



This postprint was originally published by the APA as:  
Lejarraga, T., & Hertwig, R. (2021). **How experimental methods shaped views on human competence and rationality.** *Psychological Bulletin*, 147(6), 535–564. <https://doi.org/10.1037/bul0000324>

Supplementary material to this article is available. For more information see <http://hdl.handle.net/21.11116/0000-0007-EC18-5>

**The following copyright notice is a publisher requirement:**

©American Psychological Association, 2021. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission.

The final article is available, upon publication, at:

<https://doi.org/10.1037/bul0000324>

**Provided by:**

Max Planck Institute for Human Development

Library and Research Information

[library@mpib-berlin.mpg.de](mailto:library@mpib-berlin.mpg.de)

# How Experimental Methods Shaped Views on Human Competence and Rationality

Tomás Lejarraga<sup>1</sup> and Ralph Hertwig<sup>2</sup>

<sup>1</sup>Departament d'Economia de l'Empresa, Universitat de les Illes Balears

<sup>2</sup>Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

Within just 7 years, behavioral decision research in psychology underwent a dramatic change: In 1967, Peterson and Beach (1967) reviewed more than 160 experiments concerned with people's statistical intuitions. Invoking the metaphor of the mind as an intuitive statistician, they concluded that "probability theory and statistics can be used as the basis for psychological models that integrate and account for human performance in a wide range of inferential tasks" (p. 29). Yet in a 1974 *Science* article, Tversky and Kahneman rejected this conclusion, arguing that "people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simple judgmental operations" (p. 1124). With that, they introduced the heuristics-and-biases research program, which has profoundly altered how psychology, and the behavioral sciences more generally, view the mind's competences and rationality. How was this radical transformation possible? We examine a previously neglected driver: The heuristics-and-biases program established an experimental protocol in behavioral decision research that relied on described scenarios rather than learning and experience. We demonstrate this shift with an analysis of 604 experiments, which shows that the descriptive protocol has dominated post-1974 research. Specifically, we examine two lines of research addressed in the intuitive-statistician program (Bayesian reasoning and judgments of compound events) and two lines of research spurred by the heuristics-and-biases program (framing and anchoring and adjustment). We conclude that the focus on description at the expense of learning has profoundly shaped the influential view of the error-proneness of human cognition.

---

## Public Significance Statement

Sound statistical intuitions are essential for navigating an uncertain world. The intuitive-statistician program of the 1960s concluded that probability theory and statistics can be used as the basis for psychological models of judgment. In contrast, research starting in the 1970s—spearheaded by the heuristics-and-biases program—has concluded that people lack the correct mental software for many important judgmental tasks. Our systematic review of experimental methods shows that the source of these conflicting conclusions may be traceable to a methodological shift triggered by the heuristics-and-biases program, which largely removed learning from judgment tasks.

---

*Keywords:* cognitive bias, description–experience gap, experimental methodology, intuitive statistician, learning

*Supplemental materials:* <http://hdl.handle.net/21.11116/0000-0007-EC18-5>

---

The lesson is: think about practice, not theory. (Hacking, 1983, p.274). This article examines a transformational period in psychological research on statistical reasoning, decision making, and human

---

Tomás Lejarraga  <https://orcid.org/0000-0001-6135-2243>

Ralph Hertwig  <https://orcid.org/0000-0002-9908-9556>

We are grateful to Jill de Ron and Maximilian Rank for research assistance and Ana Sofia Morais for her invaluable help in rating the studies. We thank Hal Arkes, Lee R. Beach, Jerome Busemeyer, Gerd Gigerenzer, Werner Güth, Glenn W. Harrison, Robin Hogarth, Andreas Ortmann, and Thomas Wallsten for many helpful comments. In memory of Silvia J. Camino (1944–2020).

Correspondence concerning this article should be addressed to Tomás Lejarraga, Departament d'Economia de l'Empresa, Universitat de les Illes Balears, Carretera de Valldemossa, Palma 07122, Spain. Email: [tomas.lejarraga@uib.eu](mailto:tomas.lejarraga@uib.eu)

rationality. The *heuristics-and-biases program* of the late 1960s and early 1970s (Kahneman et al., 1982; Tversky & Kahneman, 1974) was to profoundly change the way many psychologists (and later economists) thought about the quality and cognitive mechanisms of people's statistical reasoning and intuitions. Described as nothing short of a "revolution in cognitive psychology and economics" (Miller & Gelman, 2018, p. 2; see also, e.g., Gilovich & Griffin, 2002),<sup>1</sup> the *heuristics-and-biases program* has provided many behavioral scientists with a lodestar—namely, a focus on the "systematic biases that separate the beliefs that people have and the choices they make from the optimal beliefs and choices assumed in rational-agent models" (Kahneman, 2003, p. 1449). This focus offered an influential interpretation of Simon's (1956) foundational concept of bounded rationality, namely, in terms of a map of systematic deviations (Kahneman, 2003) from what the program took as the normative benchmark of decision making: the rational choice framework. Finally, the *heuristics-and-biases program* has made strong inroads into subfields of psychology (e.g., social psychology; Nisbett & Ross, 1980) as well as neighboring fields, such as management science (e.g., Bazerman & Moore, 2013) and clinical and medical decision making (e.g., Dawson & Arkes, 1987; Elstein, 1999), it has also spawned subfields of economics and law, namely, behavioral economics (e.g., Thaler, 2016) and behavioral law (e.g., Sunstein, 2000). It was recognized not once but twice in the Nobel Prize in economics—awarded to Kahneman in 2002 and to Thaler in 2017. Arguably the most prominent psychological research program of the second half of the 20th century, the *heuristics-and-biases program* continues to influence research to this day.<sup>2</sup>

The *heuristics-and-biases program* superseded a program founded by Ward Edwards—today widely considered the father of behavioral decision research. Edwards was one of the first to take an empirical approach to the behavioral implications and assumptions of economic theory (e.g., Edwards, 1954b, 1961a) in a program described as *man as an intuitive statistician* (Peterson & Beach, 1967).<sup>3</sup> The two research programs portrayed humans' intuitive ability to reckon with risk and uncertainty and their rationality—or lack thereof—in starkly contrasting terms. They invoked different theoretical metaphors and underlying cognitive mechanisms and produced diverging and sometimes opposing experimental findings. Our goal is to understand the reasons behind this transformation. We will examine one catalyst of the *heuristics and-biases program's* swift rise and eventual displacement of the *intuitive-statistician program*, focusing on a neglected dimension that may have major implications for the empirical study of human rationality: changes in the experimental culture of research in behavioral decision making and related fields in cognitive and social psychology.

Writing mostly about experiments in physics, the philosopher of science Ian Hacking (1983) was concerned with the relationship between the realism of scientific theory and that of scientific entities and experiments. Emphasizing the role of the experiment, Hacking (1984) wrote:

Different sciences at different times exhibit different relationships between "theory" and "experiment." One chief role of experiment is the creation of phenomena. Experimenters bring into being phenomena that do not naturally exist in a pure state. These phenomena are the touchstones of physics, the keys to nature, and the source of much modern technology. [ . . . ] Most of the phenomena, effects, and events created by the experimenter are like plutonium: they do not exist in nature except possibly on vanishingly rare occasions. (p. 155)

Experiments, more than theories, convince people that scientific entities are real:

We are completely convinced of the reality of electrons when we regularly set out to build—and often enough succeed in building—new kinds of device that use various well-understood causal properties of electrons to interfere in other more hypothetical parts of nature. (Hacking, 1983, p. 265)

Psychology is not physics. And experimenters in psychology rarely build new devices based on unobservable entities to create even more abstract entities. Yet experiments are likely to shape psychologists' confidence in the reality of what they manipulate and observe. Thus, experiments—and the norms, standards, and routines that inform how they are conducted—are important for understanding the body of knowledge, the scientific entities, and the theories and theoretical metaphors to which experimenters feel committed.

Our focus will be on experiments designed to study the degree to which people's intuitive statistical reasoning is rational or irrational. We first describe the profoundly different conclusions about intuitive statistical reasoning and human rationality drawn by the *heuristics-and-biases program* and the *intuitive-statistician program*. Then, using research on Bayesian reasoning as a first concrete example, we demonstrate the distinct experimental approaches and cultures of the two programs and suggest that these disparities could contribute to the programs' conflicting findings and conclusions. Next, we systematically analyze research on probability estimates of compound events as a second example. Like Bayesian reasoning, this research topic emerged before the *heuristics-and-biases program*; it received an enormous boost through Tversky and Kahneman's (1983) work on the *conjunction fallacy* and has flourished ever since. We then turn to two lines of research that were spurred by Kahneman and Tversky's groundbreaking studies: research on framing and on anchoring and adjustment. In total, we investigate the shift in experimental culture across a total of 604 studies sampled from the two research programs. To preview our key finding, we find substantial and systematic differences in the programs' experimental approaches. We argue that these differences warrant the attention of both experimentalists and theorists of human rationality. The lesson we learned is the one Hacking (1983) emphasized: Think about experimenters' practices. We begin by outlining the two research programs.

<sup>1</sup> See also the tributes to Daniel Kahneman on the occasion of his 80th birthday at [https://www.edge.org/conversation/daniel\\_kahneman-on-kahneman](https://www.edge.org/conversation/daniel_kahneman-on-kahneman).

<sup>2</sup> We are indebted to Deb Ain and Susannah Goss for editing the manuscript.

<sup>3</sup> It was Egon Brunswik (1955b) who originally coined the term *intuitive statistician* (p. 212) to describe the human perceptual system that uses uncertain (probabilistic) cues to estimate important variables in the environment (e.g., distance or size of an object)

## The Intuitive-Statistician Program: Key Tenets and Propositions

Two review articles by Edwards were the founding documents of research on “behavioral decision theory” (Edwards, 1954b, 1961a). In “The theory of decision making” published in *Psychological Review* (Edwards, 1954b), he introduced theories of riskless and risky decision making, game theory and research and, importantly, behavioral experiments capable of testing them. His goal was to persuade researchers that “all these topics represent a new and rich field for psychologists, in which a theoretical structure has already been elaborately worked out and in which many experiments need to be performed” (p. 411). His call to action was heeded. The new research endeavor grew swiftly: A total of 139 articles on behavioral decision theory were published between 1954 and 1960 (Phillips & von Winterfeldt, 2007). In addition, Peterson and Beach (1967) reviewed 110 articles by 102 authors reporting on more than 160 experiments<sup>4</sup> on people’s statistical intuitions, such as the judgment of the mean of a series of observations (Spencer, 1961), estimates of variance (Beach & Scopp, 1967), Bayesian updates of probabilistic beliefs about random events (Edwards, 1966), probability estimates of compound events (J. Cohen et al., 1958), and judgment of correlations (Erlick, 1966). They concluded that “in general, the results indicate that probability theory and statistics can be used as the basis for psychological models that integrate and account for human performance in a wide range of inferential tasks” (p. 29).

In other words, based on their review of more than 160 experiments, Peterson and Beach (1967) saw enough correspondence between the predictions of normative models (i.e., rules from probability theory and statistics) and people’s inferences, judgments, and choices to conclude that the normative models could provide a basis for the design of psychological models. The sample of conclusions taken from studies in Peterson and Beach (1967) presented in Table 1 illustrates this finding.

Peterson and Beach (1967) presented another important tenet in the introductory paragraph of their article:

... [m]an must cope with an environment about which he has only fallible information, “while God may not gamble, animals and humans do, . . . they cannot help but to gamble in an ecology that is of essence only partly accessible to their foresight” (Brunswik, 1955[a]). And man gambles well. He survives and prospers while using the fallible information to infer the states of his uncertain environment and to predict future events. (p. 29)

Peterson and Beach (1967) were not the only researchers to take inspiration from Brunswik and his conclusion that the perceptual system—and the cognitive system more generally—works like an intuitive statistician inferring the environment (Brunswik, 1957). The metaphor of the intuitive statistician had already featured in Edwards’s (1961a) review—though without attribution to Brunswik—at the end of a discussion of dynamic decision making: The upshot of these studies of man (or rather, college student) as statistician is that he makes a fairly good one. In all cases the differences are in the proper directions, though they are not always the proper amounts (p. 490).

As we shall see shortly, Edwards more than once concluded that the correspondence between behaviors and normative benchmarks is good but not perfect. The impact that Brunswik had on Edwards’s reasoning also shines through in Edwards’s only published criticism of the heuristics-and-biases program, to which we return later.

## The Heuristics-and-Biases Program: Key Tenets and Propositions

The heuristics-and-biases program rejected the idea that the normative rules of probability theory and statistics are an appropriate basis for psychological models. In a *Science* article, Tversky and Kahneman (1974) showcased four groundbreaking articles they had published between 1971 and 1973, introducing three heuristics: the availability, representativeness, and anchoring-and-adjustment heuristics. In one of those articles, they clearly broke with the intuitive-statistician program and its idea of approximate correspondence between behavior and norm:

In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors. (Kahneman & Tversky, 1973, p. 237)

The notion that intuitive probabilistic cognition does not obey the rules of probability theory and statistics, either literally or approximately, transformed experimenters’ phenomena of interest, explanatory concepts, and perspectives on the rationality of the intuitive mind. The phenomena of interest were now deviations from what was assumed to be normatively correct (but see Birnbaum, 1983; Gigerenzer et al., 1989); heuristics were the explanatory concepts that researchers used to account for these deviations. Heuristics and biases (also dubbed “cognitive illusions”) became symbiotic concepts insofar as biases revealed “some heuristics of thinking under uncertainty” (Tversky & Kahneman, 1974, p. 1124). In turn, to explain the existence of a bias, a heuristic required “a logic of its own, which departs systematically from the logic of probability” (Tversky & Kahneman, 1982, p. 88, commenting on the representativeness heuristic).

The new findings prompted stark conclusions about the prevalence of cognitive illusions, the error-proneness of human cognition, and the implications for human rationality in psychology and beyond. As Kahneman and Tversky (1972) argued, “for anyone who would wish to view man as a reasonable intuitive statistician, such results are discouraging” (p. 445). Others, including some social psychologists and economists, went further. For instance, Nisbett et al. (1983) concluded that “people commit serious errors of inference” and that it is “disturbing to learn that heuristics people use in such tasks do not respect the required statistical principles” (pp. 339–340). Nisbett and Borgida (1975) thought the experimental findings had “bleak implications for human rationality” (p. 935), and Fiske and Taylor (1991) coined the term *cognitive miser* to describe the human mind. Thaler concluded that “mental illusions should be considered the rule rather than the exception” (p. 4)

---

<sup>4</sup> The exact number of experiments is uncertain because we were unable to retrieve 10 articles for our review. The number of experiments in the documents we could retrieve was 164.

**Table 1**  
*Sample of Illustrative Conclusions About People's Intuitive Statistical Reasoning From Research Reviewed by Peterson and Beach (1967)*

Authors	Conclusions
Peterson and Beach (1967)	"Experiments that have compared human inferences with those of statistical man show that the normative model provides a good first approximation for a psychological theory of inference. Inferences made by subjects are influenced by appropriate variables and in appropriate directions" (pp. 42–43).
J. Cohen et al. (1958)	"Indeed, their estimates could be expressed as a compound of two component subjective probabilities, in a manner analogous to the combination of mathematical probabilities" (p. 323).
Shuford (1959)	"The majority of the Ss responded in such a manner as to approximate the mathematical rule for computing the probability of the joint occurrence of two independent events" (p. 14).
Spencer (1961)	"When considering the observation that people detect consistent trends with considerable accuracy, it should be noted that the average prediction for the group of subjects agreed closely with the best-fitting <i>straight line</i> prediction" (p. 327).
Peterson and Ulehla (1964)	"Thus, fairly complex inferential behavior was shown to bear orderly relations to mathematical properties of the environment" (p. 530).
Rapoport (1964)	"To summarize, it appears that the Bayesian model does moderately well as a descriptive model in predicting the asymptotic estimates, predictions, and verbal responses of most subjects. Moreover, the accuracy of the model seems to increase as the amount of information available to the decision makers increases. These results are in agreement with results of earlier studies" (p. 372).
Peterson, Hammond, and Summers (1965)	"The response system of Ss fell a little short of the optimal response system" (p. 274).
Beach and Peterson (1966)	"These results support the hypothesis that for unions of events the relationships among subjective probabilities are consistent with those relationships required by probability theory and that this consistency persists even when the subjective probabilities are inaccurate" (p. 308).

While Kahneman and Tversky (1972) noted that heuristics can "sometimes yield reasonable estimates and quite often do not" (p. 431), neither they nor their close collaborators empirically investigated the potential advantages of heuristics. This may be one reason why the idea that biases are ubiquitous, as advocated by scholars such as Thaler (1991), won the day. Our intention here is not to engage with these stark conclusions about human rationality (as many have repeatedly done over the years), but to take them as testimony of the extent to which the heuristics-and-biases program replaced the intuitive-statistician program. For instance, the contrasting citation histories of Peterson and Beach's (1967) *Psychological Bulletin* article and Tversky and Kahneman's (1974) *Science* article attest to how thoroughly the heuristics-and-biases program has shaped researchers' thinking. The two articles were published a mere 7 years apart and both presented experimental evidence—indeed, Peterson and Beach offered an extensive review. Yet Tversky and Kahneman's article has been cited far more often than Peterson and Beach (see Figure 1, left panel), with a cumulative frequency of 15,159 versus 497 citations, respectively, according to Scopus (as of August 15, 2020).<sup>5</sup> Another indication of the success of the heuristics-and-biases program can be found in the Wikipedia entry for *cognitive illusions*, which lists almost 200 distinct cognitive biases, classified as decision-making biases, social biases, and memory errors (see also Krueger & Funder, 2004, who drew up an extensive list of errors of judgment identified by social psychologists and called for a more balanced social psychology).

How could experimental psychology, in one short period, support such diverging conclusions? We argue that an unappreciated shift in experimental culture played a crucial role. We illustrate this shift first by reference to a key inference task: Bayesian reasoning as the yardstick of rational inference and a key building block of economic models of rational choice (see Figure 1, right panel).<sup>6</sup>

### The Emergence of a New Experimental Culture and Bayesian Reasoning

One of Edwards's most important research questions was whether the mind is Bayesian. Preceded only by Rouanet (1961), Edwards and the group around him at the University of Michigan examined what they called "human information processing" using Bayesian-type bookbag-and-poker-chip problems. Bayesianism provided the experimenter's framework for analyzing data<sup>7</sup> and represented the key normative benchmark against which the sequential updating of probabilistic beliefs was measured. Edwards's experimental protocol was informed by his graduate training in psychophysics, the study of how sensory and cognitive systems perceive physical stimuli. Psychophysicists are concerned with the measurement (e.g., the properties and precision) of perception; a typical experiment involves exposing respondents to a long series of visual or auditory stimuli, which they score on meaningful dimensions. Edwards's training was "providential" for his work on judgment and decision making (Schum, 1999, p. 407), his experimental protocol for Bayesian reasoning was reminiscent of that of a psychophysicist—at least in terms of the sequential and

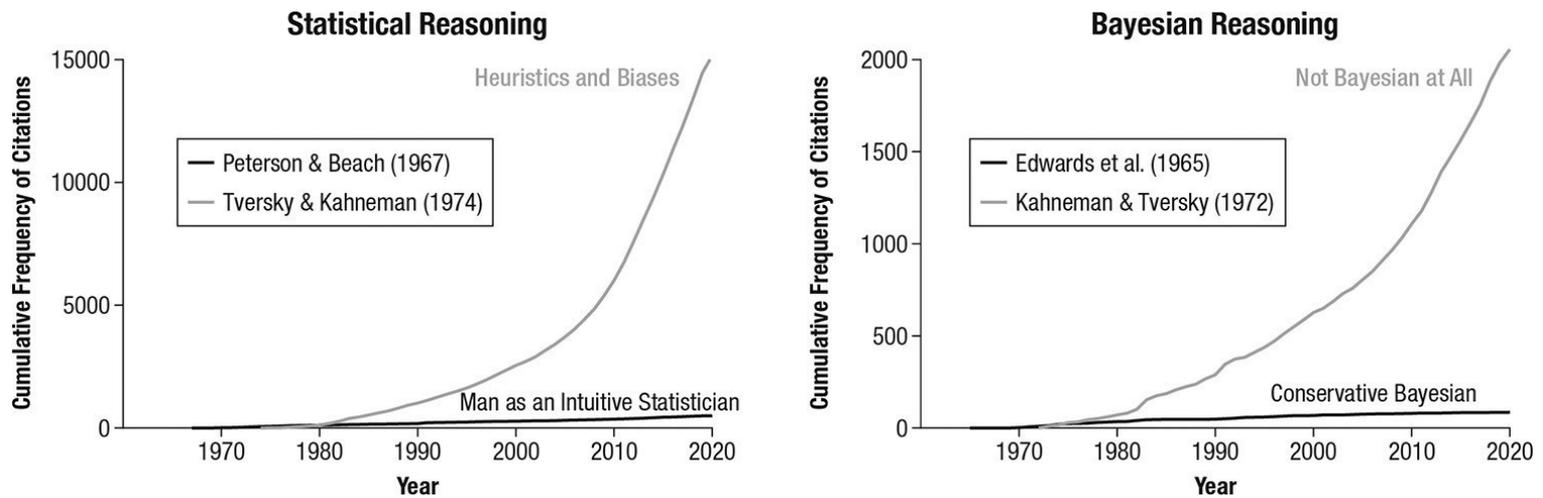
<sup>5</sup> The cumulative frequencies of citations in Google Scholar were 37,331 for Tversky and Kahneman (1974) and 1,165 for Peterson and Beach (1967) on August 15, 2020.

