Conditions for Intuitive Expertise

A Failure to Disagree

Daniel Kahneman *Princeton University*
Gary Klein *Applied Research Associates*

This article reports on an effort to explore the differences between two approaches to intuition and expertise that are often viewed as conflicting: heuristics and biases (HB) and naturalistic decision making (NDM). Starting from the obvious fact that professional intuition is sometimes marvelous and sometimes flawed, the authors attempt to map the boundary conditions that separate true intuitive skill from overconfident and biased impressions. They conclude that evaluating the likely quality of an intuitive judgment requires an assessment of the predictability of the environment in which the judgment is made and of the individual's opportunity to learn the regularities of that environment. Subjective experience is not a reliable indicator of judgment accuracy.

Keywords: intuition, expertise, overconfidence, heuristics, judgment

In this article we report on an effort to compare our views on the issues of intuition and expertise and to discuss the evidence for our respective positions. When we launched this project, we expected to disagree on many issues, and with good reason: One of us (GK) has spent much of his career thinking about ways to promote reliance on expert intuition in executive decision making and identifies himself as a member of the intellectual community of scholars and practitioners who study naturalistic decision making (NDM). The other (DK) has spent much of his career running experiments in which intuitive judgment was commonly found to be flawed; he is identified with the "heuristics and biases" (HB) approach to the field.

A surprise awaited us when we got together to consider our joint field of interest. We found ourselves agreeing most of the time. Where we initially disagreed, we were usually able to converge upon a common position. Our shared beliefs are much more specific than the commonplace that expert intuition is sometimes remarkably accurate and sometimes off the mark. We accept the commonplace, of course, but we also have similar opinions about more specific questions: What are the activities in which skilled intuitive judgment develops with experience? What are the activities in which experience is more likely to produce overconfidence than genuine skill? Because we largely agree about the answers to these questions we also favor generally similar recommendations to organizations seeking to improve the quality of judgments and decisions. In spite of all this agreement, however, we find that we are still separated in many ways: by divergent attitudes, preferences about facts, and feelings about fighting words such as "bias." If we are to understand the differences between our respective communities, such emotions must be taken into account.

We begin with a brief review of the origins and precursors of the NDM and HB approaches, followed by a discussion of the most prominent points of contrast between them (NDM: Klein, Orasanu, Calderwood, & Zsambok, 1993; HB: Gilovich, Griffin, & Kahneman, 2002; Tversky & Kahneman, 1974). Next we present some claims about the conditions under which skilled intuitions develop, followed by several suggestions for ways to improve the quality of judgments and choices.

Two Perspectives

Origins of the Naturalistic Decision Making Approach

The NDM approach, which focuses on the successes of expert intuition, grew out of early research on master chess players conducted by deGroot (1946/1978) and later by Chase and Simon (1973). DeGroot showed that chess grand masters were generally able to identify the most promising moves rapidly, while mediocre chess players often did not even consider the best moves. The chess grand masters mainly differed from weaker players in their unusual ability to appreciate the dynamics of complex positions and quickly judge a line of play as promising or fruitless. Chase and Simon (1973) described the performance of chess experts as a form of perceptual skill in which complex patterns are recognized. They estimated that chess masters acquire a repertoire of 50,000 to 100,000 immediately recognizable patterns, and that this repertoire enables them to identify a good move without having to calculate all possible contingencies. Strong players need a decade of serious play to assemble this large collection of basic patterns, but of course they achieve impressive levels

Daniel Kahneman, Woodrow Wilson School of Public and International Affairs, Princeton University; Gary Klein, Applied Research Associates, Fairborn, Ohio.

We thank Craig Fox, Robin Hogarth, and James Shanteau for their helpful comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Daniel Kahneman, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ 08544-0001. E-mail: kahneman@princeton.edu

Vol. 64, No. 6, 515–526 DOI: 10.1037/a0016755

Daniel Kahneman

of skill even earlier. On the basis of this work, Simon defined intuition as the recognition of patterns stored in memory.

The early work that led to the approach that is now called NDM was an attempt to describe and analyze the decision making of commanders of firefighting companies. Fireground commanders are required to make decisions under conditions of uncertainty and time pressure that preclude any orderly effort to generate and evaluate sets of options. Klein, Calderwood, and Clinton-Cirocco (1986) investigated how the commanders could make good decisions without comparing options. The initial hypothesis was that commanders would restrict their analysis to only a pair of options, but that hypothesis proved to be incorrect. In fact, the commanders usually generated only a single option, and that was all they needed. They could draw on the repertoire of patterns that they had compiled during more than a decade of both real and virtual experience to identify a plausible option, which they considered first. They evaluated this option by mentally simulating it to see if it would work in the situation they were facing—a process that deGroot (1946/1978) had described as progressive deepening. If the course of action they were considering seemed appropriate, they would implement it. If it had shortcomings, they would modify it. If they could not easily modify it, they would turn to the next most plausible option and run through the same procedure until an acceptable course of action was found. This recognition-primed decision (RPD) strategy was effective because it took advantage of the commanders' tacit knowledge (Klein et al., 1986). The fireground commanders were able to draw on their repertoires to anticipate how flames were likely to spread through a building, to notice signs that a house was likely to collapse, to judge when to call for additional support, and to make many other critical decisions. The RPD model is consistent with the work of deGroot (1946/1978) and Simon (1992) and has been replicated in multiple domains, including system design, military command and control, and management of offshore oil installations (see Klein, 1998, for a review). In each of these domains, the RPD model offers a generally encouraging picture of expert performance. It would be a caricature of the NDM approach, however, to describe it as being solely dedicated to praising expertise. NDM researchers have also tried to document and analyze failures in the performance of experts (Cannon-Bowers & Salas, 1998; Klein, 1998; Woods, O'Brien, & Hanes, 1987). In fact, the NDM movement was crystallized by an event that resulted from a catastrophic failure in expert decision making.

In 1988, an international tragedy occurred after the USS *Vincennes* accidentally shot down an Iranian Airbus (Fogarty, 1988). The USS *Vincennes* was an Aegis cruiser, one of the most technologically advanced systems in the Navy inventory, but the technology was not sufficient to stave off the disaster. The incident has been the subject of detailed investigation by NDM researchers (Collyer & Malecki, 1998; Klein, 1998). As a result of the disastrous error and subsequent political fallout, the U.S. Navy decided to initiate a program of research on decision making, the Tactical Decision Making Under Stress (TADMUS) program (Cannon-Bowers & Salas, 1998).

