reliance on the indicant information, probabilities less than 30% were scored as reliance on the base-rate information, and probabilities between 30% and 70% were interpreted as indicating Bayesian amalgamation. Using this classification scheme, 93 participants were classified as reliant on the indicant (responses greater than 70%), 57 participants were classified as reliant on the base rate (responses less than 30%), and 47 were classified as approximately normatively Bayesian (responses between 30% and 70%).

Table 4 displays the mean scores of these three groups on the other variables examined in Experiment 1. As indicated in Table 4, the mean total SAT scores of the three groups were significantly different. The mean SAT score of the Bayesian participants was the highest (1130), followed by the mean of the participants relying on the indicant information (1094), and the participants relying solely on the base rate information had the lowest mean SAT scores (1062). However, the three participant groups displayed no significant differences on the Raven matrices or the Nelson-Denny Comprehension measure. The three groups were not significantly different in their responses to the syllogistic reasoning task and the AET. There was a significant overall difference in selection task performance. The Bayesian and indicant participants tended to outperform the base-rate participants (although not significantly so in individual post hoc comparisons). There were no significant differences displayed on the statistical reasoning problems involving causal aggregate information. Interestingly, however, the participants who were most Bayesian on the cab problem were least likely to use causal statistical aggregate information. No differences were displayed on the AET.

The bottom half of Table 4 indicates these largely null effects were replicated in Experiment 3B. None of the overall trends were significant. As in Experiment 3A, the base-rate participants tended to have the lowest performance on most of the tasks. Also as in Experiment 3A, on no task did the Bayesian participants significantly outperform the indicant participants. Bayesian participants even failed to display more statistical reasoning on causal aggregate problems.

This lack of processing superiority on the part of the Bayesian participants was even more marked on a second noncausal base-rate task—an AIDS testing problem modeled on Casscells, Schoenberger, and Graboys (1978):

> Imagine that AIDS occurs in one in every 1,000 people. Imagine also there is a test to diagnose the disease that always gives a positive result when a person has AIDS. Finally, imagine that the test has a false positive rate of 5 percent. This means that the test wrongly indicates that AIDS is present in 5 percent of the cases where the person does not have AIDS. Imagine that we choose a person randomly, administer the test, and that it yields a positive result (indicates that the person has AIDS). What is the probability that the individual actually has AIDS, assuming that we know nothing else about the individual's personal or medical history?

Participants responded on the same response scale as the cab problem.

Table 4
*Mean Task Performance for the Groups Classified as Indicant, Base Rate, and Bayesian on the Cab Problem*

| | Group | | | |
|---|---|---|---|---|
| Task | Indicant | Base rate | Bayesian | Significance level |
| | | Experiment 3A | | |
| | ($n$ = 93) | ($n$ = 57) | ($n$ = 47) | |
| SAT total | 1093 | 1062$^a$ | 1130$^b$ | $F(2, 181) = 4.07$** |
| Raven matrices | 9.95 | 8.95 | 9.55 | $F(2, 194) = 2.03$ |
| Nelson-Denny | 19.94 | 19.51 | 19.94 | $F(2, 194) = 0.48$ |
| Syllogisms | 4.73 | 4.11 | 4.02 | $F(2, 192) = 2.20$ |
| Selection task | 1.84 | 0.72 | 1.62 | $F(2, 188) = 3.07$* |
| Statistical reasoning | .293 | .151 | −.764 | $F(2, 194) = 2.12$ |
| AET | .350 | .337 | .325 | $F(2, 191) = 0.25$ |
| | | Experiment 3B | | |
| | ($n$ = 128) | ($n$ = 38) | ($n$ = 45) | |
| SAT total | 1145 | 1096 | 1129 | $F(2, 198) = 2.85$ |
| Raven matrices | 9.47 | 8.69 | 8.95 | $F(2, 189) = 0.88$ |
| Nelson-Denny | 20.40 | 19.95 | 19.83 | $F(2, 194) = 0.90$ |
| Syllogisms | 5.23 | 4.74 | 5.44 | $F(2, 208) = 1.24$ |
| Statistical reasoning | .261 | −.866 | −.011 | $F(2, 208) = 1.80$ |

*Note.* Means with different superscripts (a, b) are significantly different (Scheffé). SAT = Scholastic Aptitude Test; Raven matrices = Raven Advanced Progressive Matrices (Set II, Raven, 1962); Nelson-Denny = Nelson-Denny Reading Test (Form F; Brown, Bennett, & Hanna, 1981); AET = argument evaluation test.
*$p < .05$. **$p < .025$.

The Bayesian posterior probability for this problem is slightly less than .02. Thus, responses of less than 10% were interpreted as indicating Bayesian amalgamation, responses of over 90% were scored as indicating strong reliance on indicant information, and responses between 10% and 90% were scored as intermediate. In Experiment 3A, 107 participants were classified as strongly reliant on indicant information (responses over 90%), 50 were classified as intermediate (responses between 10% and 90%), and 40 were classified as approximately Bayesian (responses less than 10%).

As indicated in Table 5, the three groups displayed a significant difference in their mean total SAT scores. The mean SAT scores of the participants strongly reliant on indicant information (1115) was higher than the mean score of the Bayesian participants (1071) whose mean was higher than that of the group showing moderate reliance on indicant information (1061). Significant differences were also observed on the Raven matrices, the Nelson-Denny Comprehension measure, and the syllogistic reasoning task. In each case, the indicant participants outperformed the other two groups. No significant differences were obtained on the selection task, statistical reasoning task, and AET.

The results displayed at the bottom of Table 5 indicate that exactly the same trends were apparent in Experiment 3B. The mean SAT scores of the participants strongly reliant on indicant information (1153) was significantly higher than the mean score of either the Bayesian participants (1103) or the mean score of the intermediate participants (1109). There were no significant differences on the Raven matrices, the Nelson-Denny Comprehension measure, or the syllogistic reasoning task, although the group highly reliant on indicant information had the highest mean in all cases. There was a statistically significant difference in the group mean scores on the composite score for the causal aggregate statistical reasoning problems. Most interestingly, however, were the direction of the differences. The highest mean score was achieved by the group highly reliant on the indicant information in the AIDS problem (.726), followed by the mean of the group showing moderate reliance on indicant information (−.840). The participants giving the Bayesian answer on the AIDS problem were least reliant on the aggregate information in the causal statistical reasoning problems (−1.051).

These results, taken in conjunction with the milder tendencies in the same direction in the cab problem, indicate that the noncausal base-rate problems display patterns of individual differences quite unlike those shown on the causal aggregate problems. On the latter, participants giving the statistical response scored consistently higher on measures of cognitive ability and were disproportionately likely to give the standard normative response on other rational thinking tasks (see Tables 1–3). This pattern did not hold for the two noncausal problems. There were few significant differences on the cab problem, and on the AIDS problem the significant differences were in the opposite direction: Participants strongly reliant on the indicant information scored higher on measures of cognitive ability and were more likely to give the standard normative response on other rational thinking tasks. Thus, discrepancies between descriptive and normative models on noncausal base-rate problems are not as easily explained by recourse to cognitive capacity

Table 5
*Mean Task Performance for the Groups Classified as Indicant, Intermediate, and Bayesian on the AIDS Problem*

| | Group | | | |
|---|---|---|---|---|
| Task | Indicant | Intermediate | Bayesian | Significance level |
| | | Experiment 3A | | |
| | ($n$ = 107) | ($n$ = 50) | ($n$ = 40) | |
| SAT total | 1115$^{a}$ | 1061$^{b}$ | 1071 | $F$(2, 181) = 4.26** |
| Raven matrices | 10.09$^{a}$ | 8.56$^{b}$ | 9.40 | $F$(2, 194) = 4.82*** |
| Nelson-Denny | 20.23 | 19.52 | 19.05 | $F$(2, 194) = 3.09* |
| Syllogisms | 4.79$^{a}$ | 3.66$^{b}$ | 4.21 | $F$(2, 192) = 4.65** |
| Selection task | 1.61 | 1.46 | 1.11 | $F$(2, 188) = 0.48 |
| Statistical reasoning | .421 | −.472 | −.537 | $F$(2, 194) = 2.42 |
| AET | .345 | .322 | .351 | $F$(2, 191) = 0.30 |
| | | Experiment 3B | | |
| | ($n$ = 118) | ($n$ = 57) | ($n$ = 36) | |
| SAT total | 1153$^{a}$ | 1109$^{b}$ | 1103$^{b}$ | $F$(2, 198) = 4.60** |
| Raven matrices | 9.49 | 9.04 | 8.64 | $F$(2, 189) = 0.89 |
| Nelson-Denny | 20.47 | 20.08 | 19.47 | $F$(2, 194) = 1.95 |
| Syllogisms | 5.40 | 4.93 | 4.92 | $F$(2, 208) = 1.32 |
| Statistical reasoning | .726$^{a}$ | −.840$^{b}$ | −1.051$^{b}$ | $F$(2, 208) = 7.24*** |

*Note.* Means with different superscripts (a, b) are significantly different (Scheffé). SAT = Scholastic Aptitude Test; Raven matrices = Raven Advanced Progressive Matrices (Set II, Raven, 1962); Nelson-Denny = Nelson-Denny Reading Test (Form F; Brown, Bennett, & Hanna, 1981); AET = argument evaluation test.
*$p$ < .05. **$p$ < .025. ***$p$ < .01.

differences at the algorithmic level. As mentioned previously, theorists in the heuristics and biases literature are sometimes criticized for explaining divergences between normative models and actual performance by claiming that limitations in computational capacity prevent the normative response. But Lopes and Oden (1991) have claimed that there is "no support for the view that people would choose in accord with normative prescriptions if they were provided with increased capacity" (p. 209). With regard to noncausal base rates, we have failed to find such evidence in Experiments 3A and 3B (in contrast to the tasks investigated in Experiments 1 and 2). In Experiment 4 we turned our attention to other alleged cognitive biases that have been the focus of considerable controversy.