<sup>6</sup> The article by Edwards et al. (1965) is included in the Scopus database as a secondary document. This means that it has been extracted from a Scopus document reference list, but the full document is not directly available in the Scopus database. A similar pattern of citation frequencies was observed in Google Scholar, with a total of 214 for Edwards et al. (1965) and 5,815 for Kahneman and Tversky (1972), on August 15, 2020.

<sup>7</sup> In 1963, Edwards and his colleagues proposed, unsuccessfully, that experimenters use Bayesian statistics (Edwards et al., 1963). It was not until nearly four decades later that experimenters' statistical practices slowly began to change (see Kruschke, 2010).

Figure 1

Cumulative Citation Frequencies of Four Key Articles on Statistical Reasoning and Bayesian Reasoning



Note. Left panel: Cumulative frequency of citations of Peterson and Beach (1967) and Tversky and Kahneman (1974) from 1967 to August 15, 2020, according to Scopus. Right panel: Cumulative frequency of citations of Edwards et al. (1965) and Kahneman and Tversky (1972) from 1965 to August 15, 2020, according to Scopus.

repeated presentation of stimuli. In one task (described in the left column of Figure 2)—later to be adapted by Kahneman and Tversky (1972)—participants are told to imagine that two bookbags contain 100 chips, with red poker chips predominating in one bookbag (e.g., 70% red and 30% blue), and blue chips predominating in the other (70% blue). One bookbag is chosen by flipping a fair coin, and participants are asked to estimate the probability that the chosen bookbag is predominantly red. Participants are then shown 12 chips, which they are told were drawn at random with replacement from a randomly chosen bookbag, one at a time. After each draw, the participant revises their estimate of the probability that the chosen bookbag is predominantly red (Edwards et al., 1965).<sup>8</sup> After the sequence of 12 chips, participants are told which bookbag they came from. Participants repeated this procedure several times (with different bookbags containing different proportions of chips), thus generating hundreds of estimates (e.g., 150 estimates in Edwards, 1966, Study 3; or 480 in Phillips & Edwards, 1966; Experiment 1). Evidence from this experiential protocol on people's Bayesian-reasoning performance accumulated quickly (Edwards et al., 1965; Peterson & Miller, 1965; Phillips & Edwards, 1966). In 1968, Edwards summarized his conclusions:

An abundance of research has shown that human beings are conservative processors of fallible information. Such experiments compare human behavior with the outputs of Bayes's theorem, the formally optimal rule about how opinions (that is, probabilities) should be revised on the basis of new information. It turns out that opinion change is very orderly, and usually proportional to numbers calculated from Bayes's theorem—but it is insufficient in amount. (p. 18)

In other words, people's intuitive reasoning appeared structurally Bayesian in nature, but people revised probabilities to a lesser extent than Bayes's theorem would prescribe—a phenomenon that became known as *conservatism*.

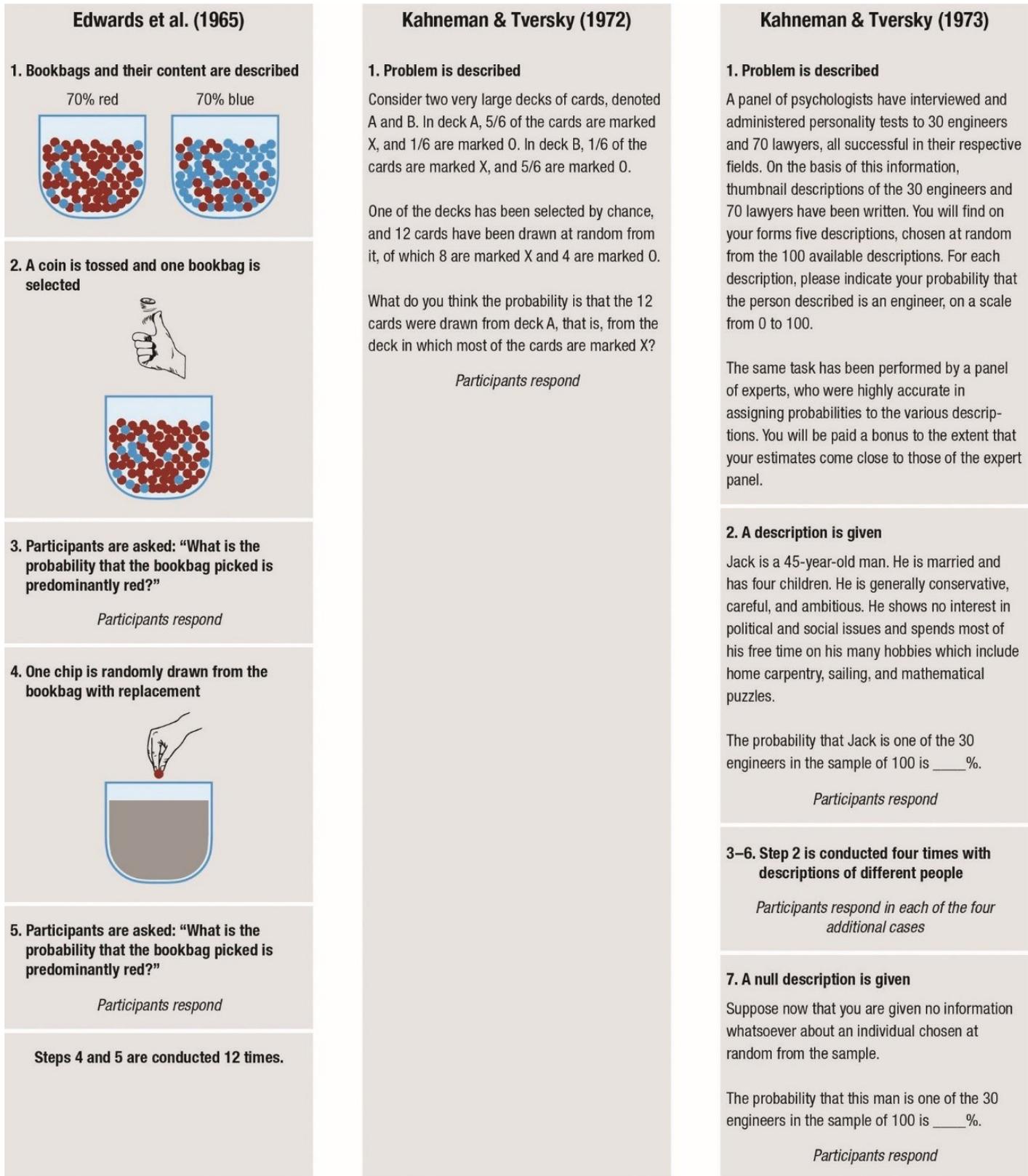
Edwards's conclusion did not remain unchallenged for long. Tversky encountered the latest research on probabilistic inference as a postdoctoral fellow in Edwards's lab (Heukelom, 2012). He then joined the psychology department at the Hebrew University of Jerusalem (Lewis, 2017), where Kahneman invited him to give a seminar on the latest findings in human judgment. Kahneman was skeptical about Tversky's argument that people are Bayesians—albeit conservatively so—in probabilistic inference, and intuitive statisticians more generally: "I knew I was a lousy intuitive statistician. And I really didn't think I was stupider than anyone else" (in Lewis, 2017, p. 148). This moment was to commence one of the most fruitful collaborations in the behavioral sciences. Their experimental approach to Bayesian reasoning was similar to Edwards's, but with crucial differences (see Figure 2, middle column).

Although structurally equivalent to the Edwards et al. (1965) experimental task, Kahneman and Tversky's (1972) task was much easier to administer. Instead of experiencing a series of 12 random draws, participants read a short description of the scenario and of a hypothetical sample. They did not respond sequentially, updating their responses in light of new evidence, but instead produced a single final response. Indeed, the task was so easy to administer that Kahneman ". . . would arrive early each morning and analyze the answers that Oregon college students had given to their questions of the day before" (Lewis, 2017, p. 180). The set-up and implementation costs of the new protocol were also minimal, making research into human judgment widely accessible: Experimentation no longer required a laboratory, materials such as chips and bookbags, software programmers, or complex incentive schemes. In some cases, all people had to do was provide a single answer to a written question. Kahneman later summarized the rationale of the "psychology of single questions," inspired by Mischel's marshmallow test (Mischel & Ebbesen, 1970), as follows (Vedantam, 2018):

And I just fell in love with that idea, the psychology of single questions. And I looked for ways to do that sort of thing. And the work

<sup>8</sup> Edwards et al. (1965) described the task used in previous research: Participants were shown urns in vivo (Peterson, Schneider, & Miller, 1965) or asked to imagine them (Phillips & Edwards, 1966), or dice were used instead of urns (Peterson & Miller, 1965). Responses were given in probabilities (Phillips & Edwards, 1966, Experiment 1) or in odds (Edwards, 1966, and Phillips & Edwards, 1966, Experiment 3; Phillips, 1966, Experiment 1). Estimates were reported by distributing 100 washers over two pegs (Phillips & Edwards, 1966, Experiment 1), 100 discs over two troughs (Phillips & Edwards, 1966, Experiment 2), or by moving a pointer (Peterson, Schneider, & Miller, 1965; Phillips & Edwards, 1966, Experiment 3).

**Figure 2**  
*Three Experimental Tasks Designed to Investigate Bayesian Reasoning*



*Note.* In Edwards et al. (1965), participants sequentially learned new evidence (sampled chips) and provided a series of probability estimates. In Kahneman and Tversky (1972, 1973), participants read a written depiction of a scenario—either a condensed version of the Edwards et al. set-up or the engineer–lawyer problem. See the online article for the color version of this figure.

with Amos on judgment turned out to lend itself to just that, that is there is a single question that elicits a funny thought, and it makes a point.

Whereas Edwards (1968) observed that people are essentially Bayesian but suggested that they give too much weight to base rates in revising their beliefs, Kahneman and Tversky (1972) concluded, using a structurally equivalent task, that “in his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all” (p. 450). Indeed, Kahneman and Tversky observed that people tend to give too little weight to base rates, a phenomenon known as the base-rate fallacy. More generally, Kahneman

and Tversky challenged—in a single article—Peterson and Beach's (1967) comprehensive review of 110 articles and Edwards's (1968) summary of numerous experiments of Bayesian reasoning.<sup>9</sup>

Before we turn to how these conflicting results may have emerged, it is worth noting Wallsten's (1972, 1976) observation that Edwards's conclusions were premised on several key assumptions: People can learn probabilities; the learned probabilities are processed and aggregated according to the Bayesian calculus (e.g., the likelihood ratio principle, the multiplying principle; Wallsten, 1972), and there is a simple one-to-one mapping between people's probability estimates and their subjective probabilities (beliefs). One way to interpret the early work on Bayesian reasoning is that both Edwards and Kahneman and Tversky focused their research primarily on the aggregation assumption, namely, on examining whether information aggregation is Bayesian in nature. The learning aspect and estimate–belief mapping fell by the wayside. Focusing the question on the Bayesian aggregation assumption, as Kahneman and Tversky did, followed naturally from Edwards.<sup>10</sup> This focus on the normative core, however, forestalled questions such as whether and how well people can learn probabilities and how accurately beliefs are translated into subjective probabilities (with a few exceptions, e.g., Wallsten, 1972, 1976).

Surprisingly, neither camp appeared to have considered the conditions and cognitive mechanisms that could give rise to both conservatism and the base-rate fallacy. Indeed, no one cognitive strategy, heuristic, or model of Bayesian reasoning has yet been proposed to explain both phenomena (although Erev et al., 2008 much later proposed that a mere-presentation effect can explain the simultaneous existence of base-rate neglect and base-rate sensitivity). Rather, it became widely accepted that people typically neglect base rates. As Bar-Hillel (1980) argued, “The genuineness, the robustness, and the generality of the base-rate fallacy are matters of established fact” (p. 215).<sup>11</sup> It may be telling that the previous findings of conservatism—a result in profound conflict with the diagnosis of universal base-rate neglect—are not mentioned once in this much-cited publication on Bayesian probability judgment. One possible explanation can be found in a technical report preceding this publication, where Bar-Hillel (1977) wrote that conservatism “isn't a property of people's probability revisions [ . . . ]. The whole finding is a fluke of the paradigm used by the Bayesian approach” (p. 3).

How did these conflicting results—base-rate neglect versus conservatism—emerge? We suggest that one key to understanding the conflicting findings—but probably not the only one<sup>12</sup>—is the change in experimental culture: The experiential protocol used by Edwards and colleagues was abandoned by Tversky and Kahneman and replaced by a more description-based protocol, in which learning receded into the background. Of course, Kahneman and Tversky were not the first to use a descriptive protocol. Symbolic representations such as text vignettes had long been used in various fields in psychology (e.g., the radiation problem in Gestalt psychology; Dunker, 1945). Yet Kahneman and Tversky's research established the descriptive protocol as the new normal in behavioral decision research and related fields.

Bayesian reasoning is just one probabilistic reasoning domain in which a descriptive protocol came to the fore. In their *Science* article, Tversky and Kahneman (1974) summarized 13 behaviors that in their view constituted systematic errors in probabilistic reasoning; these behaviors were observed in studies using 18 different tasks (e.g., the engineer–lawyer problem, the maternity-ward problem). All but two of the tasks were solely description-based. Many other description-based problems would follow: the Asian disease problem (Kahneman & Tversky, 1972, 1973), the Linda problem and other conjunction fallacy problems (Tversky & Kahneman, 1983), and described monetary lotteries invoking violations of expected utility theory (Kahneman & Tversky, 1979).

### Why Experimental Protocol Matters

Let us next outline three converging lines of research and theoretical ideas from the 1950s to the present that, taken together, suggest that the change in experimental protocol was consequential.

#### The “Ecological Normal” and Representative Design Perspective

The first line of research returns to Brunswik (1943, 1952) and his theory of probabilistic functionalism. As mentioned earlier, Brunswik held that psychological processes are adapted in a Darwinian sense (Hammond, 1996) to the environments in which they function. The methodological implication (see Dhami et al., 2004) was that stimuli should be sampled from the organism's natural ecology<sup>13</sup> to be representative of the population of stimuli to which the organism has adapted and to which the experimenter wishes to generalize (Brunswik, 1956). Brunswik opposed psychology's accepted experimental design, which he referred to as “systematic” (p. 8). In this design, experimenters systematically vary selected independent variables while holding others constant or allowing them to vary randomly, then observe the resulting changes in the dependent variable(s). This design emphasizes internal validity and therefore risks presenting participants with carefully constructed stimuli that lack realism, having not been drawn from the

---

<sup>9</sup> Kahneman and Tversky were not the first to diagnose base-rate neglect. Laplace (1814/1951) offered what may have been the first account of the base-rate fallacy (Miller & Gelman, 2020). Much later, Meehl and Rosen (1955), in their analysis of the use of psychometric devices in clinical practice, pointed out that diagnostic and clinical predictions can often be made with high accuracy using the base rates of clinical categories. Therefore, “a psychometric device, to be efficient, must make possible a greater number of correct decisions than could be made in terms of the base rates alone” (p. 194), except that “almost all contemporary research reporting neglects the base-rate factor and hence makes evaluation of test usefulness difficult or impossible” (p. 215).

<sup>10</sup> We thank one of the reviewers for raising this point.

<sup>11</sup> For illustration: “Information about base rates is generally observed to be ignored” (Evans & Bradshaw, 1986, p. 16), “it has repeatedly been shown that people commit the base-rate fallacy, that is, they ignore base rate frequencies and, instead, base their judgments solely on the similarity between the individual's personality and the prototypes of the categories under consideration” (Ginossar & Trope, 1987, p. 464), “many (possibly most) subjects generally ignore base rates completely” (Pollard & Evans, 1983, p. 124).

<sup>12</sup> Another probable contributing factor is that judgments (e.g., probability estimates) are perturbed by random error and subject to regression toward the mean (Erev et al., 1994). Depending on the specific base rates and the likelihood information stipulated in a given task, regression toward the mean can result in estimates that are primarily consistent with either base-rate neglect or conservatism (see Erev et al., 1994, 2008, for more details).

<sup>13</sup> Brunswik defined the ecology as the “natural–cultural habitat of an individual or group” (Brunswik, 1955b, p. 198). It consists, among other factors, of reference classes that define populations of stimuli (Brunswik, 1943) that can be drawn upon in experiments.

“ecological normal, located in the midst of a crowd of natural instances” (Brunswik, 1955b, p. 204). The ensuing conclusions about people’s behaviors may lead to systematic mischaracterizations (see Dhimi et al., 2004; e.g., in the context of behavioral decision-making research).

What is the ecological normal of statistical intuitions, and how can it be implemented in the experimental laboratory? Is it better represented by an experiential and learning-focused program or by a description-based program, which offers little opportunity to learn? Animals’, babies’, and young children’s learning is not based on symbolic descriptions (Schulze & Hertwig, 2021a), and professionals such as physicians may never receive explicit, thorough, and systematic instruction in statistics and probability theory (Gigerenzer & Muir Gray, 2011). Many behaviors, both mundane and consequential—crossing the street, falling in love, interviewing for a job—do not have explicitly described probabilities. Yet the world today, more than at any other point in history, is replete with written symbols and described statistics and probabilities, from probabilistic weather forecasts to medication package inserts.

It therefore seems fair to say that both designs represent aspects of the ecological normal that deserve careful study and experimentation (Hertwig, 2015). But it is crucial to also examine them in terms of Brunswik’s notion of representative design. The core premise of representative design is that the informational properties of the experimental task presented to participants represent the properties of the ecology to which experimenters wish to generalize; otherwise, drawing conclusions becomes highly problematic. In other words, one must be wary of generalizing findings and conclusions from a descriptive to an experiential experimental protocol and vice versa. The extent to which such a generalization is legitimate in any specific case must be empirically established, not simply assumed. Campbell (1957) later drew on Brunswik’s emphasis of representative sampling in the service of linking experiments to theory (see the insightful discussion by Albright & Malloy, 2000, of the links between Brunswik, Campbell, and Cronbach on matters of sampling and experimental methodology). As far as we can tell, neither the intuitive-statistician program nor the heuristics-and-biases program was ever concerned with Brunswik’s notion of representative design. This is particularly interesting in the case of the intuitive-statistician program, as Peterson and Beach (1967) explicitly referenced Brunswik’s probabilistic functionalism, which Brunswik saw as inseparably connected to representative design. Furthermore, in many of the bookbag-andpoker-chip problems employed in the studies reviewed by Peterson and Beach (1967), there was little concern for the natural– cultural habitat of the individual.