Thus it was that in 1989 a group of 30 researchers who studied decision making in natural settings met for several days in an effort to find commonalities between the decision-making processes of firefighters, nuclear power plant controllers, Navy officers, Army officers, highway engineers, and other populations. Several researchers from the judgment and decision making tradition participated in this meeting and in the preparation of a book describing the NDM perspective (Klein et al., 1993). Lipshitz (1993) identified several decision-making models that were developed to describe the strategies used in field settings, including the recognition-primed decision model (Klein, 1993), the cognitive continuum model (Hammond, Hamm, Grassia, & Pearson, 1987), image theory (Beach, 1990), the search for dominance structures (Montgomery, 1993), and the skills/rules/knowledge framework and decision ladder (Rasmussen, 1986). The NDM movement that emerged from this meeting focuses on field studies of subject-matter experts who make decisions under complex conditions. These experts are expected to successfully attain vaguely defined goals in the face of uncertainty, time pressure, high stakes, team and organizational constraints, shifting conditions, and action feedback loops that enable people to manage disturbances while trying to diagnose them (Orasanu & Connolly, 1993).

A central goal of NDM is to demystify intuition by identifying the cues that experts use to make their judgments, even if those cues involve tacit knowledge and are difficult for the expert to articulate. In this way, NDM researchers try to learn from expert professionals. Many NDM researchers use cognitive task analysis (CTA) methods to investigate the cues and strategies that skilled deci-

Gary Klein

sion makers apply (Crandall, Klein, & Hoffman, 2006; Schraagen, Chipman, & Shalin, 2000). CTA methods are semi-structured interview techniques that elicit the cues and contextual considerations influencing judgments and decisions. Researchers cannot expect decision makers to accurately explain why they made decisions (Nisbett & Wilson, 1977); CTA methods provide a basis for making inferences about the judgment and decision process. For example, Crandall and Getchell-Reiter (1993) studied nurses in a neonatal intensive care unit (NICU) who could detect infants developing life-threatening infections even before blood tests came back positive. When asked, the nurses were at first unable to describe how they made their judgments. The researchers used CTA methods to probe specific incidents and identified a range of cues and patterns, some of which had not yet appeared in the nursing or medical literature. A few of these cues were opposite to the indicators of infection in adults. Crandall and Gamblian (1991) extended the NICU work. They confirmed the findings with nurses from a different hospital and then created an instructional program to help new NICU nurses learn how to identify the early signs of sepsis in neonates. That program has been widely disseminated throughout the nursing community.

Origins of the Heuristics and Biases Approach

In sharp contrast to NDM, the HB approach favors a skeptical attitude toward expertise and expert judgment. The origins of this attitude can be traced to a famous monograph published by Paul Meehl in 1954. Meehl (1954) reviewed approximately 20 studies that compared the accuracy of forecasts made by human judges (mostly clinical psychologists) and those predicted by simple statistical models. The criteria in the studies that Meehl (1954) discussed were diverse, with outcome measures ranging from academic success to patient recidivism and propensity for violence. Although the algorithms were based on a subset of the information available to the clinicians, statistical predictions were more accurate than human predictions in almost every case. Meehl (1954) believed that the inferiority of clinical judgment was due in part to systematic errors, such as the consistent neglect of the base rates of outcomes in discussion of individual cases. In a well-known article, he later explained his reluctance to attend clinical conferences by citing his annoyance with the clinicians' uncritical reliance on their intuition and their failure to apply elementary statistical reasoning (Meehl, 1973).

Inconsistency is a major weakness of informal judgment: When presented with the same case information on separate occasions, human judges often reach different conclusions. Goldberg (1970) reported a "bootstrapping effect," which provides the most dramatic illustration of the effect of inconsistency on the validity of judgments. Goldberg required a group of 29 clinicians to make diagnostic judgments (psychotic vs. neurotic) in a set of cases, based on personality test profiles of 861 patients who had been independently assigned to one of these categories. He constructed an individual model of the predictions of each judge—using multiple regression to estimate the weights that the judge assigned to each of the 11 scales in the Minnesota Multiphasic Personality Inventory. Judges were then required to make predictions for a new set of cases; Goldberg also used the individual statistical model of each judge to generate a prediction for these new cases. The bootstrap models were almost always more accurate than the judges they modeled. The only plausible explanation of this remarkable result is that human judgments are noisy to an extent that substantially impairs their validity. In an extensive meta-analysis of judgment studies using the lens model, Karelaia and Hogarth (2008) reported strong support for the generality of the bootstrap effect and for the crucial importance of lack of consistency in explaining this effect.

Kahneman read Meehl's book in 1955 while serving in the Psychological Research Unit of the Israel Defense Forces, and the book helped him make sense of his own encounters with the difficulties of clinical judgment. One of Kahneman's duties was to assess candidates for officer training, using field tests and other observations as well as a personal interview. Kahneman (2003) described the powerful sense of getting to know each candidate and the accompanying conviction that he could foretell how well the candidate would do in further training and eventually in combat. The subjective conviction of understanding each case in isolation was not diminished by the statistical feedback from officer training school, which indicated that the validity of the assessments was negligible. Kahneman coined the term *illusion of validity* for the unjustified sense of confidence that often comes with clinical judgment. His early experience with the fallibility of intuitive impressions could hardly be more different from Klein's formative encounter with the successful decision making of fireground commanders.

The first study in the HB tradition was conducted in 1969 (Tversky & Kahneman, 1971). It described performance in a task that researchers often perform without recourse to computation: choosing the number of cases for a psychological experiment. The participants in the study were sophisticated methodologists and statisticians, including two authors of statistics textbooks. They answered realistic questions about the sample size they considered appropriate in different situations. The conclusion of the study was that sophisticated scientists reached incorrect conclusions and made inferior choices when they followed their intuitions, failing to apply rules with which they were certainly familiar. The article offered a strongly worded recommendation that researchers faced with the task of choosing a sample size should forsake intuition in favor of computation. This initial study of professionals reinforced Tversky and Kahneman (1971) in their belief (originally based on introspection) that faulty statistical intuitions survive both formal training and actual experience. Many studies in the intervening decades have confirmed the persistence of a diverse set of intuitive errors in the judgments of some professionals.

Contrasts Between the Naturalistic Decision Making and Heuristics and Biases Approaches

The intellectual traditions that we have traced to deGroot's (1946/1978) studies of chess masters (NDM) and to Meehl's (1954) research on clinicians (HB) are alive and well today. They are reflected in the approaches of our respective intellectual communities. In this section we consider three important contrasts between the two approaches: the stance taken by the NDM and HB researchers toward expert judgment, the use of field versus laboratory settings for decision-making research, and the application of different standards of performance, which leads to different conclusions about expertise.

Stance Regarding Expertise and Decision Algorithms

There is no logical inconsistency between the observations that inspired the NDM and HB approaches to professional judgment: The intuitive judgments of some professionals are impressively skilled, while the judgments of other professionals are remarkably flawed. Although not contradictory, these core observations suggest conflicting generalizations about the utility of expert judgment. Members of the HB community are of course aware of the existence of skill and expertise, but they tend to focus on flaws in human cognitive performance. Members of the NDM community know that professionals often err, but they tend to stress the marvels of successful expert performance.

