

## 2. Representativeness Revisited: Attribute Substitution in Intuitive Judgment

*Daniel Kahneman and Shane Frederick*

The program of research now known as the *heuristics and biases approach* began with a survey of 84 participants at the 1969 meetings of the Mathematical Psychology Society and the American Psychological Association (Tversky & Kahneman, 1971). The respondents, including several authors of statistics texts, were asked realistic questions about the robustness of statistical estimates and the replicability of research results. The article commented tongue-in-cheek on the prevalence of a belief that the law of large numbers applies to small numbers as well: Respondents placed too much confidence in the results of small samples and their statistical judgments showed little sensitivity to sample size.

The mathematical psychologists who participated in the survey not only should have known better – they did know better. Although their intuitive guesses were off the mark, most of them could have computed the correct answers on the back of an envelope. These sophisticated individuals apparently had access to two distinct approaches for answering statistical questions: one that is spontaneous, intuitive, effortless, and fast; and another that is deliberate, rule-governed, effortful, and slow. The persistence of large biases in the guesses of experts raised doubts about the educability of statistical intuitions. Moreover, it was known that the same biases affect choices in the real world, where researchers commonly select sample sizes that are too small to provide a fair test of their hypotheses (Cohen, 1969, 1992). Tversky and Kahneman (1971) therefore concluded that intuitions should be regarded “with proper suspicion” and that researchers should “replace impression formation by computation whenever possible” (p. 31).

To explain the judgments they had observed, Tversky and Kahneman conjectured that observers expect the statistics of a sample to closely resemble (or “represent”) the corresponding population parameters, even when the sample is small. This “representation hypothesis” soon led to the idea of a “representativeness heuristic,” according to which some probability judgments (the likelihood that  $X$  is a  $Y$ ) are mediated by assessments of resemblance (the

We thank Maya Bar-Hillel, Tom Gilovich, Dale Griffin, Ralph Hertwig, Denis Hilton, David Krantz, Barbara Mellers, Ilana Ritov, Norbert Schwarz, and Philip Tetlock for helpful comments.

degree to which  $X$  “looks like” a  $Y$ ). This was the origin of the idea of heuristics in which a difficult question is answered by substituting an answer to an easier one – a theme that we develop further in this chapter.

From its earliest days, the heuristics and biases program was guided by the idea that intuitive judgments occupy a position – perhaps corresponding to evolutionary history – between the automatic parallel operations of perception and the controlled serial operations of reasoning. The boundary between perception and judgment is fuzzy and permeable: the *perception* of a stranger as menacing is inseparable from a *prediction* of future harm. Intuitive thinking extends perception-like processing from current sensations to judgment objects that are not currently present, including mental representations that are evoked by language. However, the representations on which intuitive judgments operate retain some features of percepts: they are concrete and specific, and they carry causal propensities and an affective charge.

A slower and more controlled mode of thinking governs the performance of unfamiliar tasks, the processing of abstract concepts, and the deliberate application of rules. A comprehensive psychology of intuitive judgment cannot ignore such controlled thinking, because intuition can be overridden or corrected by self-critical operations, and because intuitive answers are not always available. But this sensible position seemed irrelevant in the early days of research on judgment heuristics. The authors of the “law of small numbers” saw no need to examine correct statistical reasoning. They believed that including easy questions in the design would insult the participants and bore the readers. More generally, the early studies of heuristics and biases displayed little interest in the conditions under which intuitive reasoning is preempted or overridden – controlled reasoning leading to correct answers was seen as a default case that needed no explaining. A lack of concern for boundary conditions is typical of young research programs, which naturally focus on demonstrating new and unexpected effects, not on making them disappear. However, the topic of boundary conditions must eventually be faced as a program develops. The question of how biases are avoided was first addressed some time ago (Kahneman & Tversky, 1982; Tversky & Kahneman, 1983); we expand on it here.