### The Discrete–Continuous Process Perspective

Also drawing on the Brunswikian notion of the ecological normal, Hogarth (1981) argued that “judgment is part of a continuous, interactive process that people use to cope with the environment, [but] most judgment research has focused on discrete incidents” (p. 197). By “most judgment research” he meant research identifying “systematic dysfunctional consequences of judgmental heuristics” (p. 197), also known as the heuristicsand-biases program. He offered human conversation as an example of a continuous, interactive process. The ability to have a meaningful conversation illustrates the human competence to adjust, craft, and accommodate behavior gradually in response to feedback and experience. Similarly, consider how people walk through a crowded space. They do not and arguably could not plan an optimal path. Rather, they seem to use a set of simple navigation heuristics (Moussaïd, 2019) that enable them to adjust their course continuously to avoid colliding with others, who are also adjusting as they go (e.g., Moussaïd et al., 2011). Continuously updating a single key piece of information (e.g., a visual angle) is also what enables people to catch a high ball (Hamlin, 2017).

In short, judgments are often an interdependent element of a continuous process. People do not have to commit to a fixed path through a crowded square or a specific spot in the field to catch a baseball. Instead, as Hogarth (1981) emphasized, people in an ecologically normal environment learn from feedback and adjust their behavior in response. By focusing on discrete incidents, research on biases resulting from heuristics has led to an underestimation of “the importance of feedback in ongoing processes and the unquestioned acceptance of several assumptions implicit in the discrete, normative models used to evaluate judgmental performance” (p. 197). For Hogarth, a discrete–continuous process perspective was necessary to evaluate the performance and rationality of heuristics.<sup>14</sup>

We suggest that a discrete–continuous process perspective is also necessary to evaluate intuitive statistical cognition. Consider Edwards et al.’s (1965) experimental protocol for studying Bayesian reasoning (see Figure 2). It begins with the toss of a coin to select one of two bookbags, each containing different proportions of red and blue poker chips. Without seeing the outcome of the coin toss, participants respond to a simple question: What is the probability that the chosen bookbag is the predominantly red bookbag? The objective answer is the prior probability of each bookbag being selected: 50%. Participants are then shown a randomly chosen chip from the selected bookbag and asked to give a new estimate. If the chip is red, the estimate should be adjusted upward. Participants make their estimates sequentially; each new chip sampled provides another opportunity to observe, correct, and adjust (without explicit feedback as to the accuracy of the estimate, but with feedback after a full sequence of 12 chips has been completed). Estimates can thus be crafted progressively as experience accumulates, and perceived errors can be corrected with larger adjustments (in either direction). Yet the decision situation is still relatively simple insofar as the environment is stationary and unaffected by the person’s estimates.

Kahneman and Tversky’s (1972) experimental protocol, in contrast, took a description-based approach in which all information is packaged and delivered at once (see Figure 2). In their protocol, participants are asked to judge the probability that a selected deck of cards is marked X based on a full description of a set of random

---

<sup>14</sup> Edwards (1962) proposed a related distinction between static and dynamic decision situations. In dynamic decision situations, the “environment in which the decision is set may be changing, either as a function of the sequence of decisions, or independently of them, or both” (p. 60). In static decision situations, “the decision maker (in principle) never gets to make a second decision in which he might apply whatever he may have learned as a consequence of the first” (p. 59).

draws. They are thus required to consider all the evidence at once and provide a discrete, one-off estimate.<sup>15</sup> Kahneman and Tversky (1973) later abandoned bookbag-and-poker-chip problems, instead devising Bayesian problems with content more familiar from the real world (right column of Figure 2). In the engineer–lawyer problem, for instance, participants receive five written descriptions of fictitious individuals, allegedly drawn at random from a population of 70 lawyers and 30 engineers. They estimate the probability that each person is one of the 70 lawyers (or 30 engineers). But like Kahneman and Tversky’s version of the bookbag-and-pokerchip problem, all relevant pieces of information are described, no feedback is provided, and sequential learning is neither necessary nor possible.

Are there reasons to believe that this difference in protocols is associated with systematically different results? Recent research on what is known as the description–experience gap (Hertwig & Erev, 2009) suggests that this could well be the case.

### The Description–Experience Perspective

People can learn about the properties and statistics of events in at least two ways. Monetary gambles, the paradigmatic tool for measuring risk preferences in psychology and economics, are a case in point. For many years, researchers have typically presented the outcomes and probabilities of each option numerically (e.g., €500 guaranteed vs. €1,000 with .5; €0 with .5) or using a spinner wheel or bar chart (see the meta-analysis by Weber et al., 2002), with all possible outcomes and their probabilities being explicitly stated. This nearly invariant choice architecture is rather odd given that “it is hard to think of an important natural decision for which probabilities are objectively known” (Camerer & Weber, 1992, p. 325). In everyday life, people rarely encounter convenient descriptions of probability distributions (the probability of rain is one exception; e.g., Gigerenzer et al., 2005). For this reason, some researchers began, in the early 2000s, to systematically compare *decisions from description*—risky choices where people were given full descriptions of gambles—with *decisions from experience*—risky choices where people learned about the gamble by drawing samples from the payoff distributions (these can occur with or without feedback; see Hertwig & Erev, 2009). Assuming that people sample sufficiently across the payoff distribution, the information they gain from the choice options in description and experience will converge. Yet samplebased, experiential information will always be associated with more uncertainty than will description. Similarly, descriptions need an “author” (Hertwig et al., 2018), which introduces the problem of whether the information given can be trusted.<sup>16</sup>

Do these two modes of learning about the probabilistic texture of the world (Hertwig et al., 2018) result in equivalent choices? This question has attracted significant attention since three articles (Barron & Erev, 2003; Hertwig et al., 2004; Weber et al., 2004) demonstrated a systematic discrepancy in description- and experience-based choices: the *description–experience gap* (for reviews, see Hertwig, 2015; Hertwig & Erev, 2009; Rakow & Newell, 2010). Recently, Wulff et al. (2018) conducted a meta-analysis of the many ensuing studies on the description–experience gap in risky choice, focusing on the sampling paradigm (i.e., studies with a nonconsequential sampling phase followed by a single consequential choice). They found that decisions from experience and decisions from description are often not equivalent. When a choice was between a risky and a safe option—the choice task often used to behaviorally measure risk preference—the description–experience gap, measured in terms of a systematic difference in the proportions of choices of the two available options, was 18.7%; when a choice was between two risky options, it was 7%. Another indicator of the description–experience gap is the likelihood to choose the option that maximizes expected return. In decisions from description, Wulff et al. (2018; see their Figure 6) found that a median of 55% of choices maximized expected value. In decisions from experience, 66% of people who experienced (via sampling) all possible outcomes and 89% of people who experienced some, but not all, outcomes<sup>17</sup> maximized the experienced mean return—that is, the “expected value” of the experienced sample of outcomes rather than of the objective properties of the options. One possible explanation for the high rate of maximization when not all outcomes were encountered is lower choice difficulty (as a result of an amplification effect; see Hertwig & Pleskac, 2010). Let us also emphasize that in other experimental paradigms studying the description–experience gap, namely those involving repeated choices between the same underlying options with feedback (unlike in the sampling paradigm studied by Wulff et al., 2018), systematic deviations from maximization are often found, and in many cases in the opposite direction to those found in description (Erev et al., 2017; Plonsky et al., 2015). Finally, as another indicator of the description–experience gap, Wulff et al. (2018, Table 1) referred to a reversal of the fourfold pattern of risk preference (Tversky & Kahneman, 1992) between decisions from experience and decisions from description. This reversal suggests that in decisions from description, people choose as if they tend to overweight rare events, whereas in decisions from experience they choose as if they underweight rare events.

Modes of learning about the probabilistic texture of the world also appear to matter in other domains of choice, judgment, and reasoning. For instance, there is evidence for description–experience gaps in intertemporal choice (Dai et al., 2019), social interaction in strategic games (Isler et al., 2020; Martin et al., 2014), ambiguity aversion (Dutt et al., 2014; Güney & Newell, 2015), consumer choice (Wulff et al., 2015), financial risk taking (Lejarraga et al., 2016), medical judgments and decisions (Armstrong & Spaniol, 2017; Fraenkel et al., 2016; Lejarraga et al., 2016; Wegier & Shaffer, 2017), adolescent risk taking (van den Bos & Hertwig, 2017), categorization (Nelson et al., 2010), visual search (Zhang & Houpt, 2020), and causal reasoning (Rehder & Waldmann, 2017).

Several factors may explain the systematic difference in behaviors arising from decisions from description versus decisions from experience. First, a sequence of experience can offer the opportunity to adjust one’s judgment with each observation. Second, the experiential format accommodates heuristics and algorithms that

<sup>15</sup> Kahneman and Tversky (1972) referred to this procedure as “evaluation of evidence” (p. 445).

<sup>16</sup> Decisions from experience have been studied for decades (Edwards, 1961a), but it is only recently that researchers have begun to contrast decisions from experience with decisions from description. It seems that the two modes of learning were previously assumed to be mathematically equivalent and therefore to result in psychologically equivalent representations and behaviors.

<sup>17</sup> Modest sampling can lead to rare events being missed.

are computationally simpler than those afforded by the description format (e.g., natural mean heuristic vs. expected value theory; Hertwig & Pleskac, 2010; Hertwig, Wulff, & Mata, 2019). Third, “experience is concrete as opposed to symbolic, and it has immediate authority for the experiencing individual. It is empirical and rests on the certitude of events that have actually occurred” (Hertwig et al., 2018, p. 124).<sup>18</sup> Fourth, experience can boost people’s memory of the probabilistic texture of the world (Lejarraga, 2010). Fifth, sequential updating in decisions from experience makes it easier to process experienced outcomes (Frey et al., 2015), which is particularly helpful when decision problems become more complex (Hogarth & Soyer, 2015; Lejarraga & Gonzalez, 2011). Finally, there are numerous important circumstances under which experience is a treacherous teacher, causing behaviors arising from experience and description to diverge. Properties of experience such as small or unrepresentative experiential samples (e.g., hot-stove effect; Denrell & March, 2001), ambiguous and incomplete feedback (March, 2010), and the long-lasting impact of extreme experiences (e.g., depression-babies effect; Ludvig et al., 2014; Malmendier & Nagel, 2011) can confound learning from experience. As March (2010) concluded: “Experience may possibly be the best teacher, but it is not a particularly good teacher” (p. 115).

To conclude, at least three lines of research and theorizing—Brunswikian notions of the ecological normal and representative design, the discrete–continuous perspective, and the description–experience gap—suggest that experimental protocol can matter tremendously when studying statistical intuitions. Specifically, the experiential protocol and stimuli used by Edwards et al. (1965) and the descriptive protocol used by Kahneman and Tversky (1972; see Figure 2) may account for the conflicting findings and conclusions.<sup>19</sup> Yet one can only conclude more generally that the experimental protocol explains the diverging findings and conclusions of the two programs if the Edwards et al. (1965) protocol and the Kahneman and Tversky (1972) protocol are representative of each program’s default protocol. If so, respondents’ statistical intuitions would be studied in fundamentally different ways, with the intuitive-statistician program focusing on statistical intuitions from experience and the heuristics-and-biases program focusing on statistical intuitions from description. Was the difference between the two protocols coincidental or representative of research in the respective traditions?

### Quantitative Assessment of the Methodological Protocols

In order to answer this question, we first examined the experimental oeuvre reviewed by Peterson and Beach (1967)—a collection of results that inspired the metaphor of the mind as “intuitive statistician.” Next, we examined the four foundational articles published by Kahneman and Tversky between 1971 and 1973 (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1971, 1973) that were featured in their Science article (Tversky & Kahneman, 1974), which introduced the notion of “heuristics and biases.” A description–experience Figure 3 shows the publication timeline of the articles examined. More precisely, we examined every published article, working paper, and report in Peterson and Beach (1967) that included primary experimental data. We did not include books or full volumes, or work mentioned in the introduction but not covered in their systematic empirical review. We identified 81 articles that reported primary experimental data from a total of 164 experiments. Kahneman and Tversky’s four articles reported 30 experiments; we excluded books, volumes, conceptual articles, discussion articles, and reviews of others’ experimental work. Almost all of these experiments have been regular features in textbooks in psychology and behavioral economics, as well as in countless popular books on human decision making. In this sense, the gestalt of these experiments heralded a new theoretical paradigm.

### Attributes of Experimental Culture

We coded each experiment according to several attributes, including those that reflect the degree to which participants were exposed to an experiential or descriptive experimental microworld. The coding protocol is available in the Supplemental Appendix. All classifications and the resulting dataset are publicly available.<sup>20</sup> The attributes coded were:

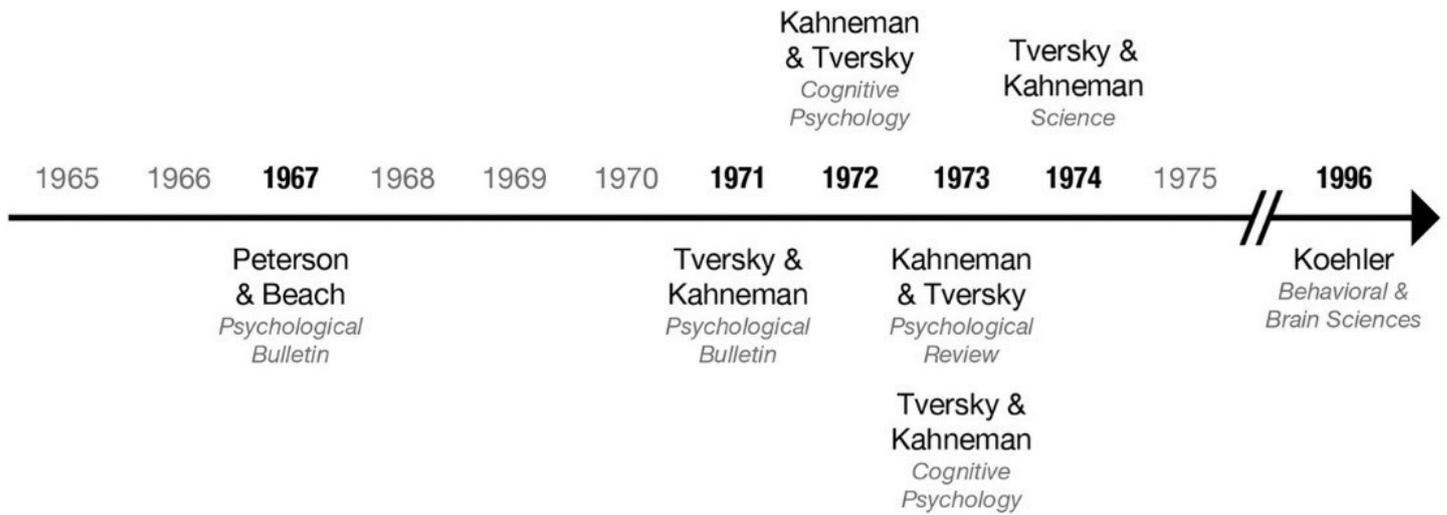
1. Task: Indicates the categories of intuitive statistical reasoning the experiment attempts to study: descriptive statistics, inferential statistics, prediction of samples, or decision making.
2. Participant type: Indicates the type of participants: grade school students, high school students, undergraduate students, graduate students, scientists, professionals, military staff, or general public.
3. Participant expertise: Indicates the participants’ academic discipline, area of expertise, or profession (e.g., business, psychology, economics, statistics).
4. Incentives: Indicates the incentive scheme used: none, fixed payment, or variable payment.
5. Incentive type: Indicates the type of incentive used: money, course credit, or vouchers.
6. Stimulus type: Indicates whether the target stimulus features “people” (e.g., babies, psychologists, engineers, judges, scientists, novelists, pilots) or “objects” (e.g., dice, cards, sounds, lines, drawings, beads, chips, urns)

<sup>18</sup> In his analysis of Bayesian reasoning studies, Koehler (1996) noted that when “base rates are directly experienced through trial-by-trial outcome feedback, their impact on judgments increases” (p. 6)—a finding observed both in the laboratory and in the real world. He speculated that directly experienced base rates may be accorded more weight “because they invoke an implicit rather than an explicit learning system” (p. 7) and that this information may be better remembered, accessed, or otherwise more meaningfully instantiated than information learned explicitly.

<sup>19</sup> When Kahneman and Tversky (1972) abstracted Edwards et al.’s (1965) experiential set-up into an isomorphic text-based representation (see Figure 2) there was no obvious reason, theoretical or empirical, to suspect that this change would have a systematic impact on people’s reasoning and inferences. This was, after all, many years before research on the description–experience gap systematically contrasted the two representations and learning modes.

<sup>20</sup> [https://osf.io/7p692/?view\\_only=e58976ccfe224d7eac90acd12bba4874](https://osf.io/7p692/?view_only=e58976ccfe224d7eac90acd12bba4874)

**Figure 3**  
Chronology of Publication of the Articles Examined in This Quantitative Analysis



discs, games of chance, letters, words, paths, geometric forms).

7. Stimulus representation: Indicates whether the target stimuli in the experiment are physically present (e.g., dice, urns, chips, coins, discs, cards, cards with written letters) or described symbolically in written or audio-recorded form.
8. Practice: Indicates whether the experiment involves practice trials or a training or learning phase.
9. Practice amount: Indicates the amount of practice in terms of number of trials, rounds, responses, problems, games, and so forth.
10. Multiple testing: Indicates whether participants in at least one experimental condition provided a response more than once, because the experiment involved (a) a repeated response procedure, (b) a within-subjects design, or (c) practice or training trials.
11. Response amount: Indicates the number of responses (e.g., answers to questions, ratings, rankings, judgments, inferences, predictions) that each participant gave in the experiment
12. Feedback: Indicates whether participants in at least one experimental condition were given feedback of any kind that allowed them to improve their responses.

Classifying experiments according to these attributes presented some challenges. For instance, some experiments were difficult to understand because the technology is now obsolete, and the materials and procedures were poorly described. In others, it was unclear from the method description where one experiment ended and another followed. For this and other reasons, all experiments were independently coded by the first author of this article and a trained rater.<sup>21</sup> Cases of disagreement were resolved in a face-to-face meeting. The coding protocol is available in the Supplemental Appendix.

### Was There a Profound Transformation in Experimental Culture?

Of the 110 references reviewed by Peterson and Beach (1967), seven were books and 10 were not retrievable.<sup>22</sup> Of the 93 references retrieved, 12 did not report primary experimental data, leaving 81 documents, most of them published articles, but also reports, chapters, unpublished doctoral dissertations, mimeographs, and research bulletins (see details of the screening process in Figure A1 in the Supplemental Appendix). The documents reported 164 experiments, but one experiment was excluded from the analysis because it used rats as subjects. The four key articles published by Kahneman and Tversky reported 30 experiments, giving a total sample of 193 experiments coded (see Figure 4).

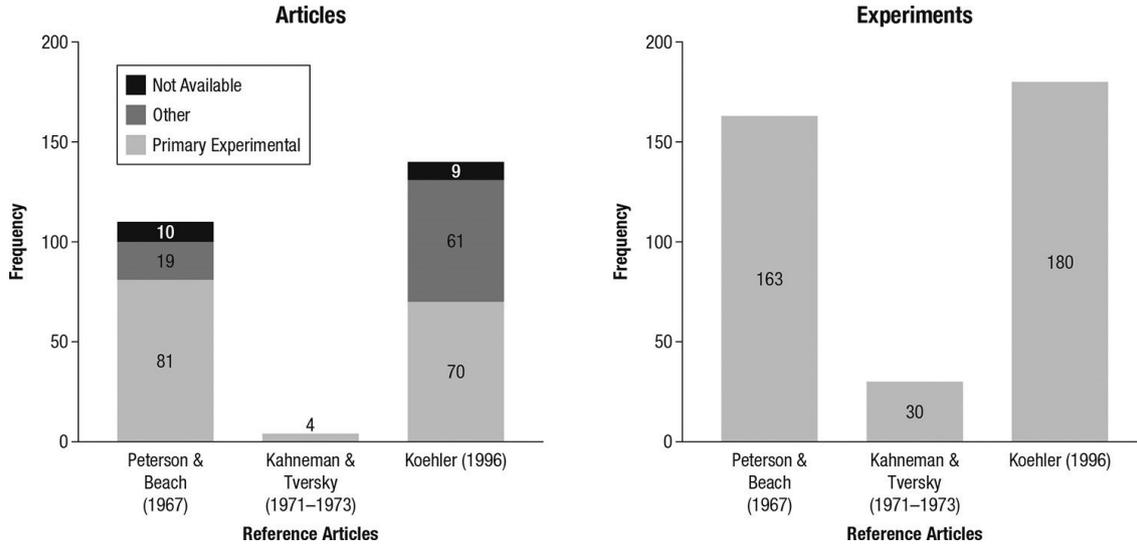
### Same Epistemological Object, Different Experimental Protocol

Was the difference between Edwards et al.'s (1965) experiential protocol and Kahneman and Tversky's (1972) descriptive protocol coincidental or representative for research in the respective traditions (intuitive-statistician vs. heuristics-and-biases)? We investigated this question by considering four attributes of experimental culture that represent the extent to which the protocols were predominantly experiential or description-based: feedback, practice, stimulus representation, and multiple testing. Provision of *feedback* allows participants to improve their responses, offering a kind of learning and experience

<sup>21</sup> In a first round of coding, Jill de Ron (B.Sc. in Psychology, Brain, and Cognition) and Maximilian Rank (B.A. in Philosophy and B.Sc. in Business Administration) served as independent raters during their research internship at the Max Planck Institute for Human Development. In response to problems these raters experienced in classifying the studies, we revised and further clarified the protocol. We also realized that the classification task required substantial experience in reading and understanding Methods sections. For this reason, we discarded the first round of coding and a final round was performed by two PhDs: Tomás Lejarra and Ana Sofia Morais.