The basic stance of HB researchers, as they consider experts, is one of skepticism. They are trained to look for opportunities to compare expert performance with performance by formal models or rules and to expect that experts will do poorly in such comparisons. They are predisposed to recommend the replacement of informal judgment by algorithms whenever possible. Researchers in the NDM tradition are more likely to adopt an admiring stance toward experts. They are trained to explore the thinking of experts, hoping to identify critical features of the situation that are obvious to experts but invisible to novices and journeymen, and then to search for ways to pass on the experts' secrets to others in the field. NDM researchers are disposed to have little faith in formal approaches because they are generally skeptical about attempts to impose universal structures and rules on judgments and choices that will be made in complex contexts.

We found that the sharpest differences between the two of us were emotional rather than intellectual. Although DK is thrilled by the remarkable intuitive skills of experts that GK and others have described, he also takes considerable pleasure in demonstrations of human folly and in the comeuppance of overconfident pseudo-experts. For his part, GK recognizes that formal procedures and algorithms sometimes outdo human judgment, but he enjoys hearing about cases in which the bureaucratization of decision making fails. Further, the nonoverlapping sets of colleagues with whom we interact generally share our attitudes and reinforce our differences. Nevertheless, as this article shows, we agree on most of the issues that matter.

Field Versus Laboratory

There is an obvious difference in the primary form of research conducted by the respective research communities. The members of the HB community are mostly based in academic departments, and they tend to favor well-controlled experiments in the laboratory. The members of the NDM community are typically practitioners who operate in "real-world" organizations. They have a natural sympathy for the ecological approach, first popularized in the late 1970s, which questions the relevance of laboratory experiments to real-world situations. NDM researchers use methods such as cognitive task analysis and field observation to investigate judgments and decision making under complex conditions that would be difficult to recreate in the laboratory.

There is no logically necessary connection between these methodological choices and the nature of the hypotheses and models being tested. As the examples of the preceding section illustrate, the view that heuristics and biases are only studied and found in the laboratory is a caricature.[1] Similarly, the RPD model could have emerged from the laboratory, and it has been tested there (Johnson & Raab, 2003; Klein, Wolf, Militello, & Zsambok, 1995). In addition, a number of NDM researchers have reported studies of the performance of proficient decision makers in realistically simulated environments (e.g., Smith, Giffin, Rockwell, & Thomas, 1986).

[1] Among many other examples, see Slovic (2000) for applications to the study of responses to risk; Guthrie, Rachlinski, and Wistrich (2007) and Sunstein (2000) for applications in the legal domain; Croskerry and Norman (2008) for medical judgment; Bazerman (2005) for managerial judgments and decision making; and Kahneman and Renshon (2007) for political decision making. The collection assembled by Gilovich, Griffin, and Kahneman (2002) includes other examples.

The Definition of Expertise

NDM researchers cannot use the same kinds of optimality criteria as the HB community to define expertise. In rare cases (e.g., the ratings of chess players based on their record of wins and losses against other rated players) the performance level of experts is determined using standardized measures. However, in most of the situations studied by NDM researchers, the criteria for judging expertise are based on a history of successful outcomes rather than on quantitative performance measures. The most common method for defining expertise in NDM research is to rely on peer judgments. The conditions for defining expertise are the existence of a consensus and evidence that the consensus reflects aspects of successful performance that are objective even if they are not quantified explicitly. If the performance of different professionals can be compared, the best practitioners define the standard. As Shanteau (1992) suggested, "Experts are operationally defined as those who have been recognized within their profession as having the necessary skills and abilities to perform at the highest level" (p. 255). For example, captains of firefighting companies are evaluated not only by their ability to extinguish fires, but also by other criteria, such as the amount of damage created before the fire is controlled. When colleagues say, "If Person X had been there instead of Person Y, the fire would not have spread as far," then Person X counts as an expert within that organization. The use of peer judgments can distinguish highly competent decision makers from mediocre ones who may have the same amount of experience and from novices who have little experience. This level of differentiation is sufficient for most NDM studies.

In several of the studies that Meehl (1954) reviewed, the quality of expert performance was evaluated by comparing the accuracy of decisions made by experts with the accuracy of optimal linear combinations. If the predictions generated by a linear combination of a few variables are more accurate (in a new sample) than those of a professional who has access to the same information, the performance of the professional is certainly suboptimal. Note that the optimality criterion is significantly more demanding than the criteria by which expertise is evaluated in NDM research. NDM researchers compare the performance of professionals with that of the most successful experts in their field, whereas HB researchers prefer to compare the judgments of professionals with the outcome of a model that makes the best possible use of available information. It is entirely possible for the predictions of experienced clinicians to be superior to those of novices but inferior to a linear model or an intelligent system.

Sources of Intuition

The judgments and decisions that we are most likely to call intuitive come to mind on their own, without explicit awareness of the evoking cues and of course without an explicit evaluation of the validity of these cues. The firefighter feels that the house is very dangerous, the nurse feels that an infant is ill, and the chess master immediately sees a promising move. Intuitive skills are not restricted to professionals: Anyone can recognize tension or fatigue in a familiar voice on the phone. In the language of the two-system (or dual process) models that have recently become popular (Evans & Frankish, 2009; see Evans, 2007, for a review of the origins of these ideas), intuitive judgments are produced by "System 1 operations," which are automatic, involuntary, and almost effortless. In contrast, the deliberate activities of System 2 are controlled, voluntary, and effortful—they impose demands on limited attentional resources. System 2 is involved, for example, when one performs a calculation ($17 \times 24 = ?$), completes a tax form, reads a map, makes a left turn into heavy traffic, or parks in a narrow space. Self-monitoring is also a System 2 operation, which is impaired by concurrent effortful tasks.

The distinction between Systems 1 and 2 plays an important role in both the HB and NDM approaches. In the RPD model, for example, the performance of experts involves both an automatic process that brings promising solutions to mind and a deliberate activity in which the execution of the candidate solution is mentally simulated in a process of progressive deepening. In the HB approach, System 2 is involved in the effortful performance of some reasoning and decision-making tasks as well as in the continuous monitoring of the quality of reasoning. When there are cues that an intuitive judgment could be wrong, System 2 can impose a different strategy, replacing intuition by careful reasoning.[2]

The NDM and HB approaches share the assumption that intuitive judgments and preferences have the characteristics of System 1 activity: They are automatic, arise effortlessly, and often come to mind without immediate justification. However, the two approaches focus on different classes of intuition. Intuitive judgments that arise from experience and manifest skill are the province of NDM, which explores the cues that guided such judgments and the conditions for the acquisition of skill. In contrast, HB researchers have been mainly concerned with intuitive judgments that arise from simplifying heuristics, not from specific experience. These intuitive judgments are less likely to be accurate and are prone to systematic biases.