## EXPERIMENT 4

One of the paradigms that has generated such controversy is the subjective probability calibration experiment (Lichtenstein, Fischhoff, & Phillips, 1982). One of the goals of this research has been to assess whether subjective probability estimates can be validated against external criteria. For example, across a set of occasions on which people assess the probability of particular events as $P$, it is possible to examine whether the events do occur $P$ proportion of the time. The knowledge calibration variant of subjective probability assessment has been the most investigated (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991; Juslin, Winman, & Persson, 1994; Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982). In this task situation, people answer multiple choice or true–false questions and, for each item, provide a judgment indicating their subjective probability that their answer is correct. It has traditionally been assumed that perfect one-to-one calibration is the normatively appropriate response—that the set of items assigned a subjective probability of .70 should be answered correctly 70% of the time, that the set of items assigned a subjective probability of .80 should be answered correctly 80% of the time, and so forth.

The standard finding of overconfidence on this task (that subjective probability estimates are consistently higher than the obtained percentage correct) has been considered normatively inappropriate (Fischhoff, 1988; Lichtenstein et al., 1982). However, the issue of whether subjective probabilities should be expected to mimic external relative frequencies has been hotly debated in the philosophical literature (Dawid, 1982; Earman, 1992; Howson & Urbach, 1993; Lad, 1984). Gigerenzer et al. (1991) have argued forcefully that many important frequentist theorists (e.g., von Mises, 1957) and subjectivist theorists (e.g., de Finetti, 1989) reject the idea that deviations of subjective probabilities from actual relative frequencies should be considered a reasoning error. Furthermore, on the basis of their theory of probabilistic mental models (PMM), which dictates that perfect calibration should occur only when the items presented to participants have the same cue validities (in a Brunswikian sense) as those in their natural environment, Gigerenzer et al. (1991) have proposed that in the environmentally unrepresentative knowledge calibration experiment properly adapted people should show overconfidence (for a review of the debates surrounding PMM, see Brenner, Koehler, Liberman, & Tversky, 1996; Griffin & Tversky, 1992; Juslin, 1994).

Another effect examined in Experiment 4 is the hindsight bias effect—that people overestimate what they would have known without outcome knowledge (Fischhoff, 1975, 1977; Hawkins & Hastie, 1990). The hindsight effect is thought to arise at least in part from egoistic or esteem-preserving motivations (J. D. Campbell & Tesser, 1983; Greenwald, 1980; Haslam & Jayasinghe, 1995; Hawkins & Hastie, 1990), and it has usually been interpreted as a nonnormative response tendency: "These experiments show that people rapidly rewrite, or fabricate, memory in situations for which this seems dubiously appropriate" (Greenwald, 1980, p. 607).

The social perception experiment—the domain of the so-called false consensus effect (Marks & Miller, 1987; Mullen et al., 1985; Ross, Greene, & House, 1977)—rivals the knowledge calibration experiment in the amount of contentious dispute that it has generated regarding normative issues. The false consensus effect is the tendency for people to project their own opinions when predicting the attitudes, opinions, and behaviors of other people, and it has traditionally been viewed as a nonnormative response tendency. The false consensus effect has usually been thought to arise at least in part from egocentric attributional biases (Gilovich, 1991; Ross, Greene, et al., 1977; Marks & Miller, 1987).

Operationally, the false consensus effect has usually been defined as occurring when participants' estimates of the prevalence of their own position exceeds the estimate of its prevalence by participants holding the opposite position (see Marks & Miller, 1987). However, several authors have argued that the term false consensus when applied to such a situation is somewhat of a misnomer because such a definition of projection says nothing about whether projecting consensus actually decreased predictive accuracy (see J. D. Campbell, 1986; Hoch, 1987). Hoch (1987) demonstrated that a Brunswikian analysis of the opinion prediction paradigm would render some degree of projection normatively appropriate in a variety of situations. Dawes (1989, 1990) demonstrated that, similarly, a Bayesian analysis renders some degree of projection as normatively appropriate because, for the majority of people, there is actually a positive correlation between their own opinion and the consensus opinion. Thus, one's own opinion is a diagnostic datum that should condition probability estimates (see Dawes, 1989, 1990; Hoch, 1987, Krueger & Zeiger, 1993). The formal analyses of the social perception experiment by Hoch (1987) and Dawes (1989, 1990) have raised doubts about the common interpretation of the consensus effect as a normatively inappropriate egocentric bias. We examined whether high or low projection of consensus is associated with greater predictive accuracy in Experiment 4, and we also explored the cognitive characteristics, thinking dispositions, and other response tendencies of people who differ in their degree of projection. Two tasks from Experiments 1 and 2 (the contrary-to-fact syllogisms and the statistical reasoning problems) were also included in Experiment 4.

## Method

### *Participants*

The participants were 211 undergraduate students (100 men and 111 women) recruited through the same introductory psychology participant pool as Experiment 1. The mean reported verbal SAT score of the 201 students who provided scores in this part of the questionnaire was 551 (*SD* = 69), the mean reported mathematical SAT score was 582 (*SD* = 83), and the mean estimated total SAT score was 1133 (*SD* = 110).

### *Tasks*

Several tasks were administered and scored in a manner identical to that employed in Experiments 1 and 2: the syllogisms task, seven statistical reasoning problems, the thinking dispositions questionnaire that was employed in Experiment 1, the Raven matrices (1962), and the Nelson-Denny Comprehension measure (Brown et al., 1981). The cognitive ability composite score was calculated as in Experiment 1. The new tasks were the following:

#### *Knowledge Calibration*

The methods and analyses used in this task were similar to those employed in the extensive literature on knowledge calibration (Fischhoff, 1982; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977, 1980; Ronis & Yates, 1987; Yates et al., 1989). Participants answered 35 general knowledge questions in multiple choice format with four alternatives. Questions were drawn from Zahler and Zahler's (1988) book *Test Your Cultural Literacy*. The items were not sampled randomly but were instead chosen to be relatively difficult. Two sets of 35 questions were compiled (Set A and Set B). One half of the participants completed Set A for the hindsight task (see next section) and Set B for the knowledge calibration task, and the other half of the participants performed the reverse. Every student completed the hindsight task before the knowledge calibration task. The mean percentage correct on Form A (52.3%, *SD* = 11.6) and Form B (56.8%, *SD* = 13.8) did differ significantly, $t(199) = 2.47, p < .025$). However, in certain analyses reported later, the scores on the two forms were standardized on the basis of their own distribution and then reconverted to percentages based on the common distribution with a mean of 54.7% and standard deviation of 13.0.

After answering each question, participants indicated their degree of confidence in their answer. Degree of confidence was indicated on a 5-point scale that was labeled with both probabilities (percentages) and verbal labels: 25% (*chance—I would just be guessing*), 25%–45% (*a little better than guessing*), 45%–65% (*fair chance that I answered correctly*), 65%–85% (*moderately high chance that I answered correctly*), 85%–100% (*high probability that I answered correctly*). For purposes of calculating the various indices of knowledge calibration, the intervals were assigned their midpoint values: .25, .35, .55, .75, and .925, respectively.

Several measures of knowledge calibration were calculated. Yates et al. (1989) and Ronis and Yates (1987) should be consulted for discussions of the computational and conceptual details of these indices. The first such index—computationally the simplest—is our focus. Termed the measure of *over/underconfidence* by Lichtenstein and Fischhoff (1977) and bias by Yates et al. (1989), it is simply the mean percentage confidence judgment minus the mean percentage correct. The mean bias score in our sample was 7.5% (*SD* = 11.1), significantly different from zero, $t(200) = 9.67, p < .001$. The positive sign of the mean score indicates that the sample as a whole displayed overconfidence, the standard finding with items of this type. An overconfidence bias was displayed by 154 (76.6%) of the 201 participants completing this task.

Other indices of performance such as calibration-in-the-small (see Lichtenstein & Fischhoff, 1977; Yates et al., 1989), and resolution (Yaniv, Yates, & Smith, 1991; Yates et al., 1989) produced values that were typical of those in the literature.

### Hindsight Bias

Participants responded to the alternate form of 35 general knowledge questions except that the answers to the items were indicated with an asterisk. Participants were given the following instructions:

> Below, you will see a series of multiple choice questions on a variety of topics. The correct answer to the item is indicated by an asterisk. We are interested in seeing how students perceive the difficulty of these items. Please read each item and indicate on the scale provided the probability that you would have answered this item correctly.

They then responded on the same scale as in the knowledge calibration task.