<sup>22</sup> These references, all as cited by Peterson and Beach (1967), were three unpublished doctoral dissertations (Azuma; Goodnow; Todd), two mimeographs (Edwards & Kramer; Peterson & Beach), four reports (Dodson; McHale & Stolorow; Rapoport; Shuford & Wiesen), and one conference paper (Schenck & Naylor).

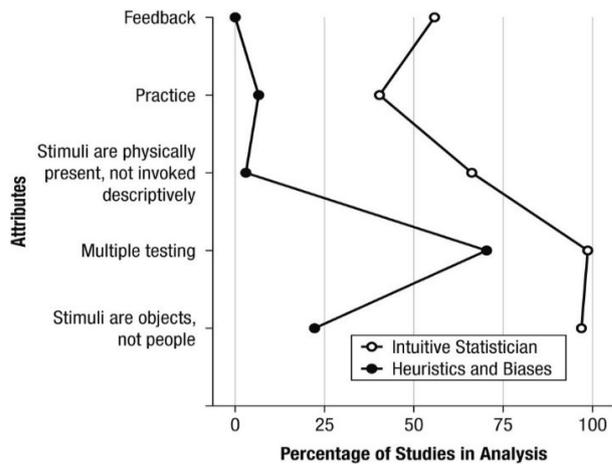
**Figure 4**  
Articles and Experiments Examined in the First Stage of This Quantitative Analysis



Note. "Kahneman and Tversky (1971-1973)" denotes the four articles published by Kahneman and Tversky between 1971 and 1973.

that is otherwise lacking. *Practice* indicates whether participants gained experience in practice trials before the experimental task. *Stimulus representation* concerns whether experimental stimuli were physically present, allowing participants to experience them in a way that is not possible when stimuli are described. *Multiple testing* describes whether an experiment required participants to provide a single response or more; eliciting more than a single response allows participants to gain more experience of the task. Figure 5 reports the coding of the two groups of experiments on these attributes, as well as on another attribute of the stimulus to which we return later: whether it featured objects or people.

**Figure 5**  
Classification of Experimental Protocols According to Five Attributes



Note. Intuitive Statistician: articles reviewed by Peterson and Beach (1967).  
Heuristics and Biases: four articles published by Kahneman and Tversky between 1971 and 1973.

Feedback is an essential aspect of experiential learning. It allows respondents to assess their previous behavior (e.g., choices, inferences, judgments) and adjust their future behavior (Hertwig et al., 2018). The two research traditions differ markedly in the provision of feedback, as Figure 5 shows. None of Kahneman and Tversky's experiments involved feedback, compared with 56% of studies in the intuitive-statistician tradition. But this was not the only striking difference. In the intuitive-statistician tradition, 40% of experiments involved practice, relative to just 7% in the heuristics-and-biases tradition. Most experiments (66%) in the intuitivestatistician tradition employed experimental stimuli that were actually physically present, such as cards, dice, or chips, and sometimes even designed especially for the experiment.<sup>23</sup> Here, practice ensures that respondents comprehend the task. In contrast, only 3% of Kahneman and Tversky's experiments used physical stimuli. Instead, most of their experimental stimuli were described in vignettes or other written representations.

Practically all experiments in the intuitive-statistician tradition (99%) and a substantial proportion in the heuristics-and-biases tradition (70%) implemented multiple testing. Note that our criterion for classifying experiments as using multiple tests was intentionally lenient, namely, that participants "provide a response more than once." Therefore, experiments classified as implementing multiple testing could range from a short vignette study requiring two responses to two questions (e.g., the replication study administered to psychologists by Tversky & Kahneman, 1971) to a probability learning study requiring 1,000 predictions per person (as in Edwards, 1961b). Indeed, the median number of responses in the heuristics-and-biases tradition was 3.0 ( $M = 8.8$ ,  $SD = 18.7$ ) while the median number in the intuitive-statistician tradition was 77.5 ( $M = 417.8$ ,  $SD = 971.8$ ).

Figure 5 includes another attribute that at first glance appears somewhat orthogonal to description versus experience: whether

<sup>23</sup> This proportion excludes 12 studies in the intuitive-statistician tradition that were coded as involving stimuli that were both physically present and invoked descriptively.

the target stimulus featured objects (e.g., dice, balls) or people. Yet it seems plausible that the change from physically presenting inanimate stimuli to describing people opened the door to a range of experimental scenarios that would otherwise be difficult to present. Excluding a small subset of studies that included both people and objects, we found that 78% of experiments in the heuristics-and-biases tradition referred to people, relative to only 3% in the intuitive-statistician tradition.<sup>24</sup> It is likely that introducing scenarios (e.g., a hit-and-run accident in the cab problem; Tversky & Kahneman, 1982), people and their personalities (e.g., engineer-lawyer problem; Kahneman & Tversky, 1973; Tversky & Kahneman, 1983), and famous people (e.g., availability heuristic; Tversky & Kahneman, 1973) made the experimental tasks more engaging than many of the experiments in the intuitive-statistician tradition that relied on physical objects.

To conclude, in the experimental protocol used in the intuitivestatistician tradition, people were afforded the opportunity to learn from practice and multiple testing, to interact physically with the experimental materials, and to learn from feedback.<sup>25</sup> In contrast, the heuristics-and-biases approach was mostly description-based, with descriptions of people rather than objects, opportunity to practice was rare, and learning from feedback not possible. These findings confirm that Tversky and Kahneman's experimental protocol abandoned learning (feedback) and the repeated measurements characteristic of the intuitive-statistician tradition. We next examine four further attributes that separate the two traditions: the type of task (category of statistical reasoning studied), the type of participants, their expertise, and the provision of incentives.

### Differences in Task Type, Incentives, and Participant Type and Expertise

An alternative or at least additional explanation for the divergent conclusions of the intuitive-statistician and the heuristics and-biases traditions is that they investigated different psychological phenomena. To examine this possibility, we classified Tversky and Kahneman's foundational articles using the categories laid out in Peterson and Beach's (1967) review: intuitive descriptive statistics (i.e., how the intuitive mind estimates proportion, mean, variance, correlation, etc.), intuitive inferential statistics (i.e., the use of samples of data to reach conclusions about characteristics of the environment, e.g., population parameters), and intuitive predictions of samples (i.e., about events to be sampled from populations). Peterson and Beach used a fourth category on nonstationary parameter values (i.e., the study of judgments, inferences, predictions, and decisions in environments that change over time). Because this additional category revisits judgments, inferences, and predictions, and adds research on decision making, we created a "decision making" category to include documents with a focus on decision making. We found that all studies by Kahneman and Tversky could be accommodated by Peterson and Beach's (1967) four categories. Yet there was high mismatch between raters that was not easily resolved. The problem for raters was how to interpret the tasks. For illustration, consider the Judgments of Word Frequency study (Study 3 in Tversky & Kahneman, 1973, pp. 211–212). Participants were given the following instruction:

The frequency of appearance of letters in the English language was studied. A typical text was selected, and the relative frequency with which various letters of the alphabet appeared in the first and third positions in words was recorded. Words of less than three letters were excluded from the count.

You will be given several letters of the alphabet, and you will be asked to judge whether these letters appear more often in the first or in the third position, and to estimate the ratio of the frequency with which they appear in these positions.

One question facing the raters was whether the estimates participants were asked to produce pertained to the positional frequencies within the "typical text" (established in the instructions) or in the English language, or whether they were asked to infer from the positional frequencies in the text to those in English language. Depending on how the raters interpreted the instructions, this task could be placed in several of the Peterson and Beach categories.<sup>26</sup>

Although such ambiguities made a clear distinction between categories difficult, all of Kahneman and Tversky's studies could be accommodated in Peterson and Beach's categories of intuitive descriptive statistics, intuitive inferential statistics, and intuitive prediction of samples. Another indication that the two research programs largely covered the same ground is that Kahneman and Tversky (1972) themselves judged that their work spoke directly to Peterson and Beach's conclusions: "[f]or anyone who would wish to view man as a reasonable intuitive statistician such results are discouraging" (p. 445). In short, although the two traditions looked at the world differently, they looked at the same referents in the world.

Another potential explanation for the divergent findings of the two programs is their use of incentives. It is possible that monetary incentives promote more accurate intuitive statistical reasoning, whereas a lack of incentives leads to more biases and errors (a hypothesis that was entertained for some time by experimental economists such as Grether & Plott, 1979; see Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001). Yet our findings showed that the frequency of use of incentives was almost identical across the two research programs, with incentives being used in 37% of studies in the intuitive-statistician program and 33% of studies in the heuristics-and-biases program. In the intuitivestatistician program, 20% of experiments had a fixed incentive scheme while the rest used a variable incentive scheme or a mix of the two; in the heuristics-and-biases tradition, 27% used a fixed incentive scheme. Thus, a greatly divergent frequency of incentives does not appear to explain the contrasting conclusions of the two programs.

Finally, we examined whether participant type and expertise explained the divergent results. Most experiments in the intuitive-

<sup>24</sup> These proportions exclude two studies in the intuitive-statistician tradition and three studies in the heuristics-and-biases tradition that were coded as involving both people and objects.

<sup>25</sup> Outside the domain of statistical reasoning, Edwards introduced a descriptive aspect into experimentation in behavioral decision research. Edwards (1953, 1954a) studied choice with lottery options in terms of decisions from description that were devoid of any learning; for instance, "If you roll a 4, you win \$6.00. If you roll anything else, you win nothing" (Edwards, 1954a, p. 59). Note that the gamble was physically implemented on a pinball machine, in contrast to many other risky choice studies in the decisions-from-description tradition.

<sup>26</sup> As one reviewer pointed out, this ambiguity also existed for participants.

statistician program had undergraduate students as participants (64%). The heuristics-and-biases tradition reported using undergraduates in only 7% of studies, but a further 43% of studies used people recruited at universities who had an unspecified level of education. Experiments with high school and grade school students were more prevalent in the heuristics-and-biases tradition (27%) than in the intuitive-statistician tradition (2%), but so were experiments with graduate students (13% vs. 4%, respectively). A minority of studies in the intuitive-statistician tradition used samples of military staff (seven studies), medically diagnosed schizophrenics (three studies), professionals (two studies), and prison inmates (one study). One study in the heuristics-and-biases tradition used scientists. Overall, there is no clear discrepancy between the types of participants that would suggest participant population to be a crucial determinant of the diverging findings and conclusions.

Likewise, the coding of participant expertise did not differentiate between the two programs. The majority of studies did not report participants' expertise (69% and 90% in the intuitive-statistician and heuristics-and-biases traditions, respectively). In those that did report the field of study of the undergraduate or graduate participants, most were psychology students (25% and 10%, respectively).

In summary, we found no systematic differences in task type, incentives, or participant type or expertise that could explain why experimenters from the two research programs appeared to have seen different worlds.

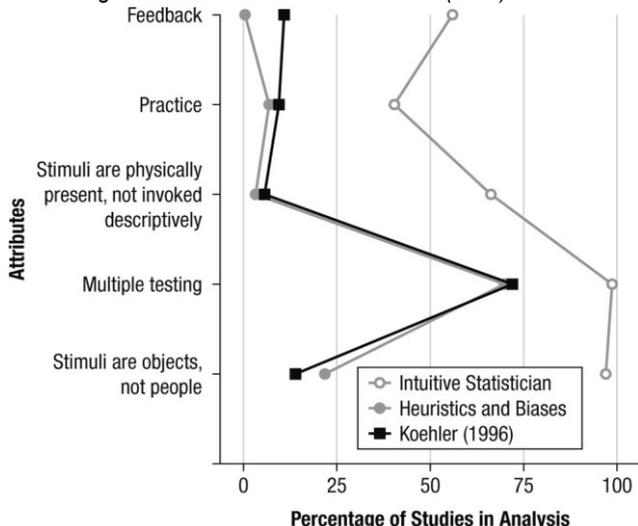
### Did the New Experimental Culture Become the Default?

To examine the possibility that the experimental protocol adopted by Kahneman and Tversky became the default protocol in future studies on statistical thinking, we extended our analysis to research conducted after 1974. As Figure 1 suggests, research on behavioral decision making increased enormously after Tversky and Kahneman (1974). We focused our initial analysis on Bayesian reasoning.

To examine how heavily experiments have relied on descriptive protocols, we drew on the comprehensive review of Bayesian reasoning published by Koehler (1996) in *Behavioral and Brain Sciences*. Koehler posed key empirical questions about Bayesian reasoning (e.g., "Are base rates ignored?," "When are base rates used?") and reviewed all the literature available until 1996 that could provide answers to each question. For each article that met our inclusion criteria, we conducted the same quantitative analysis of experimental practices as reported above (see the Supplemental Appendix for details). Of the 152 references reviewed in Koehler (1996), we included 70 in our analysis: those published after 1974 (in the wake of Kahneman and Tversky's foundational articles; Figure 3) whose primary focus was on experimentation on Bayesian reasoning and that we were able to retrieve (see Figure A1 in the Supplemental Appendix for more details).

Figure 6 shows the results. The experimental protocols in the post-1974 studies on Bayesian reasoning were strikingly similar to those introduced by Kahneman and Tversky. They rarely provided the opportunity for practice and feedback, and there was little repetition. For example, the median number of responses in the articles reviewed by Koehler (1996) was 7.0 ( $M = 42.2$ ,  $SD = 87.1$ )—similar to the median of 3.0 in the heuristics-and-biases tradition and far fewer than the median of 77.5 in the intuitive-

**Figure 6**  
Classification of Experimental Protocols in Post-1974 Bayesian Reasoning Research as Reviewed in Koehler (1996)



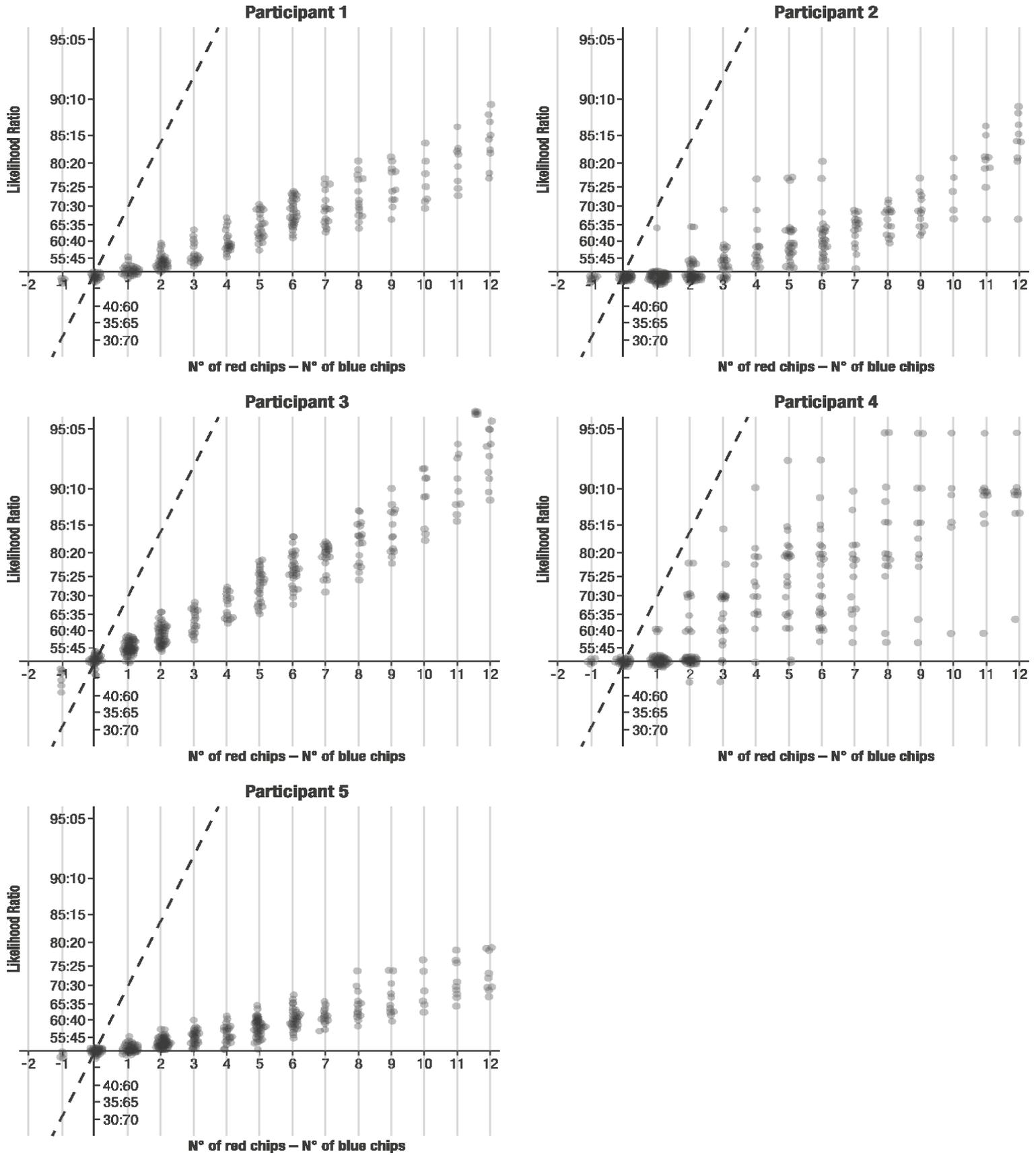
Note. For comparison, the results reported in Figure 5 are shown in gray.

statistician tradition. The descriptive protocol developed in the foundational studies of the heuristics-and-biases program has become, at least in the field of Bayesian reasoning, the default protocol in the time window covered by Koehler's review.

### Does Experience Foster Statistical Competence in Bayesian Reasoning?

One would expect experiential protocols with opportunities for learning to be more conducive to accurate statistical intuitions than descriptive protocols with no such opportunities. Is there evidence for this in studies on Bayesian reasoning? Comparing the accuracy of Bayesian reasoning in the two research traditions is not trivial (see Figure 2). The paradigm shift was so fundamental that even the ways of producing, analyzing, and reporting data changed drastically, making any comparisons almost impossible. For example, Edwards et al. (1965) were concerned with people's ability to update probabilities. Their experiments involved a few participants, many judgments, and one or a few experimental treatments. Kahneman and Tversky (1972), in contrast, aimed at uncovering the psychological mechanisms underlying probabilistic inference. They devised experiments to tease apart mechanisms, creating set-ups where different hypothesized mechanisms predicted different judgments. Thus, most of their experiments were between-subjects designs, involving many participants and no more than a few responses. One consequence of these distinct goals and methodological choices is that behavior is reported using different metrics. For example, Edwards et al. (1965; see left column of Figure 2) measured competence by comparing the individual likelihood ratio obtained from each participant's expressed probabilities with that derived from Bayes's rule. These likelihood ratios were plotted in a graph, reproduced in Figure 7, as a function of the number of red minus blue chips—that is, experienced

**Figure 7**  
*Estimates of Five Participants in Edwards et al.'s (1965) Bayesian Reasoning Task*



*Note.* Each graph depicts estimates of a single individual (expressed in log likelihood ratios) across situations of increasing evidence favoring the predominantly red bookbag over the predominantly blue bookbag (i.e., the difference between red and blue chips experienced by the subject). Estimates are expressed in log likelihood ratios because Bayesian performance is represented by the straight dashed line. "This makes it exceptionally convenient to compare actual performance with ideal performance" (Edwards et al., 1965, p. 305). Still, the gap between actual performance and ideal performance expressed as posterior probabilities, instead of log likelihoods ratios, is not as pronounced as these graphs seem to suggest (see the Supplemental Appendix for details). Adapted from "Emerging Technologies for Making Decisions," by W. Edwards, H. Lindman, and L. D. Phillips, in F. Barron, W. C. Dement, W. Edwards, H. Lindman, L. D. Phillips, J. Olds, and M. Olds (Eds.), *New Directions in Psychology* (pp. 304–308). 1965, Macmillan Publishers. Copyright 1965 by Macmillan Publishers. Adapted with permission.

successes minus failures—individually for each of the five participants.