The first section introduces a distinction between two families of cognitive operations, called *System 1* and *System 2*. The second section presents an attribute-substitution model of heuristic judgment, which elaborates and extends earlier treatments of the topic (Kahneman & Tversky, 1982; Tversky & Kahneman, 1974, 1983). The third section introduces a research design for studying attribute substitution. The fourth section discusses the controversy over the representativeness heuristic. The last section situates representativeness within a broad family of prototype heuristics, in which properties of a prototypical exemplar dominate global judgments concerning an entire set.

**TWO FAMILIES OF COGNITIVE OPERATIONS**

The ancient idea that cognitive processes can be partitioned into two main families – traditionally called *intuition* and *reason* – is now widely embraced under the general label of *dual-process theories* (Chaiken & Trope, 1999; Hammond, 1996; Sloman, 1996, Chapter 22, this volume). Dual-process models come in many flavors, but all distinguish cognitive operations that are quick and associative from others that are slow and rule-governed (Gilbert, 1999). We adopt the generic labels *System 1* and *System 2* from Stanovich and West (Chapter 24). These terms may suggest the image of autonomous homunculi, but such a meaning is not intended. We use *systems* as a label for collections of processes that are distinguished by their speed, controllability, and the contents on which they operate (Table 2.1.)

Although System 1 is more primitive than System 2, it is not necessarily less capable. On the contrary, complex cognitive operations eventually migrate from System 2 to System 1 as proficiency and skill are acquired. A striking demonstration of the intelligence of System 1 is the ability of chess masters to perceive the strength or weakness of chess positions instantly. For those experts, pattern matching has replaced effortful serial processing. Similarly, prolonged cultural exposure eventually produces a facility for social judgments – for example, an ability to recognize quickly that “a man whose dull writing is occasionally enlivened by corny puns” is more similar to a stereotypical computer programmer than to a stereotypical accountant.

In the particular dual-process model we assume, System 1 quickly proposes intuitive answers to judgment problems as they arise, and System 2 monitors the quality of these proposals, which it may endorse, correct, or override. The judgments that are eventually expressed are called *intuitive* if they retain the hypothesized initial proposal without much modification. The roles of the two systems in determining stated judgments depend on features of the task and of the individual, including the time available for deliberation (Finucane et al., 2000), the respondent’s mood (Isen, Nygren, & Ashby, 1988; Bless et al., 1996), intelligence (Stanovich & West, Chapter 24), and exposure to statistical thinking (Nisbett et al., 1983; Agnoli & Krantz, 1989; Agnoli, 1991). We assume that System 1 and System 2 can be active concurrently, that automatic and controlled cognitive operations compete for the control of overt responses, and that deliberate

**Table 2.1. Two Cognitive Systems**

<b>System 1 (Intuitive)</b>	<b>System 2 (Reflective)</b>
<b>Process Characteristics</b>	
Automatic	Controlled
Effortless	Effortful
Associative	Deductive
Rapid, parallel	Slow, serial
Process opaque	Self-aware
Skilled action	Rule application
<b>Content on Which Processes Act</b>	
Affective	Neutral
Causal propensities	Statistics
Concrete, specific	Abstract
Prototypes	Sets

52    Daniel Kahneman and Shane Frederick

judgments are likely to remain anchored on initial impressions. Our views in these regards are similar to the “correction model” proposed by Gilbert and colleagues (1989, 1991) and to other dual-process models (Epstein, 1994; Hammond, 1996; Sloman, 1996).

In the context of a dual-system view, errors of intuitive judgment raise two questions: “What features of System 1 created the error?” and “Why was the error not detected and corrected by System 2?” (cf. Kahneman & Tversky, 1982). The first question is more basic, of course, but the second should not be neglected, as illustrated next.