The mean estimated probability across the entire sample was 67.3%. This estimated probability was significantly higher than the actual percentage correct (54.7%) achieved on the items answered in the knowledge calibration task, $t(200) = 13.75, p < .001$, which the participant completed subsequently and which was composed of items of equal difficulty. This finding indicates, overall, that the sample was characterized by hindsight bias, replicating previous findings (Fischhoff, 1975, 1977; Hawkins & Hastie, 1990). Of the 201 participants who completed this task, 171 (85.1%) displayed a hindsight bias. The measure of hindsight bias employed for the purposes of individual difference analyses was simply the percentage estimate on the hindsight section minus the percentage correct on the knowledge calibration questions. For the latter, the standardized percentages, which equated the difficulty of the two forms, were employed.

### Opinion Prediction (False Consensus Stimuli)

Participants were presented with 30 statements used in previous consensus judgment research (e.g., I think I would like the work of a school teacher). Sixteen items were taken from Dawes (1990), 6 from Hoch (1987), 4 from Sanders and Mullen (1983), and 4 from J. D. Campbell (1986). The student first indicated whether they agreed with the item. Then, for each item, they answered the question, "What percentage of the students participating in this study do you think agree with the statement?" The items elicited a wide range of levels of agreement and perceived consensus.

At the group, or between-subjects, level of analysis, we replicated the false consensus effect. On 26 of the 30 items the percentage estimate of people who endorsed an item was higher than that of nonendorsers, and 24 of these 26 consensus effects were statistically significant. Thus, statistically robust indications of a consensus effect were observed in the data.

However, the between-subjects analysis tells us nothing about individual differences, and it tells us nothing about whether or not projecting consensus helps an individual more accurately predict the position of others. This was perhaps most clearly demonstrated in Hoch's (1987) Brunswikian analysis of the false consensus effect and in that of Krueger and Zeiger (1993). Instead of considering performance on individual items aggregated across participants, Hoch's (1987) analysis (and that of Krueger & Zeiger, 1993) focuses on the performance of individual participants aggregated across items. For a particular item, call the actual percentage agreement in the target population the *target position* ($T$). Call the participant's own agreement with the item (scored 0/1) *own position* ($O$). Call the participant's prediction of the level of agreement in the population the *prediction* ($P$). One measure of projection (hereafter called *Projection Index 1*) is the beta weight for own position ($O$) when $P$ is regressed on $T$ and $O$. A second measure of projection (hereafter called *Projection Index 2*) was introduced by Krueger and Zeiger (1993). They suggested that an index of overprojection (false consensus) or underprojection (false uniqueness) at the individual level that is relative to the actual accuracy achieved can be derived by correlating for each participant, across items, the agreement with the item (0/1) with the difference between the predicted percentage and the target percentage ($P - T$).

## Results

The correlations among all of the tasks in Experiment 4 are displayed in Table 6. As in Experiments 1 and 2, performance on the syllogistic and statistical reasoning tasks was significantly correlated ($r = .311, p < .001$). Also replicating previous trends, both of these tasks were correlated with the cognitive ability composite ($p < .001$).

Table 6
*Intercorrelations Among the Primary Variables in Experiment 4*

| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. Syllogisms | — | | | | | | |
| 2. Statistical reasoning | .311**** | — | | | | | |
| 3. Overconfidence bias | −.090 | −.118 | — | | | | |
| 4. Hindsight bias | −.147* | −.158* | .785**** | — | | | |
| 5. Projection Index 1 | −.060 | −.102 | .224*** | .175* | — | | |
| 6. Projection Index 2 | −.041 | −.098 | .111 | .073 | .773**** | — | |
| 7. Cognitive abil. comp. | .331**** | .255**** | −.198* | −.250**** | −.133 | −.077 | — |
| 8. TDC | .120 | .277**** | −.045 | −.092 | −.199*** | −.117 | .139* |

*Note.* Projection Index 1 = beta weight for own opinion in the analysis of performance on the opinion prediction task; Projection Index 2 = correlation of item agreement ($O$) with the difference between the predicted percentage and the target percentage ($P$ minus $T$); Cognitive abil. comp. = cognitive ability composite score; TDC = thinking dispositions composite score.
*$p < .05$, two-tailed. ***$p < .01$, two-tailed. ****$p < .001$, two-tailed.

Statistical reasoning, but not syllogistic reasoning, was correlated with the thinking dispositions composite score ($p < .001$). Neither of these tasks, however, were correlated with degree of overconfidence bias or with either index of projection on the opinion prediction task. Both tasks displayed low but statistically significant ($p < .05$) correlations with degree of hindsight bias. The latter was necessarily highly correlated with the degree of overconfidence bias because both shared the same denominator (the percentage of items on the second block of 35 items). Both biases displayed significant correlation with Projection Index 1 but not Projection Index 2. Both hindsight bias and overconfidence bias displayed significant negative correlations with the cognitive ability composite. The direction of the correlation indicates that participants displaying larger biases were lower in cognitive ability. Neither bias correlated with the thinking dispositions composite. Neither projection index correlated with the syllogistic reasoning task, statistical reasoning task, or cognitive ability composite. Projection Index 1 displayed a significant correlation with the thinking dispositions composite, but Projection Index 2 did not.

Table 7 presents the results of multiple regression analyses that examined the extent to which performance on each of the rational thinking tasks could be predicted by the cognitive ability composite and thinking dispositions composite. The multiple *R*s and standardized beta weights for both the cognitive ability composite and the thinking dispositions composite in the final simultaneous equation are presented in Table 7. Cognitive ability was a significant unique predictor of performance on each of the tasks except the opinion prediction task whereby it failed to predict the degree of projection indicated by either of the indices. The thinking dispositions composite was a significant unique predictor of performance on the statistical reasoning task and Projection Index 1.

The individual differences analyses in Experiment 4 were particularly unsuccessful in explaining the gap between the descriptive and the normative on the opinion prediction task. This finding might be viewed as reinforcing the previous critiques of the standard normative analysis of the false consensus effect (Dawes, 1989, 1990; Hoch, 1987, Krueger & Zeiger, 1993), which have questioned whether total lack of projection of one's own position is actually normative. Other aspects of our data reinforce this skepticism. For example, when predictive accuracy is analyzed, our data indicate that projection is actually efficacious.

Hoch (1987) has argued that one index of predictive accuracy is $r(T,P)$: the correlation, across items, between the actual percentage agreement in the target population and the participant's prediction of the level of agreement in the population. We calculated $r(T,P)$ for each of the 185 participants in our sample who completed this task. The mean correlation was .548 ($SD$ = .204), indicating a moderate to high predictive accuracy for our sample (see Hoch, 1987). The reason that some degree of projection might be efficacious is that most participants have positive correlations between their own opinions ($O$) and the target position ($T$). This was demonstrated empirically in Hoch's (1987) study and was true in our investigation, in which the mean $r(T,O)$ correlation was .532 ($SD$ = .170). The degree of perceived consensus is indexed by the correlation $r(P,O)$, which had a mean value of .464 ($SD$ = .238) in this sample.

Because there actually is a correlation between own position and target position for most participants in this sample, mean $r(T,O)$ = .532, in order to accurately predict the target position, participants actually did have to perceive consensus and project it. In fact, there was a positive correlation of .689 between the degree of perceived consensus, $r(P,O)$, and predictive accuracy, $r(T,P)$, in our sample. Also, the degree of perceived consensus, mean $r(P,O)$ = .464, is lower than the degree of actual consensus in the sample, mean $r(T,O)$ = .532, which itself suggests that these participants may not be overprojecting. The correlations between Projection Indices 1 and 2 and various indices of predictive accuracy are consistent with this conjecture. For example, the two indices were both positively correlated with $r(T,P)$, the correlational measure of predictive accuracy ($r = .153, p < .05$, and $r = .144, p < .05$, respectively).

Table 7
*Results of Multiple Regression Analyses Conducted on the Primary Variables in Experiment 4*

| Variable | Mult. *R* | β cog. ability | β TDC |
|---|---|---|---|
| Syllogisms | .339**** | .320**** | .076 |
| Statistical reasoning | .353**** | .221**** | .246**** |
| Overconfidence bias | .199* | −.196*** | −.019 |
| Hindsight bias | .257*** | −.242*** | −.059 |
| Projection Index 1 | .225*** | −.107 | −.184* |
| Projection Index 2 | .132 | −.061 | −.108 |

*Note.* Mult. *R* = multiple correlation; cog. ability = cognitive ability composite score; TDC = thinking dispositions composite score; β = standardized beta weight; Projection Index 1 = β weight for own opinion in the analysis of performance on the opinion prediction task; Projection Index 2 = correlation of item agreement (*O*) with the difference between the predicted percentage and the target percentage (*P* minus *T*).
*$p < .05$, two-tailed. ***$p < .01$, two-tailed. ****$p < .001$, two-tailed.