Although Kahneman and Tversky's (1972) task was very similar in structure to that of Edwards et al. (1965), their experimental protocol and analysis were profoundly different. Kahneman and Tversky explored whether posterior estimates respond to the difference between successes and failures (red and blue chips)—as prescribed by Bayes's rule for this symmetric binomial case—or whether they incorrectly depend on the sample ratio between successes and failures. Thus, as shown in Figure 8, they reported average estimates across (approximately 56) participants<sup>27</sup> for each experimental treatment. It is evident that comparing the results of Edwards et al. and Kahneman and Tversky is difficult and potentially misleading: Figure 7 plots hundreds of responses for each individual participant as a function of increasing sample differences. Figure 8 reports one posterior estimate averaged across all participants, separately for each of the 10 experimental treatments.

It is therefore not possible to determine whether experiential or descriptive protocols foster more accurate Bayesian reasoning by comparing the original studies. Nonetheless, we can make use of two other sources of data. First, in his review of Bayesian reasoning studies, Koehler (1996) paid particular attention to experiments where participants could learn from feedback. He observed that “when base rates are directly experienced through trial-by-trial outcome feedback, their impact on judgments increases” (p. 6), thereby increasing participants' ability to update probabilities in a manner consistent with Bayes's rule. For instance, in the experiment by Manis et al. (1980), “the influence of the feedback on base rate usage was quick and dramatic” (Koehler, 1996, p. 6). Furthermore, in a problem in which participants had to predict whether a person shown in a photograph was for or against the use of seatbelts, participants who learned from experience that the base rate of “for-seatbelts” was 80% matched that percentage in their predictions. Those who did not receive feedback predicted “for-seatbelts” in roughly 50% of cases. Similarly, Butt (1988) showed that “auditors learned and used the base rate for financial statement errors most easily by directly experiencing those errors” (Koehler, 1996, p. 6). Finally, Christensen-Szalanski and Bushyhead (1981) showed that physicians who learned the low base rate for pneumonia from experience “relied heavily on this base rate when making diagnoses” (Koehler, 1996, p. 6), thus improving their diagnostic skill. Koehler's (1996) conclusion was that “information that is learned implicitly [from experience] may be better remembered, more easily accessed, or otherwise more meaningfully instantiated than information learned explicitly” (p. 7).

The second source of data comes from two recent studies that systematically compared Bayesian reasoning across description and experience. Armstrong and Spaniol (2017) studied Bayesian reasoning in the context of medical screening. They started from the common observation that laypeople and experts are poor at inferring posterior probabilities when the relevant statistics are communicated descriptively. Would learning those statistics from experience improve their performance? The study presented 80 younger (aged 17–34 years) and 80 older (aged 65–87 years) adults with information about medical screening tests for two hypothetical diseases in either a description or an experience format (see Figure 9). In the description condition, participants read a passage containing statistical information. In the experience condition, they learned about the statistics sequentially from a slideshow presenting representative cases. Figure 10 (left panel) shows the results. Both younger and older adults arrived at substantially more accurate posterior probability estimates in the experience than in the description condition, where errors were about five times larger (with errors computed as the absolute difference between estimated and true positive predictive value, and the absolute difference between estimated and true negative predictive value).

Schulze and Hertwig (2021b) likewise studied Bayesian reasoning and compared the impact of description and experience on posterior probability estimates. Their results are consistent with those of Armstrong and Spaniol (2017; see right panel of Figure 10). In the description condition, adults correctly estimated the posterior probability in 18% of responses (averaged across participants and problems), in the experience condition, in 79% of responses. The same pattern of results emerged for conjunction rule problems and for children's statistical intuitions. These two studies thus establish a causal link between the mode of learning and presentation and statistical competence.

These results converge to the conclusion that, all other things being equal, statistical intuitions might be more accurate in experiential than in descriptive protocols. This conclusion is not limited to Bayesian inference. Table 2 lists recent studies that have explicitly contrasted descriptive with experiential learning across a variety of choice and inference tasks. A regularity emerges across these studies: Experience by itself or when combined with description (as in, e.g., Erev et al., 2017) can reduce and sometimes even eliminate the judgment bias observed in descriptive conditions. Let us emphasize, though, that experience is not devoid of biases; it may sometimes reverse a bias and may sometimes cause or amplify a bias (we return to this issue below).

### Did the Methodological Shift Occur Beyond Bayesian Reasoning?

Due to its pivotal role in both the intuitive-statistician and the heuristics-and-biases program, Bayesian reasoning constituted a good testbed for our argument that a methodological shift occurred. We now turn to three other lines of research. We first consider probability estimates of compound events as a topic addressed in both programs; through the heuristics-and-biases program, the conjunction fallacy—most prominently illustrated by the Linda task—eventually became a cause célèbre in the context of erroneous statistical intuitions. Next, we consider two lines of research spurred by Kahneman and Tversky's work: research on anchoring and adjustment and on framing. We employ the same analytical framework that we used for Bayesian reasoning.

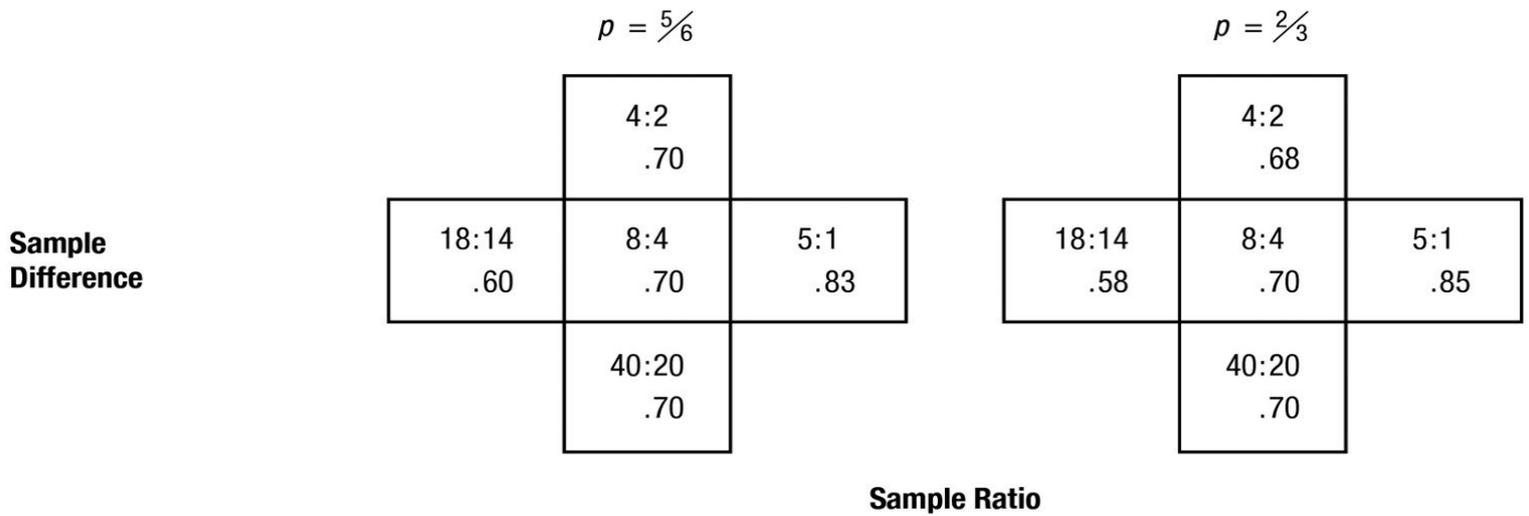
To identify relevant articles, we drew on research reviews. Specifically, to investigate the impact of the heuristics-and-biases program on research on probability estimates of compound events, we examined all work reviewed by Fisk (2004) on the conjunction fallacy, focusing on work published after Tversky and Kahneman's (1983) seminal article in *Psychological Review*. We identified 19 published articles and 43 experiments. For the analysis of anchoring and adjustment, we turned to the

<sup>27</sup> The number of participants in this study is approximate. It was described as follows: “Each of the 10 problems was presented to a different group of S[ubject]s. Size of group varied from 37 to 79 with an average of 56” (Kahneman & Tversky, 1972, p. 447).

**Figure 8**  
*Participants' Estimates in Kahneman and Tversky's (1972) Bayesian Reasoning Task*

Subjective Posterior Probability in a Symmetric Binomial Task for Two Pairs of  
 Populations:  $p = \frac{5}{6}$  and  $p = \frac{2}{3}$

The upper entry in each cell is the sample presented; the lower entry is the median subjective estimate.



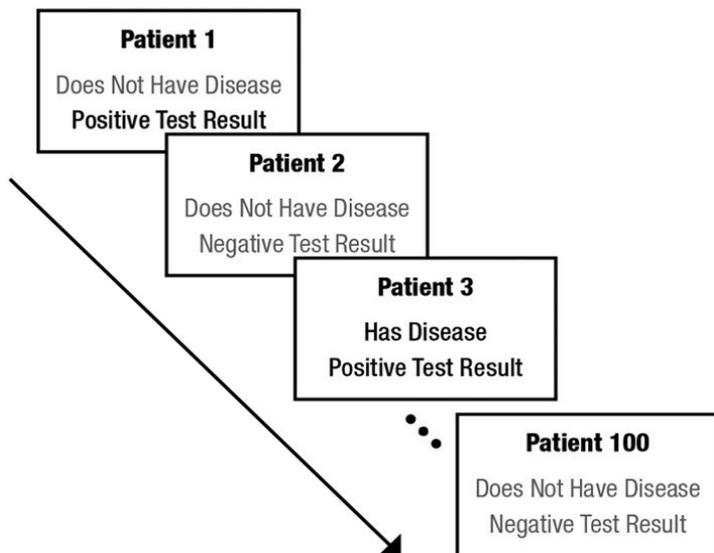
Note. Each cell entry reflects the median subjective estimate across approximately 56 participants. Cells reflect different Bayesian problems with distinct population proportions ( $\frac{5}{6}$  or  $\frac{2}{3}$ ) and evidence (4:2, 8:4, etc.). Base rates were symmetric for each deck. This was conveyed to participants as "One of the decks has been selected by chance." From "Subjective Probability: A Judgment of Representativeness," by D. Kahneman and A. Tversky, 1972, *Cognitive Psychology*, 3(3), p. 448. Copyright 1972 by Elsevier. Reprinted with permission.

literature review conducted by Furnham and Boo (2011). We examined all articles published after Tversky and Kahneman (1974) and included in the review's summary tables 1, 2, and 3. We identified 31 published articles and 96 experiments. Finally, to examine the experimental protocols in research on framing, we turned to Kühberger's (1998) meta-analysis. We focused our analysis on articles published up to 10 years after the publication of Tversky and Kahneman's (1981) *Science* article on framing, including 59 published articles and 92 experiments.

The classification of experiments (see Figure 11) shows that the impact of the heuristics-and-biases program on experimental protocols was not limited to Bayesian reasoning. Research on the conjunction fallacy, anchoring and adjustment, and framing was largely description based, with most experiments using symbolic stimuli about people and eschewing practice and feedback. The median numbers of responses

**Figure 9**  
*Experimental Material Used by Armstrong and Spaniol (2017) in the Experience and Description Formats*

**Experience Format**



**Description Format**

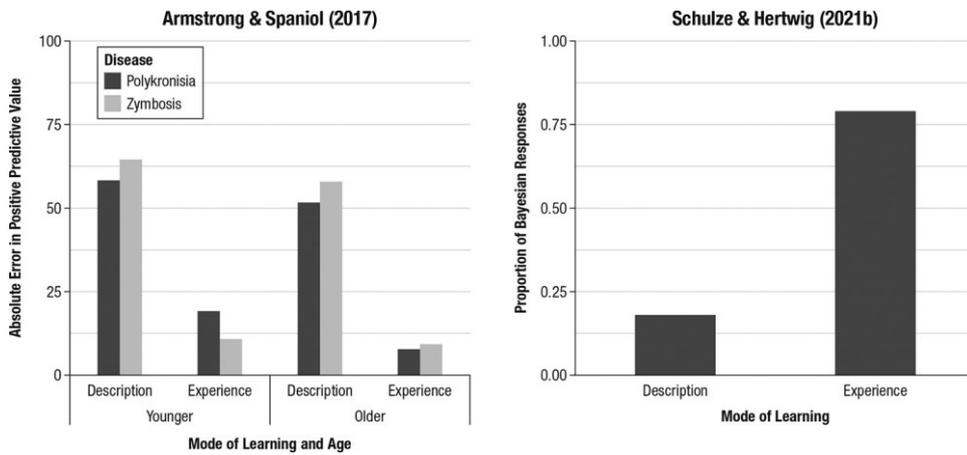
2 out of every 100 people have polykronisia.

If a person has polykronisia, 100 out of every 100 such people will have a positive result from the blood test.

If a person does not have polykronisia, it is still possible that he or she will have a positive test result from the blood test. More precisely, 8.16 out of every 100 people will have a positive result from the blood test.

Note. Modified from "Experienced Probabilities Increase Understanding of Diagnostic Test Results in Younger and Older Adults," by B. Armstrong and J. Spaniol, 2017, *Medical Decision Making*, 37(6), p. 674. Copyright 2017 by Sage. The upper-left box in the original figure stated, "Patient 2." We believe this was in error and have changed it to "Patient 1."

**Figure 10**  
**Findings Comparing the Impact of Description and Experience on Bayesian Reasoning**



Note. In Armstrong and Spaniol (2017), the measure of competence expressed in the y-axis is the absolute error in positive predictive value—the posterior probability that the patient has the disease given a positive test result. Higher bars indicate lower competence. In Schulze and Hertwig (2021b), competence is measured as the proportion of Bayesian responses. Higher bars indicate higher competence.

in the studies focusing on the conjunction fallacy, anchoring and adjustment, and framing were 16.0 ( $M = 25.9$ ,  $SD = 27.4$ ), 7.0 ( $M = 11.5$ ,  $SD = 14.9$ ), and 4.5 ( $M = 16.8$ ,  $SD = 46.1$ ), respectively, comparable to the median of 3.0 in Kahneman and Tversky's (1974) foundational work and in stark contrast to the median of 77.5 in Peterson and Beach's (1967) review. While these phenomena have largely been studied using descriptive experimental methods, they can also be studied with a more experiential approach (for framing effects from experience, see Gonzalez & Mehlhorn, 2016; for an experiential approach to probability inference, including the conjunction rule, see J. Cohen et al., 1971; Hogarth & Soyer, 2011; Peterson, Ulehla, et al., 1965).

## Summary

Kahneman and Tversky's work has transformed the behavioral sciences, the understanding of people's decision-making competences and rationality (or lack thereof), and evidence-based public policy making (Thaler & Sunstein, 2008). The persuasive nature of the heuristics-and-biases program has pushed aside the extensive body of findings on statistical intuitions prior to the 1970s. We analyzed one key variable that is likely to have made a crucial contribution to the swift transformation that occurred in the 1970s and 1980s. As our analysis of a total of 604 studies on Bayesian reasoning, the conjunction fallacy, anchoring and adjustment, and framing has shown, Kahneman and Tversky's research seems to have initiated a little-noticed but profound change in the experimental culture of behavioral decision making research—from learning and experiential involvement to simple, engaging, and symbolic descriptions of microworlds. Indeed, the word "learning" does not appear in the index of Kahneman's (2011) bestseller *Thinking, Fast and Slow*, which covers nearly four decades of research in the heuristics-and-biases tradition. This seismic methodological shift is likely to have had far-reaching implications for empirical observations about human rationality and the subsequent conclusions.

## General Discussion

### Experimental Protocols and Debates on Rationality in the Behavioral Sciences

The heuristics-and-biases program prompted dire conclusions about the human mind and its rationality. Whereas Kahneman and Tversky (1972) were restrained in how they framed their message, writing that their results were "discouraging" for anyone who wished to view people as intuitive statisticians (p. 445), other researchers went further. According to Slovic et al. (1976), the results suggested that people "lack the correct programs for many important judgmental tasks" (p. 174). Piattelli-Palmarini (1994) held that people fall victim to "inevitable illusions" when they reason about probability; indeed, he suggested a simple law: "Any probabilistic intuition by anyone not specifically tutored in probability calculus has a greater than 50% chance of being wrong" (p. 132). Bazerman and Neale (1992) maintained that "all executives have pervasive decision-making biases" (p. 2), and paleontologist and evolutionary theorist Gould (1992) conjectured that human minds "are not built (for whatever reason) to work by the rules of probability" (p. 469).

What is striking about these and many other statements is how little they square with the conclusions about people's statistical reasoning listed in Table 1, and with Peterson and Beach's (1967) summary that the "normative model provides a good first approximation for a psychological theory of inference" (p. 42). Yet despite dominating behavioral decision research since the 1970s, the conclusions of the heuristics-and-biases program have not remained unchallenged. For instance, L. J. Cohen (1981) questioned whether human rationality could ever be experimentally demonstrated, based on the argument that any normative theory of reasoning is simply an idealized theory derived from people's actual individual intuitions about reasoning. Therefore, human reasoning competences will necessarily be in accord with the norms (see also Schurz & Hertwig, 2019; Stich, 1985). From this perspective, reasoning errors do not reflect lack of

**Table 2**  
*Evidence From Research Contrasting Descriptive With Experiential Learning: Experience Tends to Reduce Judgment Bias*

Bias	Article	Conclusion	Effect
Errors in Bayesian inference	Armstrong and Spaniol (2017)	“... we found a significant format effect on estimation errors, which were significantly smaller in the experience format, compared with the description format” (p. 677).	In positive predictive value estimation, errors were 58.11% in the description format, relative to 12.64% in the experience format. In negative predictive value estimation, errors from description were 22.87%, and errors from experience were 6.3%.
	Schulze and Hertwig (2021b)	“... the odds of correctly solving all Bayesian inference problems were 32.33 times higher in experience than in description” (p. 13).	The mean proportion of logical responses in adults was 33% in description and 87% in experience.
	Hogarth and Soyer (2011)	“Our experiments demonstrate that probabilistic judgments can be more accurate when based on experience in the form of sequentially simulated outcomes as opposed to description” (p. 447).	Correct responses were 18.5% in description and 98.5% in experience.
Errors in probability judgments (birthday problem)	Hogarth and Soyer (2011)	“Experience made a dramatic difference. Whereas the actual percentage correct was less than in other problems, the answers of a clear majority were close to correct” (p. 442).	Correct responses were 1.5% in description and 60% in experience.
Loss aversion (Kahneman & Tversky, 1979)	Yechiam and Hochman (2013)	“The studies of experiential tasks indicate no asymmetric effect for losses, as the average decision maker was indifferent between alternatives that incur losses and gains and alternatives that incur either losses or gains of lower magnitude or zero ...” (p. 503).	Loss aversion was observed in four of 11 articles on decisions from description, but in none of 13 articles on decisions from experience.
Ambiguity aversion (Ellsberg, 1961)	Dutt et al. (2014)	“... participants are ambiguity-seeking in experience and ambiguity-averse in description in problems involving both gains and losses” (p. 316).	The size of the mean description–experience gap was 26%.
	Ert and Trautmann (2014)	“The current paper finds that such sampling experience reverses the pattern of ambiguity attitude observed in the static case” (p. 31). “Also noteworthy is the elimination of ambiguity aversion under the moderate probability (problem 0.5) after sampling” (p. 35).	E.g., in a problem with a 0.5 probability in the risky and ambiguous options, choices for ambiguity aversion amounts to 32% in description and to 2% in experience.
	Güney and Newell (2015)	“Overall, the results indicated that people demonstrated ambiguity-neutral attitudes when the underlying probability distributions were experienced” (p. 188).	Mean ambiguity aversion was 15% across different description conditions and 1% across different experience conditions.
	Erev et al. (2017)	“Evaluation of the effect of experience reveals that feedback eliminates these attitudes toward ambiguity” (p. 377).	Across three problems, ambiguity aversion averaged 33% in decisions from description, and 1% after experience.
Choice-pricing preference reversal	Golan and Ert (2015)	“The results show that the mode of acquiring information affects pricing: the tendency to underprice high-probability prospects and overprice low-probability ones is diminished when pricing is based on experience rather than description” (p. 9).	Mean prices deviated from expected value 94% in description and 28% in experience.
Certainty effect (Kahneman & Tversky, 1979)	Erev et al. (2017)	“... feedback reduced this difference [in choice proportions that reflect the certainty effect]” (p. 375).	The certainty effect was 19% in decisions from description and 2% in decisions from experience.
Reflection effect (Kahneman & Tversky, 1979)	Erev et al. (2017)	“Feedback reverses the results and leads to lower risk-taking rate in the loss domain ... than in the gain domain” (p. 375).	In one demonstration, the reflection effect decreased from 23% in description to 7% in experience. In another demonstration, the reflection effect reversed, from 7% in description to 20% in experience.
Conjunction fallacy (Tversky & Kahneman, 1983)	Hogarth and Soyer (2011)	“The analytic [description-based] responses to the conjunction problem were somewhat dispersed. But, experience made a difference that was largely maintained by all groups in their final responses” (p. 442). “The analytic [description-based] and experience-based responses were generally opposites for all groups (incorrect and correct, respectively)” (p. 442).	Correct responses were 25% in description and 84% in experience.