We discuss the two classes of judgment in sequence. First, we describe the process of skill acquisition that supports the intuitive judgments and preferences of genuine experts. In particular, we explore two necessary conditions for the development of skill: high-validity environments and an adequate opportunity to learn them. Next, we discuss heuristic-based intuitions and some of the biases to which they are prone. Finally, we address the question of the critique of intuition: How can skilled intuitions be distinguished from heuristic-based intuitions?

[2] The contrast between System 1 and System 2 has given rise to its own literature. For example, J. St. B. T. Evans (2007) has asserted that System 1 is affected by the tendency to contextualize problems in the light of prior knowledge and belief and that System 2 is affected by the tendency to satisfice without considering alternatives.

Skilled Intuition as Recognition

Simon (1992) offered a concise definition of skilled intuition that we both endorse: "The situation has provided a cue: This cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition" (p. 155). The model of intuition as recognition is helpful in several ways. First, it demystifies intuition. Many experts who have intuitions (and some authors who study them) endow intuition with an almost magic aura—knowledge that is not acquired by a rational process. In Simon's definition, the process by which the pediatric nurse recognizes that an infant may be gravely ill is not different in principle from the process by which she would notice that a friend looks tired or angry or from the way in which a small child recognizes that an animal is a dog, not a cat. It may be worth noting that this description of pattern recognition and the skilled pattern recognition described in the RPD model are different from the recognition heuristic discussed by Goldstein and Gigerenzer (1999), which is a special-purpose rule of thumb.

The recognition model implies two conditions that must be satisfied for an intuitive judgment (recognition) to be genuinely skilled: First, the environment must provide adequately valid cues to the nature of the situation. Second, people must have an opportunity to learn the relevant cues. For the first condition, valid cues must be specifiable, at least in principle—even if the individual does not know what they are. The child relies on valid cues to identify a dog, without any ability to state what the cues are. Similarly, the nurse and the firefighter are also guided by valid cues they find in the environment. No magic is involved. A crucial conclusion emerges: Skilled intuitions will only develop in an environment of sufficient regularity, which provides valid cues to the situation. The ways in which skilled judgments take advantage of environmental regularities have been discussed by, among others, Brunswik (1957) and Hertwig, Hoffrage, and Martingnon (1999).

Validity, as we use the term, describes the causal and statistical structure of the relevant environment. For example, it is very likely that there are early indications that a building is about to collapse in a fire or that an infant will soon show obvious symptoms of infection. On the other hand, it is unlikely that there is publicly available information that could be used to predict how well a particular stock will do—if such valid information existed, the price of the stock would already reflect it. Thus, we have more reason to trust the intuition of an experienced fireground commander about the stability of a building, or the intuitions of a nurse about an infant, than to trust the intuitions of a trader about a stock. We can confidently expect that a detailed study of how professionals think is more likely to reveal useful predictive cues in the former cases than in the latter.

Determining the validity of an environment is not always easy. When Tetlock (2005) embarked on his ambitious study of long-term forecasts of strategic and economic events by experts, the outcome of his research was not obvious. Fifteen years later it was quite clear that the highly educated and experienced experts that he studied were not superior to untrained readers of newspapers in their ability to make accurate long-term forecasts of political events. The depressing consistency of the experts' failure to outdo the novices in this task suggests that the problem is in the environment: Long-term forecasting must fail because large-scale historical developments are too complex to be forecast. The task is simply impossible. A thought experiment can help. Consider what the history of the 20th century might have been if the three fertilized eggs that became Hitler, Stalin, and Mao had been female. The century would surely have been very different, but can one know how?

In other environments, the regularities that can be observed are misleading. Hogarth (2001) introduced the useful notion of wicked environments, in which wrong intuitions are likely to develop. His most compelling example (borrowed from Lewis Thomas) is the early 20th-century physician who frequently had intuitions about patients in the ward who were about to develop typhoid. He confirmed his intuitions by palpating these patients' tongues, but because he did not wash his hands the intuitions were disastrously self-fulfilling.

High validity does not imply the absence of uncertainty, and the regularities that are to be discovered are sometimes statistical. Games such as bridge or poker count as high-validity situations. The mark of these situations is that skill, the ability to identify favorable bets, improves without guaranteeing that every attempt will succeed. The challenge of learning bridge and poker is not essentially different from the challenge of learning chess, where the uncertainty arises from the enormous number of possible developments.

As the examples of competitive games illustrate, the second necessary condition for the development of recognition (and of skilled intuition) is an adequate opportunity to learn the relevant cues. It has been estimated that chess masters must invest 10,000 hours to acquire their skills (Chase & Simon, 1973). Fortunately, most of the skills can be acquired with less practice. A child does not need thousands of examples to learn to discriminate dogs from cats. The skilled pediatric nurse has seen a sufficient number of sick infants to recognize subtle signs of disease, and the experienced fireground commander has experienced numerous fires and probably imagined many more, during years of thinking and conversing about firefighting. Without these opportunities to learn, a valid intuition can only be due to a lucky accident or to magic—and we do not believe in magic.

Two conditions must be satisfied for skilled intuition to develop: an environment of sufficiently high validity and adequate opportunity to practice the skill. Ericsson, Charness, Hoffman, and Feltovich (2006) have described a range of factors that influence the rate of skill development. These include the type of practice people employ, their level of engagement and motivation, and the self-regulatory processes they use. Even when the circumstances are favorable, however, some people will develop skilled in-

tuitions more quickly than others. Talent surely matters. Every normal child can recognize a cat or a dog, but not all dedicated chess players become grand masters. Extraordinary players such as Fischer and Kasparov were able to recognize patterns that other grand masters could not see on their own—although the weaker players could recognize the validity of the star's intuition when led through it.

Intuitions that are available only to a few exceptional individuals are often called creative. Like other intuitions, however, creative intuitions are based on finding valid patterns in memory, a task that some people perform much better than others. There are large individual differences in performance on the Remote Associations Test (RAT), which has a long history as a test of creativity. Participants in that test are instructed to search for a common associate of three words. The task has a wide range of difficulty: The item cottage/swiss/cake is easy, but few people can quickly find the answer to the item dive/light/rocket—although everyone recognizes the answer as valid (it is above us and is blue in good weather; Mednick, 1962). The RAT brings us back to Simon's observation that the regularities on which intuitions depend are represented in memory. The situation of the RAT has high validity: Widely shared patterns of associations exist, which everyone can recognize although few can find them without prompting.