## GENERAL DISCUSSION

As outlined in the introduction, there are several reasons why a descriptive account of actual reasoning performance might not accord with normative theory (see L. J. Cohen, 1981, and Stein, 1996, for extensive discussions of the various possibilities). First, performance may depart from normative standards because of performance errors—temporary lapses of attention, memory deactivation, and other sporadic information-processing mishaps. Secondly, there may be algorithmic limitations that prevent the normative response (Cherniak, 1986; Oaksford & Chater, 1993, 1995; Stich, 1990). Thirdly, in interpreting performance, we might have applied the wrong normative model to the task. Alternatively, we may have applied the correct normative model to the problem as set, but the participant might have construed the problem differently and be providing the normatively appropriate answer to a different problem (Adler, 1984; Broome, 1990; Hilton, 1995; Oaksford &

Chater, 1994; Schwarz, 1996). Finally, the participant might be systematically computing the response from a nonnormative rule (Baron, 1991b; Kahneman & Tversky, 1996).

Patterns of individual differences may have implications for some of these explanations of discrepancies between descriptive and normative models. If discrepancies from normative models are due to performance errors or to differential construals there would be little reason to expect these discrepancies to correlate across reasoning tasks or to correlate with measures of cognitive ability (particularly in the case of performance errors). The cab problem investigated in Experiment 3 comes closest to displaying this pattern. Across two experiments involving a total of 954 participants, on no other reasoning task was there a statistically significant difference between those giving an indicant-dominated response on the cab problem and those giving a Bayesian response. These two groups also did not differ on any measure of cognitive ability (SAT, the Raven matrices, and the Nelson-Denny Comprehension measure). Our interpretation of these null effects is that differential construal accounts for the indicant group's departure from the normatively appropriate Bayesian response. We have been drawn to this interpretation by the growing literature that calls into question how participants are interpreting this task (Birnbaum, 1983; Braine, Connell, Freitag, & O'Brien, 1990; Koehler, 1996; Macchi, 1995). For example, several authors (see Braine et al., 1990; Macchi, 1995) have discussed how in the phrasing "the witness correctly identified each of the two colors 80 percent of the time" the words *correctly identified* might suggest to participants that the 80% refers to $P(H/D)$ rather than to $P(D/H)$. Given this construal, the participant might be thought to be responding normatively.

In contrast to discrepancies caused by differential construal, if performance discrepancies from normative models are due to algorithmic limitations then correlations between performance on the reasoning task and measures of cognitive ability would be expected. Virtually all of the deductive reasoning, inductive reasoning, methodological thinking, and heuristic reasoning tasks examined in Experiments 1 and 2 displayed this pattern, indicating that at least to some significant degree algorithmic limitations prevent fully normative responding.

Finally, if the wrong normative model is being applied to performance, it might be expected that the correlation would go in the other direction—that those of higher cognitive ability would produce responses that are more discrepant from the incorrect normative standard being applied. A finding suggestive of this pattern was obtained on the AIDS base-rate problem of Experiment 3 in which those giving responses of over 90% (consistent with total reliance on indicant information) consistently scored higher on the cognitive ability and reasoning tasks than did those giving responses closer to the normatively appropriate Bayesian response. Of importance is the fact that this included the causal base-rate statistical reasoning problems, in which participants who were more reliant on the indicant in the AIDS problem were found to be more likely to rely on the aggregate information in the statistical reasoning problems. Interestingly, the AIDS problem (or close variants of it) has been the focus of intense debate in the literature and several authors have argued against making the automatic assumption that the indicant response is nonnormative in the version that we used in Experiment 3 (L. J. Cohen, 1979, 1981, 1982, 1986; Cosmides & Tooby, 1996; Gigerenzer, 1991, 1993, 1996; Gigerenzer & Hoffrage, 1995; Koehler, 1996; Kyburg, 1983; Levi, 1983). The result might also suggest that the Bayesian participants on the AIDS problem might not actually be arriving at their response through anything resembling Bayesian processing (see Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995) because on causal aggregate statistical reasoning problems these participants were less likely to rely on the aggregate information.

## Individual Differences and Prescriptive Models

Although there were algorithmic influences on rational thinking revealed on many of the reasoning tasks investigated here, these should not be viewed as absolute limitations because much variability remained after cognitive ability had been accounted for. As previously discussed, the magnitude of many of the correlations leaves much systematic variance unaccounted for and thus considerable room for the possibility that participants are systematically computing according to nonnormative rules and are not necessarily doing so to circumvent limitations in cognitive capacity (Baron, 1991b; Shafir, 1994). The loose connections with cognitive capacity can be illustrated by pointing to some specific instances of normative performance. For example, in Experiments 1 and 4 several individuals with total SAT scores less than 1100 answered each of the contrary-to-fact syllogisms correctly, and in Experiment 2 individuals with total SATs as low as 960 and 990 answered all of these items correctly. In Experiment 4, several individuals with SAT scores in the 1000–1100 range gave the aggregate response on at least six out of seven statistical reasoning problems. In the covariation task in Experiment 2, individuals with SAT scores as low as 910 and 960 had correlations of over .90 with the normatively appropriate $\Delta p$ strategy. Eight individuals with SAT scores less than 900 displayed no outcome bias in Experiment 2, as did many students with SAT scores in the 900–1000 range. In short, the normative responses for these particular tasks could be computed by students who had modest cognitive abilities.

Assuming SAT scores as a gross indicator of cognitive capacity (an assumption for which there is growing evidence; see Deary & Stough, 1996; Vernon, 1993), there appear to be no computational limitations preventing most of the college population from producing normative performance on any of the tasks investigated in Experiment 2. Using Baron's (1985, 1994, 1996) prescriptive–normative distinction, one might say that algorithmic-level limitations do not prevent the prescriptive response from approximating the normative one for a majority of the population examined here. Furthermore, for many of the tasks examined in Experiments 1 and 2, part of the descriptive–normative discrepancy that remains even when differences in cognitive ability are accounted for is predictable from differences in thinking dispositions. If Baron's (1985) suggestion that thinking dispositions are more malleable than cognitive capacity is accepted, then one may take this finding as additional evidence that prescriptive models should approximate normative ones.

## Thinking Styles and Rational Thought

That thinking dispositions could serve as a predictor independent of differences in cognitive ability in Experiments 1 and 2 supports the distinction between thinking dispositions and cognitive capacities that is championed by some investigators (e.g., Baron, 1985, 1988). It is possible that the cognitive styles considered in our study represent a mixture of psychological mechanisms. For example, some of the thinking styles could be conceptualized as stored rules with propositional content (e.g., "think of alternative explanations," "think of a reason against your position"). However, others may be less rule-like (e.g., dogmatism, absolutism), and their dissociation from computational capacity may suggest an interpretation similar to that given to the function of emotions by Johnson-Laird and Oatley (1992), as interrupt signals supporting goal achievement.

Johnson-Laird and Oatley's (1992) discussion of the rationality of emotions (see also, de Sousa, 1987; Oatley, 1992) emphasizes pragmatic rationality: the coordination and achievement of goals (Audi, 1993a, 1993b; Stich, 1990). However, it is also possible that some of the thinking styles discussed above may be conceived of as signals (e.g., "avoid closure," "keep searching for evidence," etc.) that have effects primarily on mechanisms that serve the ends of epistemic rationality: processes of fixing beliefs in proportion to evidence and in coherent relationships with other beliefs (Harman, 1986; Kitcher, 1993; Thagard, 1992). In short, thinking dispositions of the type we have examined may provide information about epistemic goals at the rational level of analysis (see Anderson, 1990).

The importance of thinking styles in discussions of human rationality has perhaps not received sufficient attention because of the heavy reliance on the competence–performance distinction in philosophical treatments of rational thought in which all of the important psychological mechanisms are allocated to the competence side of the dichotomy. From one such view, L. J. Cohen (1982) has argued that there really are only two factors affecting performance on rational thinking tasks: "normatively correct mechanisms on the one side, and adventitious causes of error on the other" (p. 252). Not surprisingly given such a conceptualization, the processes contributing to error (adventitious causes) are of little interest to L. J. Cohen (1981, 1982). Human performance arises from an intrinsic human competence that is impeccably rational, but responses occasionally deviate from normative correctness because of inattention, memory lapses, lack of motivation, distraction, momentary confusion, and other fluctuating but basically unimportant causes. There is nothing in such a view that would motivate any interest in patterns of errors or individual differences in such errors.

In contrast, Overton and Newman (1982) have argued for a more balanced view of the competence–performance divide, one in which competence models and models of the moderators of competence (cognitive styles, motivation, etc.) have more equal status. They have argued that competence theories have problems in accounting for variation in performance and that purely procedural theories have trouble accounting for important underlying constancies. As do Overton and Newman (1982), Johnson-Laird and Byrne (1993) have articulated a view of rational thought that parses the competence–performance distinction very differently from that of L. J. Cohen (1981, 1982, 1986) and that simultaneously leaves room for cognitive styles to play an important role in determining responses. At the heart of the rational competence that Johnson-Laird and Byrne (1993) have attributed to humans is only one meta-principle: People are programmed to accept inferences as valid provided that they have constructed no mental model of the premises that contradict the inference. Inferences are categorized as false when a mental model is discovered that is contradictory. However, the search for contradictory models is "not governed by any systematic or comprehensive principles" (p. 178).