(table continues)

**Table 2** (continued)

Bias	Article	Conclusion	Effect
Insensitivity to sample size	Hogarth and Soyer (2011)	“Experience led to almost 100% correct responses for the hospital problem . . .” (p. 442).	Correct responses were 32% in description and 98.5% in experience.
Regression toward the mean	Hogarth and Soyer (2011)	“The effect of experience was to shift answers to being more correct. However, at the final stage, the naïve group is not convinced” (pp. 442–443).	Correct responses were 28.5% in description and 69% in experience.
Overconfidence by over-precision (Moore & Healy, 2008)	Camilleri and Newell (2019)	“Across three experiments, we found that those learning from description tended to be overprecise whereas those learning from experience were underprecise. These differences were driven by a relatively better calibrated representation of the underlying outcome distribution by those presented with experience-based information. We argue that those presented with experience-based information have better learning due to more opportunities for prediction-error” (p. 62).	In Experiment 1, hit rate from description was 74% and hit rate from experience was 94%. In Experiment 2, it was 89% in description and 95% in experience.
Overweighting of small probabilities (Kahneman & Tversky, 1979)	Golan and Ert (2015)	“... in the pricing from description condition, people exhibited risk seeking for low-probability prospects and risk aversion for high-probability prospects, a pattern that reflects “overweighting rare events”. In pricing from experience, however, the results suggested risk neutrality for both low-probability and high-probability prospects” (p. 11).	The mean description–experience gap in pricing decisions was 25.5%.
	Erev et al. (2017)	“Experience reduces sensitivity to the rare event in all five problems” (p. 375); “. . . the emergence of underweighting of rare events in decisions with feedback is robust” (p. 375).	In the problems designed to examine overweighting of rare events, 56% of decisions from description and 45% of decisions from experience were consistent with overweighting of rare events.
	Wulff et al. (2018): A meta-analysis encompassing 80 datasets and 4,400 participants	“Figure 6 shows that in decisions from experience a large proportion of choices are consistent with maximization of the experienced mean return. This, in turn, is consistent with either linear weighting of probabilities or, more radically, with no weighting of probabilities at all” (p. 159).	“As Figure 6 shows, in decisions from description, a median of 55% of choices maximized expected value; in decisions from experience, 66% and 89% maximized the experienced mean return (depending on whether a sequence did or did not include all possible outcomes)” (p. 149).
Decoy effects (Huber et al., 1982)	Ert and Lejarraga (2018)	“Indeed, in two experiments we found that decoy effects did not emerge in decisions from experience” (p. 542).	The decoy effect was 14% in decisions from description and 3% in decisions from experience.

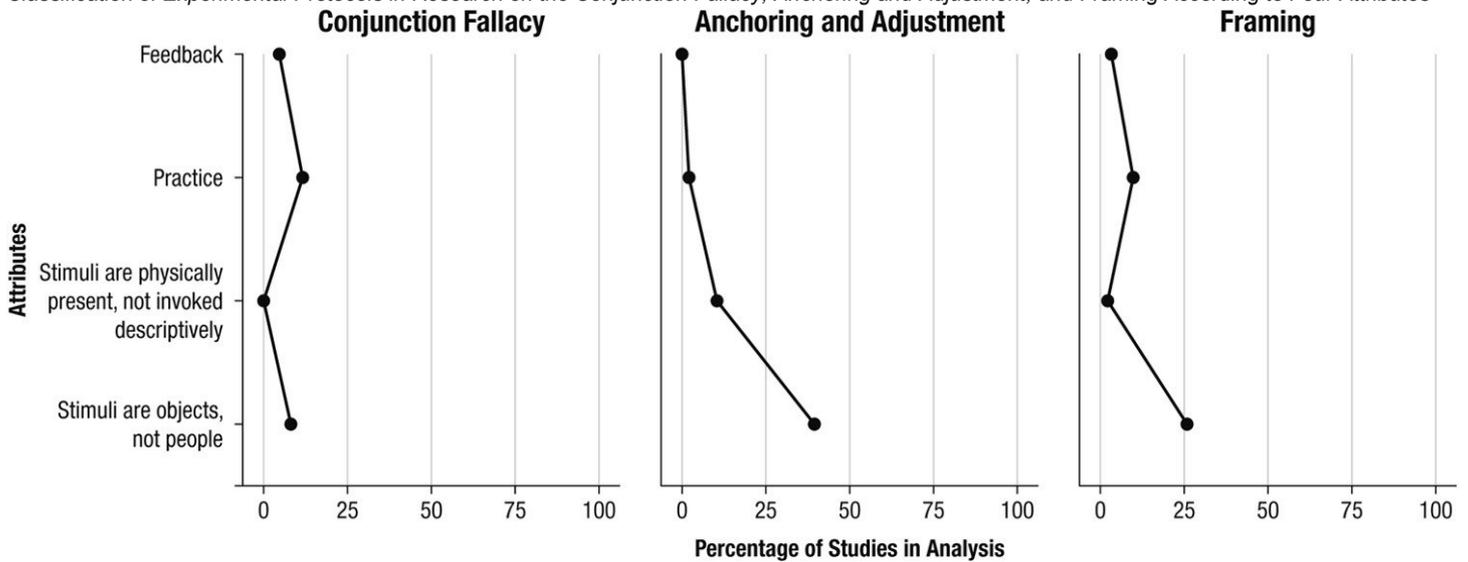
competence or inappropriate application of the normative criteria; rather they result from, for instance, performance errors. Gigerenzer (1996) criticized what he saw as the imposition of unnecessarily narrow norms of good reasoning to diagnose cognitive illusions. He argued that cognitive illusions can be reduced and sometimes even eliminated by presenting the information in text vignettes as natural frequencies rather than in terms of single-event probabilities. Pinker (1997) argued that the idea that humans could have evolved with “no instinct for probability,” making them “blind to chance” (p. 351) is singularly implausible. The debate has often been heated, leading some observers to speak of “rationality wars” (Samuels et al., 2002).

It is beyond the scope of this article to review this debate and the conceptual arguments raised. Instead, we return to Edwards (1983) who, to the best of our knowledge, wrote only once about his views on the heuristics-and-biases program.

I have been in disagreement with this line of research and thought for some time, and now I am ashamed about my own role in starting it off. I remained silent about it because I believed, wrongly, that it was a fad and would die out—though those of you who have followed my work will note that I published not a word about conservatism in probabilistic inference since about 1970. However, I now find that the ideas, without the accompanying caveats, have spread far beyond psychology. [. . .] It is time to call a halt. I have two messages. One is that psychologists have failed to heed the urging of Egon Brunswik (1955b) that generalizations from laboratory tasks should consider the degree to which the task (and the person performing it) resemble or represent the context to which the generalization is made. [. . .] My second message is that, even without tools, experts can in fact do a remarkably good job of assessing and working with probabilities. (pp. 508–511)

In our view, Edwards’s reference to Brunswik and his notion of representative design is important regardless of one’s position in the debate—though it is worth noting that the intuitive-statistician program does not systematically implement representative designs and that the frequent use of artificial chance devices (e.g., urns, poker chips) does not demonstrate much concern for the natural habitat of the decision maker. For Brunswik, representative design

**Figure 11**  
*Classification of Experimental Protocols in Research on the Conjunction Fallacy, Anchoring and Adjustment, and Framing According to Four Attributes*



Note. The attribute “multiple testing” (included in Figures 5 and 6) is reported in detail in the text.

was indispensable for measuring achievement or performance. Brunswik (1943) pointed to a widespread double standard in the way that sampling is practiced in psychological research: Whereas the procedures for sampling participants are typically scrutinized carefully, the empirical observations obtained are generalized to conditions outside those used in the laboratory without the same scrutiny or logical defense (see Dhimi et al., 2004).

What does this mean for statistical intuitions? People cannot help but rely on statistical intuitions when dealing with symbolic and descriptive information (e.g., health statistics, weather forecasts). Statistical intuitions may even become increasingly important in a digital world. Yet people also make judgments, choices, and inferences based solely on past experience and on-the-job learning. In other words, they need to reason statistically in both the symbolically described world and the world affording experiential involvement and learning. Drawing on Brunswik’s (1955b) representative design, it is clear that researchers need to consider which of these worlds they implement in the laboratory. Their choice will dictate which generalizations about people’s competences and rationality they are justified in making from their findings.

Many conclusions about the error-proneness and irrationality of people’s statistical intuitions have been made without taking the experimental protocol and the resulting limits on generalizability into account. We suggest that the debate on the rationality of the human mind must not be divorced from the choice of experimental protocol. Researchers need to study the full range of situations and circumstances under which people deal with uncertain worlds, considering dimensions such as repetition, feedback, learning opportunities, experiential engagement, and transparency of symbolic description. Different research questions will require different experimental protocols, and findings will need to be understood and interpreted contingent on the environmental circumstances implemented in the laboratory. A pluralistic approach that combines descriptive and experiential protocols is much more likely to yield findings that represent the full range of human competences. It will also allow researchers to more closely integrate research on human statistical intuitions with highly related experimental results from research on primates and babies (Schulze & Hertwig, 2021a), from experimental economics (e.g., the rich empirical and theoretical results on one-shot vs. repeated games; see Binmore, 1994), and from cognitive science’s research on Bayesian models of cognition.

### Reconciling Bayesian Models of Cognition and Poor Numerical Cognition

One might expect research on behavioral decision making and cognitive science to be highly interconnected due to their shared interest in human cognition. Yet this is not the case. Whereas the heuristics-and-biases view of human cognition—including the dictum that people are not Bayesian at all—has made strong inroads into many neighboring fields, including public policy making, it is Bayesian models of cognition that dominate in cognitive science. This development can be seen as a reversal of the heuristics-and-biases program’s view on the essence of human rationality; Bayesian models of cognition explain human learning and inductive reasoning in terms of Bayesian inference and Bayesian rationality. How can these two influential but opposing views be squared? Griffiths et al. (2010) argued that: Using probabilistic models to provide a computational-level explanation does not require that hypothesis spaces or probability distributions be explicitly represented by the underlying psychological or neural processes, or that people learn and reason by explicitly using Bayes’ rule (p. 362).

In other words, their account of cognition focuses on the computational and not on the algorithmic level. This distinction goes back to Marr (1982), who defined the computational level as referring to how a problem can be solved in functional terms. The algorithmic level, in contrast, refers to the processes that the mind implements to produce this solution. According to this rationale, on the algorithmic level, people may well be as flawed as the heuristics-and-biases program suggests. Discerning between the computational and the algorithmic level rendered it possible for Oaksford and Chater (2009) to argue that “people are typically poor at numerical reasoning about probability” (p. 69) and yet still aim to “reevaluate forty years of empirical research in the psychology of human reasoning, and cast human rationality in a new and more positive light” (p. 69). Setting aside situations in which people need to operate with symbolic descriptions of uncertainty, Tenenbaum et al. (2011) wrote:

The claim that human minds learn and reason according to Bayesian principles is not a claim that the mind can implement any Bayesian inference. Only those inductive computations that the mind is designed to perform well, where biology has had time and cause to engineer effective and efficient mechanisms, are likely to be understood in Bayesian terms. [ . . . ] In contrast, in tasks that require explicit conscious manipulations of probabilities as numerical quantities—a recent cultural invention that few people become fluent with, and only then after sophisticated training—judgments can be notoriously biased away from Bayesian norms. (p. 1280)

Cognitive scientists advancing Bayesian models of cognition evidently struggle with the finding that people are “not Bayesian at all” (Kahneman & Tversky, 1972, p. 450). One solution is to limit the Bayesian claim to cognitive tasks that do not require conscious effort and attention, and for which evolution has equipped the mind with the right software (as suggested by Tenenbaum et al., 2011). Part of the solution, however, could also be in the intuitivestatistician program. Remember what Edwards (1968) concluded about people’s statistical reasoning. He suggested that their opinion change is orderly, and usually proportional to numbers calculated from Bayes’s theorem—but that it is insufficient in amount. In other words, people update probability in a way that qualitatively, but not strictly quantitatively, obeys Bayes’s rule. Compare this conclusion with Oaksford and Chater’s (2009): “human thought is sensitive to subtle patterns of qualitative Bayesian, probabilistic reasoning” (p. 69). Despite this parallelism, Oaksford and Chater’s (2009) précis of their book “Bayesian rationality” in *Behavioral and Brain Sciences* does not draw on either Edwards’s (1968) or Peterson and Beach’s (1967) work—nor was this work featured in the widely cited Griffiths et al. (2010) article on probabilistic models of cognition or in Tenenbaum et al.’s (2011) article. We believe that those concerned with Bayesian models of cognition have something to gain from rediscovering the intuitivestatistician program and, more precisely, studies with a focus on experience and learning.

This is not meant as a naïve call to return to the intuitive-statistician approach. This research was left behind not only due to the attractions of the heuristics-and-biases program. Researchers were also dissatisfied with key properties of the intuitive-statistician research program, including its highly artificial experimental stimuli (the bookbag-and-poker-chip tradition), strong and untested assumptions, and general lack of a process theory (see Rapoport & Wallsten, 1972; Wallsten, 1972, 1976).

Furthermore, in some subsequent but little-cited research on probabilistic reasoning (still in the bookbag-and-poker-chip tradition), experience received much more explicit attention than it did before 1967. For instance, Wallsten (1976) found “large and systematic differences” (p. 171) in some parameter values of models of probabilistic inference as a function of experience. More generally, he observed that people reduced their processing requirements as they gained experience, learning to focus on the more pertinent information and to ignore redundant information. He thus foreshadowed research on ecologically rational heuristics (Gigerenzer et al., 2011), which suggests that the use of heuristics necessitates some experience, namely, proficiency in discerning the value of probabilistic information. We argue that research on Bayesian models of cognition can link past, present, and future research examining experience and learning in probabilistic reasoning. The finding that people are not Bayesian at all is only part of the story.

Let us conclude by returning briefly to Peterson and Beach (1967) and their use of the metaphor of the intuitive statistician. Much like the work of Tenenbaum et al. (2011), the intuitive-statistician program did not mean to imply that people actually implement sophisticated reasoning processes akin to those of statisticians. Instead, for researchers in this tradition, probability theory and statistics offered a ready-made class of models that should and could be modified in light of observed behaviors to be enlisted as predictive models. This approach was not unlike that of Friedman (1953) in economic theorizing; his dictum was that complete psychological realism was unattainable but that the benchmark of a model was its predictive power. Similarly, the researchers in the intuitive-statistician program focused on prediction at the expense of psychologically insightful process models—perhaps not surprisingly given that many of the researchers in Edwards’s laboratory at the University of Michigan had human engineering backgrounds and the group published regularly in human engineering journals (e.g., Edwards, 1962; Phillips et al., 1966). This engineering focus also foreshadowed the focus of Edwards and colleagues on person-machine systems and on “decision technology—the rules and tools that help us to make wiser decisions” (Edwards & Fasolo, 2001, p. 581). In hindsight, the program’s abstinence from process models, relative to the heuristics-and-biases program’s offer of psychologically intuitive heuristics, may also have crucially contributed to why “this line of research was quietly abandoned” (Fischhoff & Beyth-Marom, 1983, p. 248).

### Experience Is Not Devoid of Biases

Early researchers certainly did not glorify the intuitive statistical mind. Take, for instance, the comments by Edwards et al. (1965) on their findings of conservatism in probability updating:

Whatever the merits or demerits of a built-in tendency to conservatism in information processing in daily life, such a tendency is clearly a hindrance to human effectiveness in information-processing systems. Such systems have no need for built-in, conservative, information-processing biases; they can provide much less automatic, much more rational biases in rather different ways. (p. 310)

Although experience appears to entail properties that can improve statistical inference, it is no panacea. Some cognitive and memory biases appear to be impervious to experience. One striking example is hindsight bias, described by Tolstoy (1869/1982) in *War and Peace* as the “law of retrospectiveness, which makes all the past appear a preparation for events that occur subsequently” (p. 843). Many experimental studies (e.g., Fischhoff, 1975) have documented hindsight biases (for a recent meta-analysis, see Guilbault et al., 2004). Although one can debate the extent to which hindsight bias impedes learning or is a byproduct of learning (e.g., Hoffrage et al., 2000), it is without doubt a robust memory bias. Evidence also suggests that people can be unwilling or unprepared to go beyond the data at hand—in a kind of metacognitive myopia, they fail to take the history and validity of the sample and the sampling processes into account (Fiedler, 2000, 2012). This has led some researchers to argue that humans are

“naïve intuitive statisticians” (Fiedler & Juslin, 2006; but see Le Mens & Denrell, 2011, for a different view) in their experience-based intuitions. Research on decisions from experience has also produced ample evidence that people’s risky choices (with and without feedback) often reveal a pattern that Hertwig and Erev (2009) summarized as follows: “in decisions from experience, people behave as if the rare events have less impact than they deserve according to their objective probabilities, whereas in decisions from description people behave as if the rare events have more impact than they deserve” (p. 518; see also Erev et al., 2017; Regenwetter & Robinson, 2017). This pattern can also be interpreted as reflecting a kind of miscalibration to base rates (especially to those of rare events) that takes opposite directions in description versus experience.

Another line of research investigates when and why a person learns better by observing repeated decisions of another person, rather than experiencing the outcomes themselves (Merlo & Schotter, 2003). This may happen when the immediate experience focuses the person on the local stimulus–response unit rather than permitting them to learn about the general statistical structure of the environment. As these examples illustrate, experience is not impervious to biases—and yet, some biases in experience may not be biases at all. Let us consider the hot-hand fallacy.

### **Beyond Naïveté in Normative Benchmarks**

Many basketball fans, coaches, and players believe in the existence of a “hot hand,” where a player’s ability to make shots is noticeably better than usual. The player is thought to be on a streak, even though the data show that their “streak” is essentially random: Gilovich et al. (1985) debunked the belief that a player’s chances of hitting a shot are higher following a hit than following a miss. The hot-hand fallacy, a “massive and widespread cognitive illusion” (Kahneman, 2011, p. 117), is thought to underlie a variety of anomalies in behavior (e.g., in financial markets; see Miller & Sanjurjo, 2014) and appears to be impervious to learning.