Imperfect Intuition

We have seen that reliably skilled intuitions are likely to develop when the individual operates in a high-validity environment and has an opportunity to learn the rules of that environment. These conditions often remain unmet in professional contexts, either because the environment is insufficiently predictable (as in the long-term forecasting of political events) or because of the absence of opportunities to learn its rules (as in the case of firefighters exposed to a fire in a skyscraper with unexpected damage to the heat shielding of its structural support). We both agree that most of the intuitive judgments and decisions that System 1 produces are skilled, appropriate, and eventually successful. But we also agree that not all intuitive judgments are skilled, although our hunches about the frequency of exceptions differ. People, including experienced professionals, sometimes have subjectively compelling intuitions even when they lack true skill, either because the environment is insufficiently regular or because they have not mastered it. Lewis (2003) described the weaknesses in the ability of baseball scouts and managers to judge the capabilities, contributions, and potential of players. Despite ample opportunities to acquire judgment skill, scouts and managers were often insensitive to important variables and overly influenced by such factors as the player's appearance—a clear case of prediction by representativeness.

When intuitive judgments do not come from skill, where do they come from? This is the question that students of heuristics and biases have explored, mostly in laboratory experiments. The answer, of course, is that incorrect intuitions, like valid ones, also arise from the operations of memory. Three phenomena that have been discussed in the HB literature illustrate the sources of flawed intuitive judgments.

Frederick (2005) has studied problems such as the following: "A ball and a bat together cost $1.10. The bat costs a dollar more than the ball. How much does the ball cost?" The question invariably evokes an immediate tentative solution: 10 cents. But the intuitive response is wrong in this problem: The correct response is 5 cents. Furthermore, an easy check will quickly show that the answer is wrong: If the ball is worth 10 cents, then the bat is worth $1.10 and the total is $1.20, which is not correct. The surprising finding of Frederick's research is that many intelligent people adopt the intuitively compelling response without checking it. The incidence of intuitive errors in this question ranges from approximately 50% in top undergraduate schools (MIT, Princeton, Harvard) to 90% in somewhat less selective schools. It can be argued that the setting of this problem is not typical of the challenges that people face in the real world, but the phenomenon that Frederick studied is hardly restricted to puzzles. A common genre of business literature celebrates successful leaders who made strategic decisions on the basis of gut feelings and intuitions that they did not adequately check, but many of these successes owe more to luck than to genius (Rosenzweig, 2007).

The anchoring phenomenon is another case in which a bias in the operations of memory causes intuitions to go astray. Suppose some participants in an experiment are first asked "Is the average price of German cars more or less than $100,000?" before they are required to provide a numerical estimate of the average cost of German cars. Other respondents encounter a different anchoring question: They are first asked whether the average cost of German cars is more or less than $30,000, and then they are to give an estimate of the average. We can expect the estimates of the two groups to differ by as much as half the difference between the anchors—in this case the expected anchoring effect would be $35,000 (Jacowitz & Kahneman, 1995). The mechanism of anchoring is well understood (Mussweiler & Strack, 2000). The original question with the high anchor brings expensive cars to the respondents' mind: Mercedes, BMWs, Audis. The lower anchor is more likely to evoke the image of a beetle and the name Volkswagen. The initial question therefore biases the sample of cars that come to mind when people next attempt to estimate the average price of German cars. The process of estimating the average is a deliberate, System 2 operation, but the bias occurs in the automatic phase in which instances are retrieved from memory. The resulting anchoring effect is large and robust. The answers that come to mind are typically held with substantial confidence, and the victims of anchoring manipulations confidently deny any effect of the anchor. The common criticism of laboratory experiments hardly applies here, because large anchoring effects have been demonstrated in the courtroom, in real estate transactions, and in other real-world contexts.

For a final example, consider this question: "Julie is a graduating senior. She read fluently at age 4. What is your best guess of her GPA [grade point average]?" Most people

who think about this question report having an immediate intuitive impression of the best-fitting GPA. The value that comes to their mind is a GPA that is as impressive as Julie's precocity in reading—roughly a match of percentile scores. This intuitive prediction is clearly wrong because it is not regressive. The correlation between early reading and graduating GPA is not high and certainly does not justify nonregressive matching. The process that generates this intuitive answer has been called attribute substitution. The attribute that is to be assessed is GPA, but the answer is simply a projection onto the GPA scale of an evaluation of reading precocity. Attribute substitution has been described as an automatic process. It produces intuitive judgments in which a difficult question is answered by substituting an easier one—the essence of heuristic thinking (Kahneman & Frederick, 2002).

Of course, the mechanisms that produce incorrect intuitions will only operate in the absence of skill. If people have a skilled response to the task with which they are charged, they will apply their skill. But even in the absence of skill an intuitive response may come to their minds. The difficulty is that people have no way to know where their intuitions came from. There is no subjective marker that distinguishes correct intuitions from intuitions that are produced by highly imperfect heuristics. An important characteristic of intuitive judgments, which they share with perceptual impressions, is that a single response initially comes to mind. Most of the time we have to trust this first impulse, and most of the time we are right or are able to make the necessary corrections if we turn out to be wrong, but high subjective confidence is not a good indication of validity (Einhorn & Hogarth, 1978). Checking one's intuition is an effortful operation of System 2, which people do not always perform—sometimes because it is difficult to do so and sometimes because they do not bother.

Intuitions that originate in heuristics are not necessarily wrong. Indeed, the original statement of the HB approach asserted, "In general these heuristics are quite useful, but sometimes they lead to severe and systematic errors" (Tversky & Kahneman, 1974, p. 1124). The HB claim is not that intuitions that arise in heuristics are always incorrect, only that they are less trustworthy than intuitions that are rooted in specific experiences. Unfortunately, people are not normally aware of the origins of the thoughts that come to their minds, and the correlation between the accuracy of their judgments and the confidence they experience is not consistently high (Arkes, 2001; Griffin & Tversky, 1992). Subjective confidence is often determined by the internal consistency of the information on which a judgment is based, rather than by the quality of that information (Einhorn & Hogarth, 1978; Kahneman & Tversky, 1973). As a result, evidence that is both redundant and flimsy tends to produce judgments that are held with too much confidence. These judgments will be presented too assertively to others and are likely to be believed more than they deserve to be. The safe way to evaluate the probable accuracy of a judgment (our own or someone else's) is by considering the validity of the environment in which the judgment was made as well as the judge's history of learning the rules of that environment.

Professional Intuitions

We are of course not the first to have identified a regular environment and an adequate opportunity to learn it as preconditions for the development of skills, including intuitive skills (see, e.g., Hogarth, 2001). Other investigators have focused on attitude, motivation, talent, and deliberate practice as crucial to skill development (Ericsson, 2006; Ericsson et al., 2006).