The key point in Johnson-Laird and Byrne's (1993) account is that once an individual constructs a mental model from the premises, once the individual draws a new conclusion from the model, and once the individual begins the search for an alternative model of the premises that contradicts the conclusion, the individual *"lacks any systematic method to make this search for counter-examples"* (p. 205, italics added). In this passage, Johnson-Laird and Byrne seem to be arguing that there are no systematic control features of the search process. But epistemically related cognitive dispositions may in fact be reflecting just such control features. Individual differences in the extensiveness of the search for contradictory models could arise from a variety of cognitive factors that may be far from "adventitious," although not completely systematic (see Oatley, 1992; Overton, 1985, 1990): factors such as cognitive confidence, reflectivity, need for cognition, ideational generativity, dispositions toward confirmation bias and premature closure, etc.

It is possible that one of the characteristic cognitive styles that may be accounting for the common variance carried by several of the rational thinking tasks that did converge in the individual difference analyses is the tendency to *decontextualize* reasoning (Denny, 1991; Donaldson, 1978, 1993; Epstein, 1994; Evans, 1984, 1989; Luria, 1976; Neimark, 1987; Olson, 1977, 1994; Piaget, 1926, 1972). Decontextualization is the feature of thought that is actually emphasized by many critics of the heuristics and biases literature who, nevertheless, fail to see it as implying a research program for differential psychology. For example, if to contextualize a problem is the natural and nearly universal reasoning style of human beings (what might be called the *fundamental computational bias*), then it is not surprising that many people respond incorrectly when attempting a psychological task that is explicitly designed to require a decontextualized reasoning style (contrary-to-fact syllogisms, base-rate problems, argument evaluation, etc.). But the fact that some people do give the decontextualized response means that at least some people have available a larger repertoire of reasoning styles (they can flexibly reason so as to override the fundamental computational bias if the situation requires).[2]

## Alternative Explanations

One possible mechanism accounting for the patterns of covariance in task performance arguably may be differential educational experience. Perhaps, for example, more intelligent participants have been exposed to superior educational experiences: experiences in which they were more likely to be taught the normatively appropriate responses to the tasks. The plausibility of this argument varies greatly across the various tasks that were investigated. For example, the causal aggregate statistical problems are presented to the participants as having no right or wrong answers. These problems did not require any numerical calculation, and nothing like them appears in any curriculum known to us. Empirically, the correlation between the composite score on the aggregate reasoning problems and formal statistics exposure at the college level was .110 in Experiment 2: a relationship not large enough to sustain a strong explanation in terms of educational experience. As Table 3 indicates, across the seven tasks in Experiment 2, the mathematics–statistics background composite variable displayed a significant correlation with only one task (the AET). The absolute magnitude of the seven correlations ranged from .045 to .137. Even the composite variable of all seven rational thinking tasks displayed a correlation of only .162 with the mathematics–statistics background variable, considerably lower than that with the TDC (.442) or SAT total score (.547).

Finally, it is not the case that more intelligent students were simply more sensitive to demand characteristics that encouraged a socially desirable normative response. Across all of the experiments, a consistent finding was that the students giving the normative response were not more prone to make socially desirable responses. More often, the trend was in the opposite direction.

## Conclusions

In reply to L. J. Cohen's (1981) well-known critique of the heuristics and biases literature—surely the most often cited of such critiques—Jepson et al. (1983) have argued that "Cohen postulates far too broad a communality in the reasoning processes of the 'untutored' adult" (p. 495). Jepson et al., we have argued, were right on the mark—but their argument has been largely ignored in more recent debates about human rationality and the tasks that we have used to assess it (see Roberts, 1993; Yates et al., 1996). Although on all tasks many participants displayed the characteristic biases that have been observed in the literature, we consistently uncovered enormous individual variation on each of the tasks that we investigated, and there were almost always a few participants whose performance was almost perfectly optimal from a normative point of view: Some participants got all the syllogisms correct, some were almost perfectly correlated with the experts' judgments on the AET, some were nearly perfectly correlated with $\Delta p$ in the covariation task, a few chose the more diagnostic statistical evidence and ignored the vivid case evidence nearly every time, a few made consistently correct choices in the selection task, a reasonable number displayed no outcome bias or if–only thinking, and some chose the VARY-ONE strategy every time.

In fact, critics of the heuristics and biases literature have sometimes mentioned an individual differences result to bolster their position. L. J. Cohen (1982) has critiqued the older "bookbag and poker chip" literature on Bayesian conservatism (Phillips & Edwards, 1966; Slovic, Fischhoff, & Lichtenstein, 1977) by noting that

> if so-called "conservatism" resulted from some inherent inadequacy in people's information-processing systems one might expect that, when individual differences in information-processing are measured on independently attested scales, some of them would correlate with degrees of "conservatism." In fact, no such correlation was found by Alker and Hermann (1971). And this is just what one would expect if "conservatism" is not a defect, but a rather deeply rooted virtue of the system. (pp. 259–260)

This is precisely how Alker and Hermann (1971) themselves argued in their article:

> Phillips et al. (1966) have proposed that conservatism is the result of intellectual deficiencies. If this is the case, variables such as rationality, verbal intelligence, and integrative complexity should have related to deviation from optimality—more rational, intelligent, and complex individuals should have shown less conservatism. The lack of relationship raises questions about the type of ideal decision maker the Bayesian model denotes. (p. 40)

Funder (1987), like L. J. Cohen (1982), has used a finding about individual differences to argue that a particular attribution error should not be considered suboptimal or nonnormative. Block and Funder (1986) analyzed the role effect observed by Ross, Amabile, and Steinmetz (1977): that people rated questioners more knowledgeable than contestants in a quiz game. Although Ross, Amabile, et al. (1977) viewed the role effect as an attributional error—people allegedly failed to adjust their estimates of the knowledge displayed by a consideration of the individual's role—Block and Funder (1986) demonstrated that individuals most susceptible to this attributional "error" were more socially competent, more well adjusted, and more intelligent. Because more socially competent individuals were more prone to this attributional effect, Funder (1987) has argued that the view that this attributional pattern is an error is undermined. Thus, both Funder (1987) and L. J. Cohen (1982) have had recourse to patterns of individual differences to pump our intuitions (see Dennett, 1980) in the direction of undermining our standard normative analysis of the tasks under consideration.

Funder (1987) has compared decision-making biases to optical illusions, arguing that an individual experiencing such an illusion is not characterized as displaying a flaw in

[2] The tendency to decontextualize and to respond normatively are not the same thing, however, although the two often coincide (Baron, 1994; Stanovich, 1996).

judgment or a dysfunctional cognitive mechanism. Indeed, the clear inference to be drawn from this example is that the information-processing system of anyone not subject to the illusion would be suspect. This is how, for example, researchers interpret the finding that mentally retarded individuals are less susceptible to certain illusions (Spitz, 1979). Those less susceptible to the illusion are expected to be characterized by less efficient information processing because the occurrence of the illusion is seen to result from the operation of adaptive cognitive mechanisms.

The use of the visual illusion analogy invites the same inference with regard to rational thinking tasks whereby the modal response departs from the normative response. That is, it is apparently to be inferred that the modal response reflects the operation of adaptive cognitive mechanisms and that the normative response is reflective of inefficient information processing. In short, according to the visual illusion analogy argument, individuals less susceptible to the cognitive illusions in these rational thinking tasks have less efficient cognitive mechanisms—just as in the case of mentally retarded individuals who prove less susceptible to visual illusions. The visual illusion analogy seems appropriate in the case of the AIDS base-rate problem of Experiment 3 whereby participants displaying base-rate neglect (i.e., those more susceptible to the illusion) were indeed higher in cognitive capacity than those less susceptible to base-rate neglect. The problem here is that the results of Experiments 1 and 2 suggest that the analogy to visual illusions might be inappropriate as applied to a number of other tasks (see Baron, 1994). Individuals less susceptible to several cognitive illusions (belief bias, vividness effects, matching bias in the selection task, etc.) were, contrary to the visual illusion analogy, more cognitively competent than those more susceptible. This is not what would be expected if susceptibility to the illusion reflected adaptive functioning (Anderson, 1991; Cooper, 1987; Oaksford & Chater, 1994, 1995; Payne et al., 1993; Shanks, 1995).

Philosopher Nicholas Rescher (1988) has argued that

> to construe the data of these interesting experimental studies [of probabilistic reasoning] to mean that people are systematically programmed to fallacious processes of reasoning—rather than merely that they are inclined to a variety of (occasionally questionable) substantive suppositions—is a very questionable step. . . . While all (normal) people are to be credited with the capacity to reason, they frequently do not exercise it well. (p. 196)

There are two parts to Rescher's (1988) point here: the "systematically programmed" part and the "inclination toward questionable suppositions" part. Rescher's (1988) focus—like that of many who have dealt with the philosophical implications of the idea of human irrationality (L. J. Cohen, 1981, 1982, 1986; Davidson, 1980; Dennett, 1987; Goldman, 1986; Harman, 1995; Kornblith, 1993; Stein, 1996; Stich, 1990)—is on the issue of how humans are systematically programmed. Inclinations toward questionable suppositions are only of interest to those in the philosophical debates as mechanisms that allow one to drive a wedge between competence and performance (L. J. Cohen, 1981, 1982; Rescher, 1988), thus maintaining a theory of near-optimal human rational competence in the face of a host of responses that seemingly defy explanation in terms of standard normative models (Baron, 1993a, 1994; Capon & Kuhn, 1979; Dawes, 1988; Gilovich, 1991; Griffiths, 1994; Shafir, 1994; Shafir & Tversky, 1992; Sutherland, 1992; Thaler, 1992; Wagenaar, 1988).