In recent years, however, the human mind has been acquitted from this particular charge of irrationality. According to Miller and Sanjurjo (2014, 2018) the “hot hand” is not a myth after all. Let us assume that basketball players have a 50% probability of hitting a shot. Miller and Sanjurjo (2014) demonstrated that, in a finite number of attempts, the probability of a hit is always 50%, but the conditional probability of a hit immediately following a hit is not. In a finite number of attempts, sequences with consecutive identical outcomes can only be arranged in so many ways. And it turns out that a hit is more likely to be followed by a miss than by another hit. For instance, in a sequence of three shot attempts, there are eight possible sequences of hits (H) and misses (M; e.g., MMM, MMH, . . . , HHH). Among these sequences, the expected proportion of hits on shots following a hit is not .50 but .42 (5/12; see Table 1 in Miller & Sanjurjo, 2018). So, players who have a .5 success rate after a hit are scoring above expectation and indeed have a hot hand. Miller and Sanjurjo reanalyzed the data from Gilovich et al. (1985) and estimated a hot-hand effect of around 12%. In other words, the difference in probability of making a shot following three hits in a row, relative to three misses in a row, was estimated to be 12% higher than would be expected for a player with a constant success rate of .50. Therefore, what has been cast as a “general misconception of chance” (Gilovich et al., 1985, p. 295) actually reflects an accurate perception. Ironically, the bias resides in researchers, who lack the experience of basketball players and who compute expected probabilities from description.

This reevaluation of the hot-hand fallacy raises a broader issue. Experience is inevitably limited, but these limits are not necessarily reflected in descriptions, which can have far-reaching normative implications. As Hahn (2014) pointed out, the “statistics of small samples are often quite different from those of large samples, and this needs to be taken into account in assessing the rationality of human behavior” (p. 229). In the descriptions used in the lab, researchers often invoke benchmarks of statistical reasoning that hold in large samples or at the limit. Yet these principles do not always hold in small or finite samples—the samples of human experience. Consequently, “evaluations of rationality based on long-run considerations or limit properties may lead to a distorted picture” (Hahn, 2014, p. 239) of human rationality. In other words, informed decisions on whether to use descriptive or experiential protocols will likely also advance the debate over which standards are suitable for judging the appropriateness of statistical intuition.

### **Which Properties of Experience May Foster Statistical Intuitions?**

We already highlighted that experience is neither a panacea nor devoid of biases. Yet there are indications that it results, by and large, in statistical intuitions that are not as replete with biases as has been reported in the context of descriptive experimental protocols. Which qualities of experience make it different from description and thus potentially foster statistical intuitions? Hertwig et al. (2018; see also Schulze & Hertwig, 2021a) suggested that interaction with the world affords myriad concurrent dimensions of information that symbolic description lacks or can convey only in condensed form (see Hertwig et al., 2018). A learner experiencing a sequence of events may, for instance, simultaneously receive sensory and motor feedback (potentially triggering affective or motivational processes); obtain temporal, structural, and sample size information; and gain firsthand insights into conditions for statistical inferences (e.g., randomness or independence). These properties must be explicitly stipulated by the author of the description or assumed by the reader. But what are the specific mechanisms by which experience facilitates statistical inference? And under which cognitive and environmental conditions can properties of experience improve or impair accurate statistical intuitions? Several mechanisms are conceivable and different factors may determine the influence of experience on statistical intuitions, depending on the task at hand. Here we briefly examine two factors (also discussed in Schulze & Hertwig, 2021a): computational ease and incremental learning.

#### **Computational Ease**

Experiencing events sequentially can ease computational demands, reducing difficulty and fostering good judgments. Consider the simple method of computing a gamble’s expected value. In experience-based problems, such as  $n$ -armed bandits (Sutton & Barto, 1998), the computation involves adding up all the experienced rewards and dividing the sum by the number of

experiences. In described gambles, by contrast, the computation of expected values involves multiplying each possible consequence by its corresponding probability, and then summing all probability-weighted consequences. Thus, computing the natural mean of a series of experiences involves a much simpler calculus than does description-based expected value theory (Hertwig & Pleskac, 2010). This simpler calculus may be one explanation for Wulff et al.'s (2018) finding—from their metaanalysis of thousands of description- and (sampling-based) experience-based choices between risky lottery options—that far fewer choices were consistent with a maximization process in decisions from description than in decisions from experience. Finally, computational ease arises from simple updating processes such as those captured by reinforcement learning. They can alleviate the burden of storing all the experienced outcomes, as only one value (the current one) is assumed to be held in memory (Frey et al., 2015). Therefore, learning from experience may become more critical as decision problems become more complex (Hogarth & Soyer, 2015; Lejarraga, 2010).

### **Incremental Learning**

Experience also allows people to learn and decide in a step-by-step manner, thus adapting to the demands of the environment. Researchers in the heuristics-and-biases program, which has focused on one-shot situations, therefore exclude “one of the most important determinants of the behavior they purport to explain” (Hogarth, 1981, p. 213). Perhaps Binmore (1994) put this criticism most bluntly:

But how much attention should we pay to experiments that tell us how inexperienced people behave when placed in situations with which they are unfamiliar, and in which the incentives for thinking things through carefully are negligible or absent altogether? [ . . . ] Does it [the participant's behavior] survive after the subjects have had a long time to familiarize themselves with all the wrinkles of the unusual situation in which the experimenter has placed them? (pp. 184–85)

One need not agree with Binmore's (1994) criticism of snapshot studies to appreciate that these studies leave little room for people to learn—to observe, correct, and adjust, to craft their responses progressively as experience accumulates. To use the metaphor employed by Connolly (1988), decision makers learning from sequential experience can clip away at a hedge incrementally, whereas decision makers confronted with one-shot descriptive situations are expected to fell a tree in a single pass (see Connolly, 1988). Judgment errors in experience can potentially be remedied; in description, they cannot.

Let us end this review of properties of experience that foster statistical competence with an important clarification. Throughout this article, we have used the shortcut of referring to protocols as descriptive or experiential. This dichotomy is an oversimplification. There are many variants of both experience and description, and experimental protocols can implement more or fewer aspects of experience. Description and experience can be seen as end points of a continuum (see also Camilleri & Newell, 2013; Hertwig, Pleskac, et al., 2019; Rakow & Newell, 2010; Wulff et al., 2018). Variations in an experimental protocol that “move” it up or down the description–experience continuum can be expected to lead, all other things being equal, to systematic changes in performance. Take natural frequencies. Symbolic descriptions that reflect more of the underlying natural experience (e.g., the sample size)—as described natural frequency formats do—should lead to better statistical inferences. Such findings have been reported, for instance, in research on Bayesian reasoning (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; see McDowell & Jacobs, 2017, for a meta-analysis and a discussion of the underlying mechanisms).

### **The Policy Implications of This Debate**

Investigating the nature of people's statistical intuitions is important. These intuitions are a building block for judgments and decisions, and they come into play whenever people think about the future, reckon with uncertainty, or make potentially highly consequential decisions about their health, wealth, and well-being. The way researchers assess the intuitive statistical mind is directly relevant to the question of how much help people need to make good decisions—and what kind of help is likely to be most efficient. In recent years, the most influential approach to helping people make better decisions has been *nudging* (Thaler & Sunstein, 2008). This approach is deeply rooted in the findings, metaphors, and theoretical commitments of the heuristics-and-biases program (see Grüne-Yanoff & Hertwig, 2016). Thaler and Sunstein (2008) held that research into human behavior raises “serious questions about the rationality of many judgments and decisions that people make” (p. 7) and that “hundreds of studies confirm that human forecasts are flawed and biased. Human decision making is not so great either” (p. 7).

Taking this error-proneness as the starting point, the nudging approach does not aim to remove people's biases, let alone to build lasting cognitive or motivational competences. Nudges (at least of the noneducative kind; Sunstein, 2016) target the chooser's external choice architecture and vary features that people typically claim not to care about (e.g., position in a list, default, framing; see also Rebonato, 2012). By harnessing these external features, the choice architect steers the chooser away from the behavior implied by the cognitive or motivational shortcoming and toward the chooser's ultimate goal or preference (e.g., healthier food choices). In other words, the nudging approach embraces Thaler's (1991) view that mental illusions are people's standard mode of operation, and that their error-proneness is so incorrigible that it is more efficient to work with these illusions than to work to overcome them.

Yet as our analysis demonstrates, this dire portrayal of human decision-making competences is not the only legitimate conception. As Hertwig and Grüne-Yanoff (2017) noted, several other frameworks exist, including ecological rationality (Gigerenzer et al., 1999; Hertwig, Pleskac, et al., 2019), the probabilistic mind (Griffiths et al., 2010), and natural decision making (Klein, 1999). These frameworks' conclusions about human decision-making competences are generally less disquieting in terms of the error-proneness of people's statistical intuitions and decision-making competences. They also justify more optimism about the success and efficiency of interventions designed to foster people's competences. Based on these frameworks, Hertwig and Grüne-Yanoff (2017) proposed a different class of behavioral science interventions: *boosts*. In contrast to nudges, the objective of boosts is to foster people's

competence to make their own choices—that is, to exercise their own agency.

The debate about people's statistical intuitions is a debate among academics about the adaptivity and rationality of human cognition, but more is at stake. The debate has immediate consequences for how people can be supported to make better decisions (see also the notion of simulated experience; Hogarth & Soyer, 2011; Kaufmann et al., 2013). It affects the extent to which policymakers should invest in building people's decision-making competences while preserving their autonomy, or in implementing paternalistic interventions (albeit libertarian in nature; Thaler & Sunstein, 2008). Relatedly, there is the rarely addressed issue (for an exception, see Erev & Roth, 2014) of whether mechanism design<sup>28</sup> should rely primarily on decisions from experience, decisions from description, or, depending on the domain, both. The debate between the intuitive-statistician program and the heuristics-and-biases program is as necessary, timely, and relevant today as ever, with important methodological, theoretical, and policymaking implications.

## Conclusions

In the late 1960s and early 1970s, research into people's intuitive statistical reasoning underwent a dramatic conceptual change. Peterson and Beach (1967), drawing on an extensive review of more than 160 experiments, concluded that the mind is an intuitive statistician—one whose processes can be modeled using probability theory and statistics. Tversky and Kahneman (1974), drawing on a set of 30 experiments they had conducted between 1971 and 1973, unequivocally rejected this conclusion. Their arguments prevailed. Moreover, their research program became the most successful research program in psychology in the second half of the 20th century. One important key to its signature finding—the extensive catalogue of systematic deviations between people's reasoning and assumed norms of rationality—stems from a largely overlooked but profound change in experimental practice: from a largely experiential protocol in which learning was often required to a largely descriptive protocol offering few, if any, opportunities for learning.

The experimental methods scientists choose are not neutral tools. As Hacking (1983) emphasized, experiments, more than theories, convince people that scientific entities are real. Perhaps it is the very simplicity of the experiments in the heuristics-and-biases tradition that explains their persuasive power. But this experimental approach exacts a price. For instance, descriptive protocols make it difficult to establish crucial structural components for statistical reasoning (e.g., random sampling), largely remove learning from the equation, and typically use people rather than objects as experimental stimuli, thus potentially introducing sources of ambiguity that cannot be resolved through trial and error.

The diverging conclusions about the mind drawn by proponents of the two research programs highlight the risk of relying exclusively on one class of experimental methods. We advocate for a pluralist methodological approach encompassing both descriptive and experiential protocols; we further propose that the empirical findings of the intuitive-statistician program be integrated into the collective canon of knowledge about people's statistical cognition. Doing so would not take away from the many insights of the heuristics-and-biases paradigm. Instead, it would offer a more complete understanding of humans' bounded rationality by including one of the great human strengths: the ability to learn.

## References

- Albright, L., & Malloy, T. E. (2000). Experimental validity: Brunswik, Campbell, Cronbach, and enduring issues. *Review of General Psychology*, 4(4), 337–353. <https://doi.org/10.1037/1089-2680.4.4.337>
- Armstrong, B., & Spaniol, J. (2017). Experienced probabilities increase understanding of diagnostic test results in younger and older adults. *Medical Decision Making*, 37(6), 670–679. <https://doi.org/10.1177/0272989X17691954>
- Bar-Hillel, M. (1977). *The base-rate fallacy in probability judgments* (Technical report PTR.1042-77.4/1). Hebrew University.
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233. [https://doi.org/10.1016/0001-6918\(80\)90046-3](https://doi.org/10.1016/0001-6918(80)90046-3)
- Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3), 215–233. <https://doi.org/10.1002/bdm.443>
- Bazerman, M. H., & Moore, D. (2013). *Judgment in managerial decision making* (8th ed.). Wiley.
- Bazerman, M. H., & Neale, M. A. (1992). *Negotiating rationally*. Free Press.
- Beach, L. R., & Peterson, C. R. (1966). Subjective probabilities for unions of events. *Psychonomic Science*, 5(8), 307–308. <https://doi.org/10.3758/BF03328412>
- Beach, L. R., & Scopp, T. S. (1967). Intuitive statistical inferences about variances. *Organizational Behavior and Human Performance*, 3(2), 109–123.
- Binmore, K. (1994). *Game theory and the social contract: Vol. I. Playing fair*. MIT Press.
- Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *The American Journal of Psychology*, 96(1), 85–94. <https://doi.org/10.2307/1422211>
- Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, 50(3), 255–272. <https://doi.org/10.1037/h0060889>
- Brunswik, E. (1952). The conceptual framework of psychology. *Psychological Bulletin*, 49(6), 654–656. <https://doi.org/10.1037/h0049873>
- Brunswik, E. (1955a). In defense of probabilistic functionalism: A reply. *Psychological Review*, 62(3), 236–242. <https://doi.org/10.1037/h0040198>
- Brunswik, E. (1955b). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193–217. <https://doi.org/10.1037/h0047470>
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.
- 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). Harvard University Press.
- Butt, J. (1988). Frequency judgments in an auditing-related task. *Journal of Accounting Research*, 26(2), 315–330. <https://doi.org/10.2307/2491106>
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework.

<sup>28</sup> In economic policy making, mechanism design theory examines when market or market-based institutions yield desirable outcomes and when other institutions are likely more efficient in yielding the desirable outcomes.

- Journal of Risk and Uncertainty*, 19(1/3), 7–42. <https://doi.org/10.1023/A:1007850605129>
- Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 5(4), 325–370. <https://doi.org/10.1007/BF00122575>
- Camilleri, A. R., & Newell, B. R. (2013). Mind the gap? Description, experience, and the continuum of uncertainty in risky choice. In V. S. C. Pammi & N. Srinivasan (Eds.), *Progress in brain research* (pp. 55–71). Elsevier.
- Camilleri, A. R., & Newell, B. R. (2019). Better calibration when predicting from experience (rather than description). *Organizational Behavior and Human Decision Processes*, 150, 62–82. <https://doi.org/10.1016/j.obhdp.2018.10.006>
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. <https://doi.org/10.1037/h0040950>
- Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4), 928–935. <https://doi.org/10.1037/0096-1523.7.4.928>
- Cohen, J., Chesnick, E. I., & Haran, D. (1971). Evaluation of compound probabilities in sequential choice. *Nature*, 232(5310), 414–416. <https://doi.org/10.1038/232414a0>
- Cohen, J., Deamaley, E. J., & Hansel, C. E. M. (1958). Skill and chance: Variation in estimates of skill with an increasing element of chance. *British Journal of Psychology*, 49(4), 319–323. <https://doi.org/10.1111/j.2044-8295.1958.tb00670.x>
- Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, 4(3), 317–331. <https://doi.org/10.1017/S0140525X00009092>
- Connolly, T. (1988). Hedge-clipping, tree-felling, and the management of ambiguity: The need for new images of decision-making. In M. B. McCaskey, L. R. Pondy, & H. Thomas (Eds.), *Managing the challenge of ambiguity and change* (pp. 37–50). Wiley and Sons.
- 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), 1–73. [https://doi.org/10.1016/0010-0277\(95\)00664-8](https://doi.org/10.1016/0010-0277(95)00664-8)
- Dai, J., Pachur, T., Pleskac, T. J., & Hertwig, R. (2019). What the future holds and when: A description–experience gap in intertemporal choice. *Psychological Science*, 30(8), 1218–1233. <https://doi.org/10.1177/0956797619858969>
- Dawson, N. V., & Arkes, H. R. (1987). Systematic errors in medical decision making. *Journal of General Internal Medicine*, 2(3), 183–187. <https://doi.org/10.1007/BF02596149>
- Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, 12(5), 523–538. <https://doi.org/10.1287/orsc.12.5.523.10092>
- Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130(6), 959–988. <https://doi.org/10.1037/0033-2909.130.6.959>
- Dunker, K. (1945). On problem solving. *Psychological Monographs*, 58(5), i–113. <https://doi.org/10.1037/h0093599>
- Dutt, V., Arló-Costa, H., Helzner, J., & Gonzalez, C. (2014). The description–experience gap in risky and ambiguous gambles. *Journal of Behavioral Decision Making*, 27, 316–327. <https://doi.org/10.1002/bdm.1808>
- Edwards, W. (1953). Probability-preferences in gambling. *The American Journal of Psychology*, 66(3), 349–364. <https://doi.org/10.2307/1418231>
- Edwards, W. (1954a). Probability-preferences among bets with differing expected values. *The American Journal of Psychology*, 67(1), 56–67. <https://doi.org/10.2307/1418071>
- Edwards, W. (1954b). The theory of decision making. *Psychological Bulletin*, 51(4), 380–417. <https://doi.org/10.1037/h0053870>
- Edwards, W. (1961a). Behavioral decision theory. *Annual Review of Psychology*, 12, 473–498. <https://doi.org/10.1146/annurev.ps.12.020161.002353>
- Edwards, W. (1961b). Probability learning in 1,000 trials. *Journal of Experimental Psychology*, 62(4), 385–394. <https://doi.org/10.1037/h0041970>
- Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors*, 4(2), 59–74. <https://doi.org/10.1177/001872086200400201>
- Edwards, W. (1966). *Non-conservative probabilistic information processing systems* (ESD-TR-66-404, Report 05893-22-F). University of Michigan, Institute of Science and Technology, Electronic Systems Division.
- Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). Wiley and Sons.
- Edwards, W. (1983). Human cognitive capacities, representativeness, and ground rules for research. In P. Humphreys, O. Svenson, & A. Vari (Eds.), *Analyzing and aiding decision processes* (pp. 507–513). Akademiai Kiado.
- Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual Review of Psychology*, 52(1), 581–606. <https://doi.org/10.1146/annurev.psych.52.1.581>
- Edwards, W., Lindman, H., & Phillips, L. D. (1965). Emerging technologies for making decisions. In F. Barron, W. C. Dement, W. Edwards, H. Lindman, L. D. Phillips, J. Olds, & M. Olds (Eds.), *New directions in psychology II* (pp. 261–325). Holt, Rinehart and Winston.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193–242. <https://doi.org/10.1037/h0044139>
- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4), 643–669. <https://doi.org/10.2307/1884324>
- Elstein, A. S. (1999). Heuristics and biases: Selected errors in clinical reasoning. *Academic Medicine*, 74(7), 791–794. <https://doi.org/10.1097/00001888-199907000-00012>
- Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, 124(4), 369–409. <https://doi.org/10.1037/rev0000062>
- Erev, I., & Roth, A. E. (2014). Maximization, learning, and economic behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 111(Supplement 3), 10818–10825. <https://doi.org/10.1073/pnas.1402846111>
- Erev, I., Shimonovich, D., Schurr, A., & Hertwig, R. (2008). Base rates: How to make the intuitive mind appreciate or neglect them. In H. Plessner, C. Betsch, & T. Betsch (Eds.), *Intuition in judgment and decision making* (pp. 135–148). Erlbaum.
- Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3), 519–527. <https://doi.org/10.1037/0033-295X.101.3.519>
- Erlick, D. E. (1966). Human estimates of statistical relatedness. *Psychonomic Science*, 5(10), 365–366. <https://doi.org/10.3758/BF03328441>
- Ert, E., & Lejarraga, T. (2018). The effect of experience on context-dependent decisions. *Journal of Behavioral Decision Making*, 31(4), 535–546. <https://doi.org/10.1002/bdm.2064>
- Ert, E., & Trautmann, S. T. (2014). Sampling experience reverses preferences for ambiguity. *Journal of Risk and Uncertainty*, 49(1), 31–42. <https://doi.org/10.1007/s11166-014-9197-9>
- Evans, J. S. B. T., & Bradshaw, H. (1986). Estimating sample-size requirements in research design: A study of intuitive statistical judgment. *Current Psychological Research & Reviews*, 5(1), 10–19. <https://doi.org/10.1007/BF02686592>

- Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, 107(4), 659–676. <https://doi.org/10.1037/0033-295X.107.4.659>
- Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 57, pp. 1–55). Elsevier Academic Press.
- Fiedler, K., & Juslin, P. (2006). *Information sampling and adaptive cognition*. Cambridge University Press.
- Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288–299. <https://doi.org/10.1037/0096-1523.1.3.288>
- Fisk, J. E. (2004). Conjunction fallacy. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment, and memory* (pp. 23–42). Psychology Press.
- Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). McGraw-Hill.
- Fraenkel, L., Peters, E., Tyra, S., & Oelberg, D. (2016). Shared medical decision making in lung cancer screening: Experienced versus descriptive risk formats. *Medical Decision Making*, 36(4), 518–525. <https://doi.org/10.1177/0272989X15611083>
- Frey, R., Mata, R., & Hertwig, R. (2015). The role of cognitive abilities in decisions from experience: Age differences emerge as a function of choice set size. *Cognition*, 142, 60–80. <https://doi.org/10.1016/j.cognition.2015.05.004>
- Friedman, M. (1953). *Essays in positive economics*. University of Chicago Press.
- Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *Journal of Socio-Economics*, 40(1), 35–42. <https://doi.org/10.1016/j.socec.2010.10.008>
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103(3), 592–596. (1996). <https://doi.org/10.1037/0033-295X.103.3.592>
- Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press.
- Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). A 30% chance of rain tomorrow: How does the public understand probabilistic weather forecasts? *Risk Analysis*, 25(3), 623–629. <https://doi.org/10.1111/j.1539-6924.2005.00608.x>
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704. <https://doi.org/10.1037/0033-295X.102.4.684>
- Gigerenzer, G., & Muir Gray, J. A. (2011). *Better doctors, better patients, better decisions: Envisioning health care 2020*. MIT Press.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge University Press.
- Gigerenzer, G., Todd, P. M., & the ABC Group. (1999). *Simple heuristics that make us smart*. Oxford University Press.
- Gilovich, T., & Griffin, D. (2002). Introduction—Heuristics and biases: Then and now. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1–18). Cambridge University Press.
- Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3), 295–314. [https://doi.org/10.1016/0010-0285\(85\)90010-6](https://doi.org/10.1016/0010-0285(85)90010-6)
- Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, 52(3), 464–473. <https://doi.org/10.1037/0022-3514.52.3.464>
- Golan, H., & Ert, E. (2015). Pricing decisions from experience: The roles of information-acquisition and response modes. *Cognition*, 136, 9–13. <https://doi.org/10.1016/j.cognition.2014.11.008>
- Gonzalez, C., & Mehlohorn, K. (2016). Framing from experience: Cognitive processes and predictions of risky choice. *Cognitive Science*, 40(5), 1163–1191. <https://doi.org/10.1111/cogs.12268>
- Gould, S. J. (1992). *Bully for Brontosaurus: Further reflections in natural history*. Penguin Books.
- Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review*, 69(4), 623–638.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357–364. <https://doi.org/10.1016/j.tics.2010.05.004>
- Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, 26(1–2), 149–183. <https://doi.org/10.1007/s11023-015-9367-9>
- Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology*, 26(2–3), 103–117. <https://doi.org/10.1080/01973533.2004.9646399>
- Güney, S., & Newell, B. R. (2015). Overcoming ambiguity aversion through experience. *Journal of Behavioral Decision Making*, 28(2), 188–199. <https://doi.org/10.1002/bdm.1840>
- Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.
- Hacking, I. (1984). Experimentation and scientific realism. In J. Leplin (Ed.), *Scientific realism* (pp. 154–172). University of California Press.
- Hahn, U. (2014). Experiential limitation in judgment and decision. *Topics in Cognitive Science*, 6(2), 229–244. <https://doi.org/10.1111/tops.12083>
- Hamlin, R. P. (2017). The gaze heuristic: Biography of an adaptively rational decision process. *Topics in Cognitive Science*, 9(2), 264–288. <https://doi.org/10.1111/tops.12253>
- Hammond, K. R. (1996). Upon reflection. *Thinking & Reasoning*, 2(2–3), 239–248. <https://doi.org/10.1080/135467896394537>
- Hertwig, R. (2015). Decisions from experience. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 240–267). Wiley-Blackwell.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8), 534–539. <https://doi.org/10.1111/2Fi.0956-7976.2004.00715.x>
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12), 517–523. <https://doi.org/10.1016/j.tics.2009.09.004>
- Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12(6), 973–986. <https://doi.org/10.1177/1745691617702496>
- Hertwig, R., Hogarth, R. M., & Lejarraga, T. (2018). Experience and description: Exploring two paths to knowledge. *Current Directions in Psychological Science*, 27(2), 123–128. <https://doi.org/10.1177/0963721417740645>
- Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, 24(3), 383–403. <https://doi.org/10.1017/S0140525X01004149>
- Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115(2), 225–237. <https://doi.org/10.1016/j.cognition.2009.12.009>
- Hertwig, R., Pleskac, T. J., Pachur, T., & the Center for Adaptive Rationality. (2019). *Taming uncertainty*. MIT Press.
- Hertwig, R., Wulff, D. U., & Mata, R. (2019). Three gaps and what they may mean for risk preference. *Philosophical Transactions of the Royal Society of London: B, Biological Sciences*, 374(1766), 20180140. <https://doi.org/10.1098/rstb.2018.0140>

- Heuvelom, F. (2012). Three explanations for the Kahneman–Tversky Programme of the 1970s. *European Journal of the History of Economic Thought*, 19(5), 797–828. <https://doi.org/10.1080/09672567.2010.540350>
- Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A byproduct of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(3), 566–581. <https://doi.org/10.1037/0278-7393.26.3.566>
- Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, 90(2), 197–217. <https://doi.org/10.1037/0033-2909.90.2.197>
- Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus nontransparent description. *Journal of Experimental Psychology: General*, 140(3), 434–463. <https://doi.org/10.1037/a0023265>
- Hogarth, R. M., & Soyer, E. (2015). Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory & Cognition*, 4(3), 221–228. <https://doi.org/10.1016/j.jarmac.2014.01.005>
- Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *The Journal of Consumer Research*, 9(1), 90–98. <https://doi.org/10.1086/208899>
- Isler, O., Kopsacheilis, O., & van Dolder, D. (2020). *The description–experience gap in cooperation*. Manuscript in preparation.
- Kahneman, D. (2003). *Daniel Kahneman: Biographical. The Nobel Prize*. <https://www.nobelprize.org/prizes/economic-sciences/2002/kahneman/biographical/>
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454. [https://doi.org/10.1016/0010-0285\(72\)90016-3](https://doi.org/10.1016/0010-0285(72)90016-3)
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. <https://doi.org/10.1037/h0034747>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–391. <https://doi.org/10.2307/1914185>
- Kaufmann, C., Weber, M., & Haisley, E. (2013). The role of experience sampling and graphical displays on one's investment risk appetite. *Management Science*, 59(2), 323–340. <https://doi.org/10.1287/mnsc.1120.1607>
- Klein, G. A. (1999). *Sources of power: How people make decisions*. MIT Press.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(1), 1–53. <https://doi.org/10.1017/S0140525X00041157>
- Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27(3), 313–327. <https://doi.org/10.1017/S0140525X04000081>
- Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14(7), 293–300. <https://doi.org/10.1016/j.tics.2010.05.001>
- Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 75(1), 23–55. <https://doi.org/10.1006/obhd.1998.2781>
- Laplace, P. S. (1951). A philosophical essay on probabilities (F. W. Truscott & F. L. Emory, Trans.). Dover. (Original work published 1814)
- Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, 118(2), 379–392. <https://doi.org/10.1037/a0023010>
- Lejarraga, T. (2010). When experience is better than description: Time delays and complexity. *Journal of Behavioral Decision Making*, 23(1), 100–116. <https://doi.org/10.1002/bdm.666>
- Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes*, 116(2), 286–295. <https://doi.org/10.1016/j.obhdp.2011.05.001>
- Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2016). Decisions from experience: From monetary to medical gambles. *Journal of Behavioral Decision Making*, 29(1), 67–77. <https://doi.org/10.1002/bdm.1877>
- Lejarraga, T., Woike, J. K., & Hertwig, R. (2016). Description and experience: How experimental investors learn about booms and busts affects their financial risk taking. *Cognition*, 157, 365–383. <https://doi.org/10.1016/j.cognition.2016.10.001>
- Lewis, M. (2017). *The undoing project: A friendship that changed our minds*. WW Norton & Company.
- Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making*, 27(2), 146–156. <https://doi.org/10.1002/bdm.1792>
- Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, 126(1), 373–416. <https://doi.org/10.1093/qje/qjq004>
- Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology*, 38(2), 231–248. <https://doi.org/10.1037/0022-3514.38.2.231>
- March, J. G. (2010). *The ambiguities of experience*. Cornell University Press.
- Marr, D. (1982). *Vision*. Freeman.
- Martin, J. M., Gonzalez, C., Juvina, I., & Lebiere, C. (2014). A description–experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making*, 27, 349–362. <https://doi.org/10.1002/bdm.1810>
- McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143(12), 1273–1312. <https://doi.org/10.1037/bul0000126>
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52(3), 194–216. <https://doi.org/10.1037/h0048070>
- Merlo, A., & Schotter, A. (2003). Learning by not doing: An experimental investigation of observational learning. *Games and Economic Behavior*, 42(1), 116–136. [https://doi.org/10.1016/S0899-8256\(02\)00537-7](https://doi.org/10.1016/S0899-8256(02)00537-7)
- Miller, J. B., & Gelman, A. (2018). *Laplace's theories of cognitive illusions, heuristics, and biases*. <https://ssrn.com/abstract=3149224>
- Miller, J. B., & Gelman, A. (2020). Laplace's theories of cognitive illusions, heuristics and biases. *Statistical Science*, 35(2), 159–170. <https://doi.org/10.1214/19-STS696>
- Miller, J. B., & Sanjurjo, A. (2014). *A cold shower for the hot hand fallacy: Robust evidence that belief in the hot hand is justified* (IGIER Working Paper No. 518). <https://ssrn.com/abstract=2450479>
- Miller, J. B., & Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, 86(6), 2019–2047. <https://doi.org/10.3982/ECTA14943>
- Mischel, W., & Ebbesen, E. B. (1970). Attention in delay of gratification. *Journal of Personality and Social Psychology*, 16(2), 329–337. <https://doi.org/10.1037/h0029815>
- Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502–517. <https://doi.org/10.1037/0033-295X.115.2.502>
- Moussaïd, M. (2019). Crowds on the move. In R. Hertwig, T. J. Pleskac, T. Pachur, & the Center for Adaptive Rationality (Eds.), *Taming uncertainty* (pp. 263–282). MIT Press.
- Moussaïd, M., Helbing, D., & Theraulaz, G. (2011). How simple rules determine pedestrian behavior and crowd disasters. *Proceedings of the*

- National Academy of Sciences of the United States of America*, 108(17), 6884–6888. <https://doi.org/10.1073/pnas.1016507108>
- Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, 21(7), 960–969. <https://doi.org/10.1177/0956797610372637>
- Nisbett, R. E., & Borgida, K. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, 32(5), 932–943. <https://doi.org/10.1037/0022-3514.32.5.932>
- Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90(4), 339–363. <https://doi.org/10.1037/0033-295X.90.4.339>
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Prentice Hall.
- Oaksford, M., & Chater, N. (2009). Précis of “Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, 32(1), 69–84. <https://doi.org/10.1017/S0140525X09000284>
- Peterson, C. R., & Beach, L. R. (1967). *Man as an intuitive statistician*. *Psychological Bulletin*, 68(1), 29–46. <https://doi.org/10.1037/h0024722>
- Peterson, C. R., Hammond, K. R., & Summers, D. A. (1965). Optimal responding in multiple-cue probability learning. *Journal of Experimental Psychology*, 70, 270–276. <https://doi.org/10.1037/h0022261>
- Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, 70(1), 117–121. <https://doi.org/10.1037/h0022023>
- Peterson, C. R., Schneider, R. J., & Miller, A. J. (1965). Sample size in the revision of subjective probabilities. *Journal of Experimental Psychology*, 69, 522–527. <https://doi.org/10.1037/h0021720>
- Peterson, C. R., & Ulehla, Z. J. (1964). Uncertainty, inference difficulty, and probability learning. *Journal of Experimental Psychology*, 67, 523–530. <https://doi.org/10.1037/h0046858>
- Peterson, C. R., Ulehla, Z. J., Miller, A. J., Bourne, L. E., Jr., & Stilson, D. W. (1965). Internal consistency of subjective probabilities. *Journal of Experimental Psychology*, 70(5), 526–533. <https://doi.org/10.1037/h0022546>
- Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3), 346–354. <https://doi.org/10.1037/h0023653>
- Phillips, L. D., Hays, W. L., & Edwards, W. (1966). Conservatism in complex probabilistic inference. *IEEE Transactions on Human Factors in Electronics*, HFE-7(1), 7–18. <https://doi.org/10.1109/THFE.1966.231978>
- Phillips, L. D., & von Winterfeldt, D. (2007). Reflections on the contributions of Ward Edwards to decision analysis and behavioral research. In W. Edwards, R. F. Miles, Jr., & D. von Winterfeldt (Eds.), *Advances in decision analysis: From foundations to applications* (pp. 71–80). Cambridge University Press. <https://doi.org/10.1017/CBO9780511611308.006>
- Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. Wiley.
- Pinker, S. (1997). *How the mind works*. Norton.
- Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review*, 122(4), 621–647. <https://doi.org/10.1037/a0039413>
- Pollard, P., & Evans, J. S. B. T. (1983). The role of ‘representativeness’ in statistical inference: A critical appraisal. In J. St. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 117–144). Routledge/KeganPaul.
- Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, 23(1), 1–14. <https://doi.org/10.1002/bdm.681>
- Rapoport, A. (1964). Sequential decision-making in a computer-controlled task. *Journal of Mathematical Psychology*, 1(2), 351–374. [https://doi.org/10.1016/0022-2496\(64\)90008-2](https://doi.org/10.1016/0022-2496(64)90008-2)
- Rapoport, A., & Wallsten, T. S. (1972). Individual decision behavior. *Annual Review of Psychology*, 23(1), 131–176. <https://doi.org/10.1146/annurev.ps.23.020172.001023>
- Rebonato, R. (2012). *Taking liberties: A critical examination of libertarian paternalism*. Palgrave Macmillan.
- Regenwetter, M., & Robinson, M. M. (2017). The construct–behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, 124(5), 533–550. <https://doi.org/10.1037/rev0000067>
- Rehder, B., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition*, 45(2), 245–260. <https://doi.org/10.3758/s13421-016-0662-3>
- Rouanet, H. (1961). Études de décisions expérimentales et calcul de probabilités [Studies of experimental decision making and the probability calculus]. *Colloques internationaux du Centre National de la Recherche Scientifique* (pp. 33–43). Éditions du Centre National de la Recherche Scientifique.
- Samuels, R., Stich, S., & Bishop, M. (2002). Ending the rationality wars: How to make disputes about human rationality disappear. In R. Elio (Ed.), *Common sense, reasoning, and rationality* (pp. 236–268). Oxford University Press.
- Schulze, C., & Hertwig, R. (2021a). A description–experience gap in statistical intuitions: Of smart babies, risk-savvy chimps, intuitive statisticians, and stupid grown-ups. *Cognition*, 210, Article 104580. <https://doi.org/10.1016/j.cognition.2020.104580>
- Schulze, C., & Hertwig, R. (2021b). *Statistical intuitions: Experiencing information improves children’s and adults’ probabilistic inferences* [Manuscript submitted for publication]. Max Planck Institute for Human Development.
- Schum, D. A. (1999). Some reflections on Ward Edwards. In J. Shanteau, B. A. Mellers, & D. A. Schum (Eds.), *Decision science and technology* (pp. 405–420). Springer.
- Schurz, G., & Hertwig, R. (2019). Cognitive success: A consequentialist account of rationality in cognition. *Topics in Cognitive Science*, 11(1), 7–36. <https://doi.org/10.1111/tops.12410>
- Shuford, E. H. (1959). *A comparison of subjective probabilities for elementary and compound events* (Report No. 20). University of North Carolina, Psychometric Laboratory.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2), 129–138. <https://doi.org/10.1037/h0042769>
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1976). Cognitive processes and societal risk taking. In J. S. Carol & J. W. Payne (Eds.), *Cognition and social behavior* (pp. 1654–1684). Erlbaum.
- Spencer, J. (1961). Estimating averages. *Ergonomics*, 4(4), 317–328. <https://doi.org/10.1080/00140136108930533>
- Stich, S. P. (1985). Could man be an irrational animal? *Synthese*, 64(1), 115–135. <https://doi.org/10.1007/BF00485714>
- Sunstein, C. R. (2000). *Behavioral law and economics*. Cambridge University Press.
- Sunstein, C. R. (2016). *The ethics of influence: Government in the age of behavioral science*. Cambridge University Press.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285. <https://doi.org/10.1126/science.1192788>
- Thaler, R. H. (1991). *Quasi rational economics*. Russel Sage Foundation.
- Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *The American Economic Review*, 106(7), 1577–1600. <https://doi.org/10.1257/aer.106.7.1577>
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Tolstoy, L. (1982). *War and peace* (R. Edmonds, Eds.). Greenwich House. (Original work published 1869)

- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110. <https://doi.org/10.1037/h0031322>
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232. [https://doi.org/10.1016/0010-0285\(73\)90033-9](https://doi.org/10.1016/0010-0285(73)90033-9)
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453–458. <https://doi.org/10.1126/science.7455683>
- Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–100). Cambridge University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4), 293–315. <https://doi.org/10.1037/0033-295X.90.4.293>
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323. <https://doi.org/10.1007/BF00122574>
- van den Bos, W., & Hertwig, R. (2017). Adolescents display distinctive tolerance to ambiguity and to uncertainty during risky decision making. *Scientific Reports*, 7, 40962. <https://doi.org/10.1038/srep40962>
- Vedantam, S. (Host). (2018, March 12). Daniel Kahneman on misery, memory, and our understanding of the mind (no. 100) [Audio podcast episode]. *Hidden Brain*. NPR. <https://www.npr.org/2018/03/12/592986190/daniel-kahneman-on-misery-memory-and-our-understanding-of-the-mind>
- Wallsten, T. S. (1972). Conjoint-measurement framework for the study of probabilistic information processing. *Psychological Review*, 79(3), 245–260. <https://doi.org/10.1037/h0032687>
- Wallsten, T. S. (1976). Using conjoint-measurement models to investigate a theory about probabilistic information processing. *Journal of Mathematical Psychology*, 14(2), 144–185. [https://doi.org/10.1016/0022-2496\(76\)90044-4](https://doi.org/10.1016/0022-2496(76)90044-4)
- Weber, E. U., Blais, A., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, 15(4), 263–290. <https://doi.org/10.1002/bdm.414>
- Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, 111(2), 430–445. <https://doi.org/10.1037/0033-295X.111.2.430>
- Wegier, P., & Shaffer, V. A. (2017). Aiding risk information learning through simulated experience (ARISE): Using simulated outcomes to improve understanding of conditional probabilities in prenatal Down syndrome screening. *Patient Education and Counseling*, 100(10), 1882–1889. <https://doi.org/10.1016/j.pec.2017.04.016>
- Wulff, D. U., Hills, T. T., & Hertwig, R. (2015). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making*, 28(3), 214–223. <https://doi.org/10.1002/bdm.1841>
- Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description–experience gap. *Psychological Bulletin*, 144(2), 140–176. <https://doi.org/10.1037/bul0000115>
- Yechiam, E., & Hochman, G. (2013). Losses as modulators of attention: Review and analysis of the unique effects of losses over gains. *Psychological Bulletin*, 139(2), 497–518. <https://doi.org/10.1037/a0029383>
- Zhang, H., & Hout, J. W. (2020). Exaggerated prevalence effect with the explicit prevalence information: The description–experience gap in visual search. *Attention, Perception & Psychophysics*, 82(7), 3340–3356. <https://doi.org/10.3758/s13414-020-02045-8>