The importance of predictable environments and opportunities to learn them was apparent in an early review of professions in which expertise develops. Shanteau (1992) reviewed evidence showing that expertise was found in livestock judges, astronomers, test pilots, soil judges, chess masters, physicists, mathematicians, accountants, grain inspectors, photo interpreters, and insurance analysts. In contrast, Shanteau noted poor performance by experienced professionals in another large set of occupations: stockbrokers, clinical psychologists, psychiatrists, college admissions officers, court judges, personnel selectors, and intelligence analysts. Shanteau searched for task characteristics that distinguished the domains in which experts did well from those in which experts did poorly. The factors that we identified—the predictability of outcomes, the amount of experience, and the availability of good feedback—were included in his list. In addition, Shanteau pointed to static (as opposed to dynamic) stimuli as favorable to good performance.

Three professions—nurses, physicians, and auditors—appeared on both of Shanteau's (1992) lists. These professionals exhibited genuine expertise in some of their activities but not in others. We refer to such mixed grades for professionals as "fractionated expertise," and we believe that the fractionation of expertise is the rule, not an exception. For example, auditors who have expertise in "hard" data such as accounts receivable may do much less well with "soft" data such as indications of fraud (J. Shanteau, personal communication, February 12, 2009).

There are a few activities, such as chess, in which a master will not encounter challenges that are genuinely new. In most domains, however, professionals will occasionally have to deal with situations and tasks that they have not had an opportunity to master. Physicians, as is well known, encounter from time to time diagnostic problems that are entirely new to them—they have expertise in some diagnoses but not in others. Similarly, weather forecasters are more successful in the routine prediction of temperature and precipitation than in forecasting hail (Stewart, Roebber, & Bosart, 1997).

Characteristically, we came to the topic of fractionated expertise with different examples in mind. GK focuses on the experts who perform a constant task (e.g., putting out fires; establishing a diagnosis) but encounter unfamiliar situations. The ability to recognize that a situation is anomalous and poses a novel challenge is one of the manifestations of authentic expertise. Descriptions of diagnostic thinking in medicine emphasize the intuitive ability of

some physicians to realize that the characteristics of a case do not fit into any familiar category and call for a deliberate search for the true diagnosis (Gawande, 2002; Groopman, 2007).

DK is particularly interested in cases in which professionals who know how to use their knowledge for some purposes attempt to use the same knowledge for other purposes. He views the fractionation of expertise as one element in the explanation of the illusion of validity: the overconfidence that professionals sometimes experience in dealing with problems in which they have little or no skill. Finance professionals, psychotherapists, and intelligence analysts may know a great deal about a particular company, patient, or international conflict, and they may have received ample feedback supporting their confidence in the performance of some tasks—typically those that deal with the short term—but the feedback they receive from their failures in long-term judgments is delayed, sparse, and ambiguous. The experience of the professionals that DK has thought about is therefore conducive to overconfidence.

These professionals may have strong subjective confidence in their judgments, but we do not believe that subjective confidence reliably indicates whether intuitive judgments or decisions are valid. When experts recognize anomalies, using judgments of typicality and familiarity, they are detecting violations of patterns in the *external* situation. In contrast, people do not have a strong ability to distinguish correct intuitions from faulty ones. People, even experts, do not appear to be skilled in detecting patterns in the *internal* situation in order to identify the basis for their judgments. Therefore, reliance on subjective confidence may contribute to overconfidence.

The experts that GK has studied seem less susceptible to overconfidence, perhaps in part because of the direct personal risks it poses. Weather forecasters, engineers, and logistics specialists typically resist requests to make judgments about matters that fall outside their area of competence. People in professions marked by standard methods, clear feedback, and direct consequences for error appear to appreciate the boundaries of their expertise. These experts know more knowledgeable experts exist. Weather forecasters know there are people in another location who better understand the local dynamics. Structural engineers know that chemical engineers, or even structural engineers working with different types of models or materials, are the true experts who should be consulted.

As in the other topics that we have considered, we find no reason to disagree about either fractionation of expertise or overconfidence. As usual, different rules apply to different tasks.

Augmenting Professional Judgment: The Use of Algorithms

The attitude toward the Meehl paradigm, in which intuitions and professional judgments are set in competition, is a sore point in conversations between adherents of NDM and HB. The idea of algorithms that outdo human judges is a source of pride and joy for members of the HB tribe, but algorithms are usually distrusted by the NDM community.

There is compelling evidence that under certain conditions mechanical and analytical judgments outperform human judgment. Grove, Zald, Lebow, Snitz, and Nelson (2000) reported a meta-analysis of 136 studies that compared the accuracy of clinical and mechanical judgments, most within the domains of clinical psychology and medicine. Their review excluded studies involving nonhuman outcomes such as horse races and weather. The preponderance of data favored the algorithms (i.e., the "mechanical" judgments), which were superior in about half the studies ($n = 63$). The other half of the studies showed no difference ($n = 65$), and only a few studies showed better performance by the clinical judgments ($n = 8$). For example, the tasks for which there was at least a 17-point difference in effect size favoring mechanical over clinical judgments included the following: college academic performance, presence of throat infections, diagnosis of gastrointestinal disorders, length of psychiatric hospitalization, job turnover, suicide attempts, juvenile delinquency, malingering, and occupational choice.

Findings in which the performance of human judges is inferior to that of simple algorithms are often cited as evidence of cognitive ineptitude, but this conclusion is unwarranted. The correct conclusion is that people perform significantly more poorly than algorithms *in low-validity environments*. The tasks reviewed by Grove et al. (2000) generally involved noisy and/or highly complex situations. The forecasts made by the algorithms were often wrong, albeit less often than the clinical predictions. The studies in the Meehl paradigm have not produced "smoking gun" demonstrations in which clinicians miss highly valid cues that the algorithm detects and uses. Indeed, such an outcome would be unlikely, because human learning is normally quite efficient. Where simple and valid cues exist, humans will find them if they are given sufficient experience and enough rapid feedback to do so—except in the environments that Hogarth (2001) labeled "wicked," in which the feedback is misleading. A statistical approach has two crucial advantages over human judgment when available cues are weak and uncertain: Statistical analysis is more likely to identify weakly valid cues, and a prediction algorithm will maintain above-chance accuracy by using such cues consistently. The meta-analysis performed by Karelaia and Hogarth (2008) showed that consistency accounted for much of the advantage of algorithms over humans.

The evaluation and approval of personal loans by loan officers is an example of a situation in which algorithms should be used to replace human judgment. Identifying the relatively small number of defaulting loans is a low-validity task because of the low base rate of the critical outcome. Algorithms have largely replaced human judges in this task, using as inputs objective demographic and personal data rather than subjective impression of reliability. The result is an unequivocal improvement: We have fairer loan judgments (i.e., judgments that are not improperly influenced by gender or race), faster decisions, and reduced expenses.