One of the purposes of our research program is to reverse the figure and ground in this dispute, which has tended to be dominated by the particular way that philosophers frame the competence–performance distinction. Developmental psychologists have long held a more balanced view of the matter. For example, Neimark (1981; see also, Overton, 1990) has argued that

> in focusing upon inferred competence as a central topic of theoretical concern, individual difference factors, and all the stimulus components of ambiguous experimental procedures which give rise to them, may be dismissed as sources of noisy variability or bias to be eliminated at all cost. However, when interest shifts to description of performance as a legitimate topic of theoretical interest, then all these banished "undesirables" come flooding back demanding attention. A complete psychological theory, and its application to everyday problems, requires a full account of how competence is translated into performance and how it is masked or amplified by a plethora of varying conditions. (p. 187)

In short, from a psychological standpoint, there may be important implications in precisely the aspects of performance that have been backgrounded in the controversy about basic reasoning competence. That is, whatever the outcome of the disputes about how humans are systematically programmed (Cosmides, 1989; Johnson-Laird & Byrne, 1991, 1993; Johnson-Laird, Byrne, & Schaeken, 1994; Oaksford & Chater, 1993, 1994; O'Brien, Braine, & Yang, 1994; Rips, 1994), variation in the inclination toward questionable suppositions is of psychological interest as a topic of study in its own right (see Roberts, 1993). The experiments reported here indicate, at least for certain subsets of tasks, that the inclination toward questionable suppositions has some degree of domain generality, that it is in some cases linked to computational limitations, and that it is predicted by thinking dispositions that can be related to the epistemic and pragmatic goals of rational thought.

## References

Adler, J. E. (1984). Abstraction is uncooperative. *Journal for the Theory of Social Behaviour, 14,* 165–181.

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35,* 303–314.

Alker, H., & Hermann, M. (1971). Are Bayesian decisions artificially intelligent? The effect of task and personality on conservatism in information processing. *Journal of Personality and Social Psychology, 19,* 31–41.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15,* 147–149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin, 114,* 435–448.

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

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14,* 471–517.

Arkes, H., & Hammond, K. (Eds.). (1986). *Judgment and decision making.* Cambridge, England: Cambridge University Press.

Audi, R. (1993a). *Action, intention, and reason.* Ithaca, NY: Cornell University Press.

Audi, R. (1993b). *The structure of justification.* Cambridge, England: Cambridge University Press.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

Bar-Hillel, M. (1990). Back to base rates. In R. M. Hogarth (Ed.), *Insights into decision making: A tribute to Hillel J. Einhorn* (pp. 200–216). Chicago: University of Chicago Press.

Baron, J. (1985). *Rationality and intelligence.* Cambridge: Cambridge University Press.

Baron, J. (1988). *Thinking and deciding.* Cambridge, England: Cambridge University Press.

Baron, J. (1991a). Beliefs about thinking. In J. Voss, D. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 169–186). Hillsdale, NJ: Erlbaum.

Baron, J. (1991b). Some thinking is irrational. *Behavioral and Brain Sciences, 14,* 486–487.

Baron, J. (1993a). *Morality and rational choice.* Dordrecht, The Netherlands: Kluwer.

Baron, J. (1993b). Why teach thinking?—An essay. *Applied Psychology: An International Review, 42,* 191–214.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences, 17,* 1–42.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning, 1,* 221–235.

Baron, J. (1996). Situated cognition, prescriptive theory, evolution, and something. *Behavioral and Brain Sciences, 19,* 324–326.

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

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *Quarterly Journal of Experimental Psychology, 40A,* 269–297.

Bell, D., Raiffa, H., & Tversky, A. (Eds.). (1988). *Decision making: Descriptive, normative, and prescriptive interactions.* Cambridge, England: Cambridge University Press.

Berkeley, D., & Humphreys, P. (1982). Structuring decision problems and the "bias heuristic." *Acta Psychologica, 50,* 201–252.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology, 96,* 85–94.

Block, J., & Funder, D. C. (1986). Social roles and social perception: Individual differences in attribution and "error." *Journal of Personality and Social Psychology, 51,* 1200–1207.

Braine, M. D. S., Connell, J., Freitag, J., & O'Brien, D. P. (1990). Is the base rate fallacy an instance of asserting the consequent? In K. Gilhooly, M. Keane, R. Logie, & G. Erdos (Eds.), *Lines of thinking* (Vol. 1, pp. 165–180). New York: John Wiley.

Bratman, M. E., Israel, D. J., & Pollack, M. E. (1991). Plans and resource-bounded practical reasoning. In J. Cummins & J. Pollock (Eds.), *Philosophy and AI: Essays at the interface* (pp. 7–22). Cambridge, MA: MIT Press.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65,* 212–219.

Brodzinsky, D. M. (1985). On the relationship between cognitive styles and cognitive structures. In E. D. Neimark, R. DeLisi, & J. L. Newman (Eds.), *Moderators of competence* (pp. 147–174). Hillsdale, NJ: Erlbaum.

Broome, J. (1990). Should a rational agent maximize expected utility? In K. S. Cook & M. Levi (Eds.), *The limits of rationality* (pp. 132–145). Chicago: University of Chicago Press.

Brown, J., Bennett, J., & Hanna, G. (1981). *The Nelson-Denny Reading Test.* Lombard, IL: Riverside Publishing Co.

Cahan, S., & Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development, 60,* 1239–1249.

Campbell, D. T. (1987). Evolutionary epistemology. In G. Radnitzky & W. W. Bartley (Eds.), *Evolutionary epistemology, rationality, and the sociology of knowledge* (pp. 47–89). La Salle, IL: Open Court.

Campbell, J. D. (1986). Similarity and uniqueness: The effects of attribute type, relevance, and individual differences in self-esteem and depression. *Journal of Personality and Social Psychology, 50,* 281–294.

Campbell, J. D., & Tesser, A. (1983). Motivational interpretation of hindsight bias: An individual difference analysis. *Journal of Personality, 51,* 605–640.

Capon, N., & Kuhn, D. (1979). Logical reasoning in the supermarket: Adult females' use of a proportional reasoning strategy in an everyday context. *Developmental Psychology, 15,* 450–452.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97,* 404–431.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press.

Caryl, P. G. (1994). Early event-related potentials correlate with inspection time and intelligence. *Intelligence, 18,* 15–46.

Casscells, W., Schoenberger, A., & Graboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine, 299,* 999–1001.

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

Cherniak, C. (1986). *Minimal rationality.* Cambridge, MA: MIT Press.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, L. J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition, 7,* 385–407.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4,* 317–370.

Cohen, L. J. (1982). Are people programmed to commit fallacies? Further thoughts about the interpretation of experimental data on probability judgment. *Journal for the Theory of Social Behavior, 12,* 251–274.

Cohen, L. J. (1986). *The dialogue of reason.* Oxford, England: Oxford University Press.

Cooper, W. S. (1987). Decision theory as a branch of evolutionary theory: A biological derivation of the Savage axioms. *Psychological Review, 94,* 395–411.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31,* 187–276.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition, 50,* 41–77.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1–73.

Davidson, D. (1980). *Essays on actions & events.* Oxford, England: Oxford University Press.

Dawes, R. M. (1988). *Rational choice in an uncertain world.* San Diego, CA: Harcourt Brace Jovanovich.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25,* 1–17.

Dawes, R. M. (1990). The potential nonfalsity of the false consensus effect. In R. M. Hogarth (Ed.), *Insights into decision making* (pp. 179–199). Chicago: University of Chicago Press.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association, 77,* 605–613.

Deary, I. J. (1995). Auditory inspection time and intelligence: What is the direction of causation? *Developmental Psychology, 31,* 237–250.

Deary, I. J., & Stough, C. (1996). Intelligence and inspection time. *American Psychologist, 51,* 599–608.

de Finetti, B. (1989). Probabilism. *Erkenntnis, 31,* 169–223. (Original work published 1931)

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66,* 819–829.

Dennett, D. (1980). The milk of human intentionality. *Behavioral and Brain Sciences, 3,* 428–430.

Dennett, D. (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Denny, J. P. (1991). Rational thought in oral culture and literate decontextualization. In D. R. Olson & N. Torrance (Eds.), *Literacy and orality* (pp. 66–89). Cambridge, England: Cambridge University Press.

de Sousa, R. (1987). *The rationality of emotion.* Cambridge, MA: MIT Press.

Detterman, D. K. (1994). Intelligence and the brain. In P. A. Vernon (Ed.), *The neuropsychology of individual differences* (pp. 35–57). San Diego, CA: Academic Press.

Donaldson, M. (1978). *Children's minds.* London: Fontana Paperbacks.

Donaldson, M. (1993). *Human minds: An exploration.* New York: Viking Penguin.

Dougherty, T. M., & Haith, M. M. (1997). Infant expectations and reaction time as predictors of childhood speed of processing and IQ. *Developmental Psychology, 33,* 146–155.

Earman, J. (1992). *Bayes or bust.* Cambridge, MA: MIT Press.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9–26). New York: Freeman.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49,* 709–724.

Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems. *Journal of Personality and Social Psychology, 62,* 328–339.

Erwin, T. D. (1981). *Manual for the Scale of Intellectual Development.* Harrisonburg, VA: Developmental Analytics.

Erwin, T. D. (1983). The Scale of Intellectual Development: Measuring Perry's scheme. *Journal of College Student Personnel, 24,* 6–12.

Estes, W. K. (1982). Learning, memory, and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 170–224). Cambridge, England: Cambridge University Press.

Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology, 75,* 451–468.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* London: Erlbaum Associates.

Evans, J. St. B. T., Barston, J., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11,* 295–306.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hove, England: Erlbaum.

Evans, J. St. B. T., Over, D. E., & Manktelow, K. (1993). Reasoning, decision making and rationality. *Cognition, 49,* 165–187.

Facione, P. (1992). *California Critical Thinking Dispositions Inventory.* La Cruz, CA: California Academic Press.

Farris, H., & Revlin, R. (1989). Sensible reasoning in two tasks: Rule discovery and hypothesis evaluation. *Memory & Cognition, 17,* 221–232.

Fetzer, J. (1990). Evolution, rationality, and testability. *Synthese, 82,* 423–439.

Finocchiaro, M. A. (1980). *Galileo and the art of reasoning.* Dordrecht, The Netherlands: D. Reidel.

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

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

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, England: Cambridge University Press.

Fischhoff, B. (1988). Judgment and decision making. In R. J. Sternberg & E. E. Smith (Eds.), *The psychology of human thought* (pp. 153–187). Cambridge, England: Cambridge University Press.

Foley, R. (1987). *The theory of epistemic rationality.* Cambridge, MA: Harvard University Press.

Foley, R. (1988). Some different conceptions of rationality. In E. McMullin (Eds.), *Construction and constraint* (pp. 123–152). Notre Dame, IN: Notre Dame University Press.

Foley, R. (1991). Rationality, belief, and commitment. *Synthese, 89,* 365–392.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253–292.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence. *Psychological Science, 7,* 237–241.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75–90.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality & Individual Differences, 7,* 385–400.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General, 115,* 16–25.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98,* 254–267.

Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 284–313). London: Routledge.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103,* 592–596.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Gilovich, T. (1991). *How we know what isn't so.* New York: Free Press.

Globerson, T. (1989). What is the relationship between cognitive style and cognitive development? In T. Globerson & T. Zelniker (Eds.), *Cognitive style and cognitive development* (pp. 71–85). Norwood, NJ: Ablex.

Goldman, A. I. (1978). Epistemics: The regulative theory of cognition. *Journal of Philosophy, 55,* 509–523.

Goldman, A. I. (1986). *Epistemology and cognition.* Cambridge, MA: Harvard University Press.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist, 35,* 603–618.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

Griffiths, M. D. (1994). The role of cognitive bias and skill in fruit machine gambling. *British Journal of Psychology, 85,* 351–369.

Harman, G. (1986). *Change in view.* Cambridge, MA: MIT Press.

Harman, G. (1995). Rationality. In E. E. Smith & D. N. Osherson (Eds.), *Thinking* (Vol. 3, pp. 175–211). Cambridge, MA: MIT Press.

Haslam, N., & Jayasinghe, N. (1995). Negative affect and hindsight bias. *Journal of Behavioral Decision Making, 8,* 127–135.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107,* 311–327.

Hays, R., Hayashi, T., & Stewart, A. (1989). A five-item measure of socially desirable response set. *Educational and Psychological Measurement, 49,* 629–636.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118,* 248–271.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53,* 221–234.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach.* Chicago: Open Court.

Hunt, E. (1978). Mechanics of verbal ability. *Psychological Review, 85,* 109–130.

Hunt, E. (1987). The next word on verbal ability. In P. A. Vernon (Ed.), *Speed of information-processing and intelligence* (pp. 347–392). Norwood, NJ: Ablex.

Jackson, S. I., & Griggs, R. A. (1988). Education and the selection task. *Bulletin of the Psychonomic Society, 26,* 327–330.

Jepson, C., Krantz, D., & Nisbett, R. (1983). Inductive reasoning: Competence or skill? *Behavioral and Brain Sciences, 6,* 494–501.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction.* Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., & Byrne, R. M. J. (1993). Models and deductive rationality. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 177–210). London: Routledge.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1994). Why models rather than rules give a better account of propositional reasoning. *Psychological Review, 101,* 734–739.

Johnson-Laird, P., & Oatley, K. (1992). Basic emotions, rationality, and folk theory. *Cognition and Emotion, 6,* 201–223.

Johnson-Laird, P. N., & Shafir, E. (1993). The interaction between reasoning and decision making: An introduction. *Cognition, 49,* 1–9.

Jones, W., Russell, D., & Nickel, T. (1977). Belief in the paranormal scale: An objective instrument to measure belief in magical phenomena and causes. *JSAS Catalog of Selected Documents in Psychology, 7*(100), Manuscript No. 1577, 1–32.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226–246.

Juslin, P., Winman, A., & Persson, T. (1994). Can overconfidence be used as an indicator of reconstructive rather than retrieval processes? *Cognition, 54,* 99–130.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American, 246,* 160–173.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103,* 582–591.

Kao, S. F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective çell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1363–1386.

Kitcher, P. (1993). *The advancement of science.* New York: Oxford University Press.

Klaczynski, P. A., & Gordon, D. H. (1996). Self-serving influences on adolescents' evaluations of belief-relevant evidence. *Journal of Experimental Child Psychology, 62,* 317–339.

Klaczynski, P. A., & Laipple, J. (1993). Role of content domain, logic training, and IQ in rule acquisition and transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 653–672.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53.

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

Kornblith, H. (1993). *Inductive inference and its natural ground.* Cambridge, MA: MIT University Press.

Kramer, D. A., Kahlbaugh, P., & Goldston, R. (1992). A measure of paradigm beliefs about the social world. *Journal of Gerontology: Psychological Sciences, 47,* P180–P189.

Krueger, J., & Zeiger, J. (1993). Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology, 65,* 670–680.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing the mind: "Seizing" and "freezing." *Psychological Review, 103,* 263–283.

Kuhn, D. (1991). *The skills of argument.* Cambridge, England: Cambridge University Press.

Kuhn, D. (1993). Connecting scientific and informal reasoning. *Merrill-Palmer Quarterly, 38,* 74–103.

Kyburg, H. E. (1983). Rational belief. *Behavioral and Brain Sciences, 6,* 231–273.

Lad, F. (1984). The calibration question. *British Journal for the Philosophy of Science, 35,* 213–221.

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (1993). Focussing in reasoning and decision making. *Cognition, 49,* 37–66.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effect of graduate training on reasoning. *American Psychologist, 43,* 431–442.

Levelt, W. (1995). Chapters of psychology. In R. L. Solso & D. W. Massaro (Eds.), *The science of the mind: 2001 and beyond* (pp. 184–202). New York: Oxford University Press.

Levi, I. (1983). Who commits the base rate fallacy? *Behavioral and Brain Sciences, 6,* 502–506.

Liberman, N., & Klar, Y. (1996). Hypothesis testing in Wason's selection task: Social exchange cheating detection or task understanding. *Cognition, 58,* 127–156.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 149–171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration and probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Lopes, L. L., & Oden, G. C. (1991). The rationality of intelligence. In E. Eells & T. Maruszewski (Eds.), *Probability and rationality: Studies on L. Jonathan Cohen's philosophy of science* (pp. 199–223). Amsterdam: Editions Rodopi.

Lowe, E. J. (1993). Rationality, deduction and mental models. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 211–230). London: Routledge.

Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations.* Cambridge, MA: Harvard University Press.

Lyon, D., & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica, 40,* 287–298.

Macchi, L. (1995). Pragmatic aspects of the base-rate fallacy. *Quarterly Journal of Experimental Psychology, 48A,* 188–207.

Markovits, H., & Bouffard-Bouchard, T. (1992). The belief-bias effect in the reasoning: The development and activation of competence. *British Journal of Developmental Psychology, 10,* 269–284.

Markovits, H., & Nantel, G. (1989). The belief–bias effect in the production and evaluation of logical conclusions. *Memory & Cognition, 17,* 11–17.

Marks, G., & Miller, N. (1987). Ten years of research on the false consensus effect: An empirical and theoretical review. *Psychological Bulletin, 102,* 72–90.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore: Williams & Wilkins.

Miller, D. T., Turnbull, W., & McFarland, C. (1990). Counterfactual thinking and social perception: Thinking about what might have been. In M. P. Zanna (Eds.), *Advances in experimental social psychology* (pp. 305–331). San Diego: Academic Press.

Mullen, B., Atkins, J., Champion, D., Edwards, C., Hardy, D., Story, J., & Vanderklok, M. (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology, 21,* 262–283.

Neimark, E. D. (1981). Confounding with cognitive style factors: An artifact explanation for the apparent nonuniversal incidence of formal operations. In E. Sigel, D. M. Brodzinsky, & R. M. Golinkoff (Eds.), *New directions in Piagetian theory and practice* (pp. 177–189). Hillsdale, NJ: Erlbaum.

Neimark, E. D. (1985). Moderators of competence: Challenges to the universality of Piagetian theory. In E. D. Neimark, R. DeLisi, & J. L. Newman (Eds.), *Moderators of competence* (pp. 1–14). Hillsdale, NJ: Erlbaum.

Neimark, E. D. (1987). *Adventures in thinking.* San Diego, CA: Harcourt Brace Jovanovich.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18,* 87–127.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Newstead, S. E., & Evans, J. St. B. T. (Eds.). (1995). *Perspectives on thinking and reasoning.* Hove, England: Erlbaum.

Nickerson, R. (1987). Why teach thinking? In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 27–40). New York: Freeman.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning, 2,* 1–31.

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

Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking.* Pacific Grove, CA: Midwest Publications.

Nozick, R. (1993). *The nature of rationality.* Princeton, NJ: Princeton University Press.

Oaksford, M., & Chater, N. (1993). Reasoning theories and bounded rationality. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 31–60). London: Routledge.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631.

Oaksford, M., & Chater, N. (1995). Theories of reasoning and the computational explanation of everyday inference. *Thinking and Reasoning, 1,* 121–152.

Oatley, K. (1992). *Best laid schemes: The psychology of emotions.* Cambridge, England: Cambridge University Press.

O'Brien, D. P., Braine, M., & Yang, Y. (1994). Propositional reasoning by mental models? Simple to refute in principle and in practice. *Psychological Review, 101,* 711–724.

Olson, D. R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review, 47,* 257–281.

Olson, D. R. (1994). *The world on paper.* Cambridge, England: Cambridge University Press.

Overton, W. F. (1985). Scientific methodologies and the competence-moderator performance issue. In E. D. Neimark, R. DeLisi, & J. L. Newman (Eds.), *Moderators of competence* (pp. 15–41). Hillsdale, NJ: Erlbaum.

Overton, W. F. (1990). Competence and procedures: Constraints on the development of logical reasoning. In W. F. Overton (Ed.), *Reasoning, necessity, and logic* (pp. 1–32). Hillsdale, NJ: Erlbaum.

Overton, W. F., Byrnes, J. P., & O'Brien, D. P. (1985). Developmental and individual differences in conditional reasoning: The role of contradiction training and cognitive style. *Developmental Psychology, 21,* 692–701.

Overton, W. F., & Newman, J. L. (1982). Cognitive development: A competence-activation/utilization approach. In T. Field, A. Huston, H. Quay, L. Troll, & G. Finley (Eds.), *Review of human development* (pp. 217–241). New York: Wiley.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60,* 307–317.

Payne, J. W., Bettman, J. R., & Johnson, E. (1993). *The adaptive* *decision maker.* Cambridge, England: Cambridge University Press.

Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. Voss, D. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 83–105). Hillsdale, NJ: Erlbaum.

Perkins, D. N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly, 39,* 1–21.

Perry, W. G. (1970). *Forms of intellectual and ethical development in the college years: A scheme.* New York: Holt, Rinehart & Winston.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72,* 346–354.

Piaget, J. (1926). *The language and thought of the child.* London: Routledge & Kegan Paul.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15,* 1–12.

Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds.* New York: John Wiley.

Pollard, P., & Evans, J. S. B. T. (1987). Content and context effects in reasoning. *American Journal of Psychology, 100,* 41–60.

Pylyshyn, Z. (1984). *Computation and cognition.* Cambridge, MA: MIT Press.

Raven, J. C. (1962). *Advanced Progressive Matrices* (Set II). London: H. K. Lewis & Co.

Raven, J. C., Court, J. H., & Raven, J. (1977). *Manual for Advanced Progressive Matrices* (Sets I & II). London: H. K. Lewis & Co.

Rescher, N. (1988). *Rationality: A philosophical inquiry into the nature and rationale of reason.* Oxford, England: Oxford University Press.

Rips, L. J. (1994). *The logic of proof.* Cambridge, MA: MIT Press.

Roberts, M. J. (1993). Human reasoning: Deduction rules or mental models, or both? *Quarterly Journal of Experimental Psychology, 46A,* 569–589.

Rokeach, M. (1960). *The open and closed mind.* New York: Basic Books.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40,* 193–218.

Ross, L., Amabile, T., & Steinnetz, J. (1977). Social roles, social control, and biases in the social perception process. *Journal of Personality and Social Psychology, 35,* 485–494.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13,* 279–301.

Sanders, G. S., & Mullen, B. (1983). Accuracy in perceptions of consensus: Differential tendencies of people with majority and minority positions. *European Journal of Social Psychology, 13,* 57–70.

Schick, F. (1987). Rationality: A third dimension. *Economics and Philosophy, 3,* 49–66.

Schick, F. (1997). *Making choices: A recasting of decision theory.* Cambridge, England: Cambridge University Press.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

Schoemaker, P. (1991). The quest for optimality: A positive heuristic of science? *Behavioral and Brain Sciences, 14,* 205–245.

Schommer, M. (1993). Epistemological development and academic performance among secondary students. *Journal of Educational Psychology, 85,* 406–411.

Schommer, M. (1994). Synthesizing epistemological belief research: Tentative understandings and provocative confusions. *Educational Psychology Review, 6,* 293–319.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* Mahwah, NJ: Erlbaum.

Shafir, E. (1994). Uncertainty and the difficulty of thinking through disjunctions. *Cognition, 50,* 403–430.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24,* 449–474.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology, 48A,* 257–279.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28,* 1–39.

Spitz, H. H. (1979). Beyond field theory in the study of mental deficiency. In N. Ellis (Ed.), *Handbook of mental deficiency, psychological theory and research* (pp. 121–141). Hillsdale, NJ: Erlbaum.

Stankov, L., & Dunn, S. (1993). Physical substrata of mental energy: Brain capacity and efficiency of cerebral metabolism. *Learning and Individual Differences, 5,* 241–257.

Stanovich, K. E. (1994). Reconceptualizing intelligence: Dysrationalia as an intuition pump. *Educational Researcher, 23*(4), 11–22.

Stanovich, K. E. (1996). Decentered thought and consequentialist decision making. *Behavioral and Brain Sciences, 19,* 323–324.

Stanovich, K. E. (in press). *Who is rational? Studies of individual differences in reasoning.* Mahwah, NJ: Erlbaum.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89,* 342–357.

Stanovich, K. E., West, R. F., & Harrison, M. (1995). Knowledge growth and maintenance across the life span: The role of print exposure. *Developmental Psychology, 31,* 811–826.

Stein, E. (1996). *Without good reason: The rationality debate in philosophy and cognitive science.* Oxford, England: Oxford University Press.

Sterelny, K. (1990). *The representational theory of mind: An introduction.* Oxford, England: Basil Blackwell.

Stich, S. (1990). *The fragmentation of reason.* Cambridge, MA: MIT Press.

Straughn, C. S., & Straughn, B. L. (Eds.). (1995). *Lovejoy's college guide* (23rd ed.). New York: Simon & Schuster.

Sutherland, S. (1992). *Irrationality: The enemy within.* London: Constable.

Thagard, P. (1992). *Conceptual revolutions.* Princeton, NJ: Princeton University Press.

Thaler, R. H. (1992). *The winner's curse: Paradoxes and anomalies of economic life.* New York: Free Press.

Tobacyk, J., & Milford, G. (1983). Belief in paranormal phenomena. *Journal of Personality and Social Psychology, 44,* 1029–1037.

Troldahl, V., & Powell, F. (1965). A short-form dogmatism scale for use in field studies. *Social Forces, 44,* 211–215.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51,* 1–10.

Tversky, A. (1975). A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis, 9,* 163–173.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge, England: Cambridge University Press.

Vernon, P. A. (1991). The use of biological measures to estimate behavioral intelligence. *Educational Psychologist, 25,* 293–304.

Vernon, P. A. (1993). *Biological approaches to the study of human intelligence.* Norwood, NJ: Ablex.

Vernon, P. A., & Mori, M. (1992). Intelligence, reaction times, and peripheral nerve conduction velocity. *Intelligence, 16,* 273–288.

von Mises, R. (1957). *Probability, statistics, and truth.* New York: Dover.

Wagenaar, W. A. (1988). *Paradoxes of gambling behavior.* Hove, England: Erlbaum.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmonsworth, England: Penguin:

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 509–521.

West, R. F., & Stanovich, K. E. (1991). The incidental acquisition of information from reading. *Psychological Science, 2,* 325–330.

Wetherick, N. E. (1995). Reasoning and rationality: A critique of some experimental paradigms. *Theory & Psychology, 5,* 429–448.

Yaniv, I., Yates, F. J., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110,* 611–617.

Yates, J. F., Lee, J., & Shinotsuka, H. (1996). Beliefs about overconfidence, including its cross-national variation. *Organizational Behavior and Human Decision Processes, 65,* 138–147.

Yates, J. F., Zhu, Y., Ronis, D., Wang, D., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes, 43,* 145–171.

Zahler, D., & Zahler, K. (1988). *Test your cultural literacy.* New York: Simon & Schuster.

Zechmeister, E. B., & Johnson, J. (1992). *Critical thinking: A functional approach.* Pacific Grove: CA: Brooks/Cole.

Received July 23, 1996
Revision received February 4, 1997
Accepted July 10, 1997 ■