Our analysis suggests that algorithms significantly outperform humans under two quite different conditions: (a) when validity is so low that human difficulties in detecting weak regularities and in maintaining consistency of judgment are critical and (b) when validity is very high, in highly predictable environments, where ceiling effects are encountered and occasional lapses of attention can cause humans to fail. Automatic transportation systems in airports are an example in that class.

NDM proponents correctly emphasize that the conditions necessary for the construction and use of an algorithm are stringent. These conditions include (a) confidence in the adequacy of the list of variables that will be used, (b) a reliable and measurable criterion, (c) a body of similar cases, (d) a cost/benefit ratio that warrants the investment in the algorithmic approach, and (e) a low likelihood that changing conditions will render the algorithm obsolete. We also agree that algorithms that substitute for human judgment must remain under human supervision, to provide continuous monitoring of their performance and of relevant changes in the environment. Maintaining adequate supervision of algorithms can be difficult, because there is evidence that human operators become more passive and less vigilant when algorithms are in charge—a phenomenon that has been labeled "automation bias" (Skitka, Mosier & Burdick, 1999, 2000).

We agree that the introduction of algorithms and other formal decision aids in organizations will often encounter opposition and unexpected problems of implementation. Few people enjoy being replaced by mechanical devices or by mathematical algorithms, and many devices and algorithms function less well in the real world than on the planning board (Yates, Veinott, & Patalano, 2003). Even decision aids and procedures that leave the authority of the decision maker intact—decision analysis is a salient example—are often resisted, for both good and bad reasons. Naturally, we have somewhat different attitudes toward these problems of implementation, with DK usually viewing them as obstacles to be overcome and GK seeing them as reasons to be skeptical about the value of formal methods.

Despite our different attitudes toward formal methods, we agree on the potential of semi-formal strategies. An example is the premortem method (Klein, 2007) for reducing overconfidence and improving decisions. Project teams using this method start by describing their plan. Next they imagine that their plan has failed and the project has been a disaster. Their task is to write down, in two minutes, all the reasons why the project failed. The facilitator goes around the table, getting reasons from each of the team members, starting with the leader. The rationale for the method is the concept of prospective hindsight (Mitchell, Russo, & Pennington, 1989)—that people can generate more criticisms when they are told that an outcome is certain. It also offers a solution to one of the major problems of decision making within organizations: the gradual suppression of dissenting opinions, doubts, and objections, which is typically observed as an organization commits itself to a major plan. The premortem method is consistent with the HB concern for overconfidence while drawing on and respecting the expertise of decision makers, a hallmark of the NDM approach. We expect that there are additional methods that can synthesize the strengths of the two traditions.

Conclusions

In an effort that spanned several years, we attempted to answer one basic question: Under what conditions are the intuitions of professionals worthy of trust? We do not claim that the conclusions we reached are surprising (many were anticipated by Shanteau, 1992, Hogarth, 2001, and Myers, 2002, among others), but we believe that they add up to a coherent view of expert intuition, which is more than we expected to achieve when we began.

- Our starting point is that intuitive judgments can arise from genuine skill—the focus of the NDM approach—but that they can also arise from inappropriate application of the heuristic processes on which students of the HB tradition have focused.
- Skilled judges are often unaware of the cues that guide them, and individuals whose intuitions are not skilled are even less likely to know where their judgments come from.
- True experts, it is said, know when they don't know. However, nonexperts (whether or not they think they are) certainly do not know when they don't know. Subjective confidence is therefore an unreliable indication of the validity of intuitive judgments and decisions.
- The determination of whether intuitive judgments can be trusted requires an examination of the environment in which the judgment is made and of the opportunity that the judge has had to learn the regularities of that environment.
- We describe task environments as "high-validity" if there are stable relationships between objectively identifiable cues and subsequent events or between cues and the outcomes of possible actions. Medicine and firefighting are practiced in environments of fairly high validity. In contrast, outcomes are effectively unpredictable in zero-validity environments. To a good approximation, predictions of the future value of individual stocks and long-term forecasts of political events are made in a zero-validity environment.
- Validity and uncertainty are not incompatible. Some environments are both highly valid and substantially uncertain. Poker and warfare are examples. The best moves in such situations reliably increase the potential for success.
- An environment of high validity is a necessary condition for the development of skilled intuitions. Other necessary conditions include adequate opportunities for learning the environment (prolonged practice and feedback that is both rapid and unequivocal). If an environment provides valid cues and good feedback, skill and expert intuition will eventually develop in individuals of sufficient talent.
- Although true skill cannot develop in irregular or unpredictable environments, individuals will some-

times make judgments and decisions that are successful by chance. These "lucky" individuals will be susceptible to an illusion of skill and to overconfidence (Arkes, 2001). The financial industry is a rich source of examples.

- The situation that we have labeled fractionation of skill is another source of overconfidence. Professionals who have expertise in some tasks are sometimes called upon to make judgments in areas in which they have no real skill. (For example, financial analysts may be skilled at evaluating the likely commercial success of a firm, but this skill does not extend to the judgment of whether the stock of that firm is underpriced.) It is difficult both for the professionals and for those who observe them to determine the boundaries of their true expertise.
- We agree that the weak regularities available in low-validity situations can sometimes support the development of algorithms that do better than chance. These algorithms only achieve limited accuracy, but they outperform humans because of their advantage of consistency. However, the introduction of algorithms to replace human judgment is likely to evoke substantial resistance and sometimes has undesirable side effects.

Another conclusion that we both accept is that the approaches of our respective communities have built-in limitations. For historical and methodological reasons, HB researchers generally find errors more interesting and instructive than correct performance; but a psychology of judgment and decision making that ignores intuitive skill is seriously blinkered. Because their intellectual attitudes developed in reaction to the HB tradition, members of the NDM community have an aversion to the word *bias* and to the corresponding concept; but a psychology of professional judgment that neglects predictable errors cannot be adequate. Although we agree with both of these conclusions, we have yet to move much beyond recognition of the problem. DK is still fascinated by persistent errors, and GK still recoils when biases are mentioned. We hope, however, that our effort may help others do more than we have been able to do in bringing the insights of both communities to bear on their common subject.

REFERENCES

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 495–516). Boston: Kluwer Academic.

Bazerman, M. H. (2005). *Judgment in managerial decision making* (6th ed.). Hoboken, NJ: Wiley.

Beach, L. R. (1990). *Image theory: Decision making in personal and organizational contexts.* Chichester, England: Wiley.

Brunswik, E. (1957). Scope and aspects of the cognitive problem. In H. Gruber, K. R. Hammond, & R. Jessor (Eds.), *Contemporary approaches to cognition* (pp. 5–31). Cambridge, MA: Harvard University Press.

Cannon-Bowers, J. A., & Salas, E. (Eds.). (1998). *Making decisions under stress: Implications for individual and team training.* Washington, DC: American Psychological Association.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

Collyer, S. C., & Malecki, G. S. (1998). Tactical decision making under stress: History and overview. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 3–15). Washington, DC: American Psychological Association.

Crandall, B., & Gamblian, V. (1991). *Guide to early sepsis assessment in the NICU* [Instruction manual prepared for the Ohio Department of Development under the Ohio Small Business Innovation Research Bridge Grant program]. Fairborn, OH: Klein Associates.

Crandall, B., & Getchell-Reiter, K. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the "intuition" of NICU nurses. *Advances in Nursing Sciences, 16*(1), 42–51.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis.* Cambridge, MA: MIT Press.

Croskerry, P., & Norman, G. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine, 121,* S24–S29.

deGroot, A. D. (1978). *Thought and choice in chess.* The Hague: Mouton. (Original work published 1946)

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 395–416.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, R. R. Hoffman, & P. J. Feltovich (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 39–68). New York: Cambridge University Press.

Ericsson, K. A., Charness, N., Hoffman, R. R., & Feltovich, P. J. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance.* New York: Cambridge University Press.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment.* Hove, East Sussex, England: Psychology Press.

Evans, J. St. B. T., & Frankish, K. (Eds.). (2009). *In two minds: Dual processes and beyond.* New York: Oxford University Press.

Fogarty, W. M. (1988). *Formal investigation into the circumstances surrounding the downing of a commercial airliner by the U. S. S. Vincennes (CG 49) on 3 July 1988* [Unclassified Letter Ser. 1320 of 28 July 1988, to Commander in Chief, U.S. Central Command]. Washington, DC: U.S. Department of the Navy.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Gawande, A. (2002). *Complications: A surgeon's notes on an imperfect science.* London: Profile Books.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment.* New York: Cambridge University Press.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & A. B. C. Research Group (Eds.), *Simple heuristics that make us smart* (pp. 37–58). New York: Oxford University Press.

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

Groopman, J. (2007). *How doctors think.* New York: Houghton Mifflin.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2007). Blinking on the bench: How judges decide cases. *Cornell Law Review, 93,* 1–43.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-17*(5), 753–770.

Hertwig, R., Hoffrage, U., & Martingnon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P. M. Todd, & A. B. C. Research Group (Eds.), *Simple heuristics that make us smart* (pp. 37–58). New York: Oxford University Press.

Hogarth, R. M. (2001). *Educating intuition.* Chicago: University of Chicago Press.
Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21,* 1161–1166.
Johnson, J. G., & Raab, M. (2003). Take the first: Option generation and resulting choices. *Organizational Behavior and Human Decision Processes, 91*(2), 215–229.
Kahneman, D. (2003). Autobiography. In T. Frangsmyr (Ed.), *Les Prix Nobel 2002* [Nobel Prizes 2002]. Stockholm, Sweden: Almqvist & Wiksell International.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.
Kahneman, D., & Renshon, J. (2007). Why hawks win. *Foreign Policy, 158,* 34–38.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.
Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134,* 404–426.
Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138–147). Norwood, NJ: Ablex.
Klein, G. (1998). *Sources of power: How people make decisions.* Cambridge, MA: MIT Press.
Klein, G. (2007, September). Performing a project premortem. *Harvard Business Review,* pp. 18–19.
Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fireground. In *Proceedings of the Human Factors and Ergonomics Society 30th Annual Meeting* (Vol. 1, pp. 576–580). Norwood, NJ: Ablex.
Klein, G. A., Orasanu, J., Calderwood, R., & Zsambok, C. E. (1993). *Decision making in action: Models and methods.* Norwood, NJ: Ablex.
Klein, G., Wolf, S., Militello, L., & Zsambok, C. (1995). Characteristics of skilled option generation in chess. *Organizational Behavior and Human Decision Processes, 62,* 63–69.
Lewis, M. (2003). *Moneyball: The art of winning an unfair game.* New York: Norton.
Lipshitz, R. (1993). Converging themes in the study of decision making in realistic settings. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 103–137). Norwood, NJ: Ablex.
Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69,* 220–232.
Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.
Meehl, P. E. (1973). Why I do not attend case conferences. In P. E. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225–302). Minneapolis: University of Minnesota Press.
Mitchell, D., Russo, J., & Pennington, N. (1989). Back to the future: Temporal perspective in the explanation of events. *Journal of Behavioral Decision Making, 2,* 25–38.
Montgomery, H. (1993). The search for a dominance structure in decision making: Examining the evidence. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 182–187). Norwood, NJ: Ablex.
Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology, 78,* 1038–1052.
Myers, D. G. (2002). *Intuition: Its powers and perils.* New Haven, CT: Yale University Press.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.
Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 3–20). Norwood, NJ: Ablex.
Rasmussen, J. (1986). *Information processing and human–machine interaction: An approach to cognitive engineering.* Amsterdam: North-Holland.
Rosenzweig, P. (2007). *The halo effect . . . and the eight other business delusions that deceive managers.* New York: Free Press.
Schraagen, J. M. C., Chipman, S. F., & Shalin, V. J. (Eds.). (2000). *Cognitive task analysis.* Mahwah, NJ: Erlbaum.
Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53,* 252–262.
Simon, H. A. (1992). What is an explanation of behavior? *Psychological Science, 3,* 150–161.
Skitka, L. J., Mosier, K., & Burdick, M. D. (1999). Does automation bias decision-making? *International Journal of Human–Computer Studies, 51,* 991–1006.
Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human–Computer Studies, 52,* 701–717.
Slovic, P. (Ed.). (2000). *The perception of risk.* London: Earthscan.
Smith, P. J., Giffin, W. C., Rockwell, T. H., & Thomas, M. (1986). Modeling fault diagnosis as the activation and use of a frame system. *Human Factors, 28,* 703–716.
Stewart, T. R., Roebber, P. J., & Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision Processes, 69,* 205–219.
Sunstein, C. R. (Ed.). (2000). *Behavioral law and economics.* New York: Cambridge University Press.
Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.
Woods, D. D., O'Brien, J., & Hanes, L. F. (1987). Human factors challenges in process control: The case of nuclear power plants. In G. Salvendy (Ed.), *Handbook of human factors/ergonomics* (pp. 1724–1770). New York: Wiley.
Yates, J. F., Veinott, E. S., & Patalano, A. L. (2003). Hard decisions, bad decisions: On decision quality and decision aiding. In S. L. Schneider & J. C. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 13–63). New York: Cambridge University Press.