The notions of heuristic and bias were introduced by Tversky and Kahneman (1974; p. 3 in Kahneman, Slovic and Tversky, 1982) in the following paragraph:

The subjective assessment of probability resembles the subjective assessment of physical quantities such as distance or size. These judgments are all based on data of limited validity, which are processed according to heuristic rules. For example, the apparent distance of an object is determined in part by its clarity. The more sharply the object is seen, the closer it appears to be. This rule has some validity, because in any given scene the more distant objects are seen less sharply than nearer objects. However, the reliance on this rule leads to systematic errors in the estimation of distance. Specifically, distances are often overestimated when visibility is poor because the contours of objects are blurred. On the other hand, distances are often underestimated when visibility is good because the objects are seen sharply. Thus the reliance on clarity as an indication leads to common biases. Such biases are also found in intuitive judgments of probability. (p. xxx)

This statement was intended to extend Brunswik’s (1943) analysis of the perception of distance to the domain of intuitive thinking and to provide a rationale for using biases to diagnose heuristics. However, the analysis of the effect of haze is flawed: It neglects the fact that an observer looking at a distant mountain possesses two relevant cues, not one. The first cue is the blur of the contours of the target mountain, which is positively correlated with its distance, when all else is equal. This cue should be given positive weight in a judgment of distance, and it is. The second relevant cue, which the observer can readily assess by looking around, is the ambient or general haziness. In an optimal regression model for estimating distance, general haziness is a suppressor variable, which must be weighted negatively because it contributes to blur but is uncorrelated with distance. Contrary to the argument made in 1974, using blur as a cue does not inevitably lead to bias in the judgment of distance – the illusion could just as well be described as a *failure to assign adequate negative weight to ambient haze*. The effect of haziness on *impressions* of distance is a failing of System 1; the perceptual system is not designed to correct for this variable. The effect of haziness on *judgments* of distance is a separate failure of System 2. Although people are capable of consciously correcting their impressions of distance for the effects of ambient haze, they commonly fail to do so. A similar analysis applies to some of the judgmental biases we discuss later, in which errors and biases only occur when both systems fail.

### ATTRIBUTE SUBSTITUTION

Early research on the representativeness and availability heuristics was guided by a simple and general hypothesis: when confronted with a difficult question people often answer an easier one instead, usually without being aware of the substitution. A person who is asked “What proportion of long-distance relationships break up within a year?” may answer as if she had been asked “Do instances of swift breakups of long-distance relationships come readily to mind?” This would be an application of the availability heuristic. A professor who has heard a candidate’s job talk and now considers the question “How likely is it that this candidate could be tenured in our department?” may answer the much easier question: “How impressive was the talk?”. This would be an example of the representativeness heuristic.

The heuristics and biases research program has focused primarily on representativeness and availability – two versatile attributes that are automatically computed and can serve as candidate answers to many different questions. It has also focused on thinking under uncertainty. However, the restriction to particular heuristics and to a specific context is largely arbitrary. We will say that judgment is mediated by a heuristic when an individual assesses a specified *target attribute* of a judgment object by substituting another property of that object – the *heuristic attribute* – which comes more readily to mind. Many judgments are made by this process of *attribute substitution*. For an example, consider the well-known study by Strack, Martin, & Schwarz (1988), in which college students answered a survey that included these two questions: “How happy are you with your life in general?” and “How many dates did you have last month?”. The correlation between the two questions was negligible when they occurred in the order shown, but it rose to 0.66 when the dating question was asked first. We suggest that thinking about the dating question automatically evokes an affectively charged evaluation of one’s satisfaction in that domain of life, which lingers to become the heuristic attribute when the happiness question is subsequently encountered. The observed value of 0.66 certainly underestimates the true correlation between the target and heuristic attributes, because dating frequency is not a perfect proxy for romantic satisfaction and because of measurement error in all variables. The results suggest that respondents had little besides love on their mind when they evaluated their overall well-being.

### Biases

Because the target attribute and the heuristic attribute are different, the substitution of one for the other inevitably introduces systematic biases. In this chapter we are mostly concerned with *weighting biases*, which arise when cues available to the judge are given either too much or too little weight. Criteria for determining optimal weights can be drawn from several sources. In the classic lens model, the optimal weights associated with different cues are the

regression weights that optimize the prediction of an external criterion, such as physical distance or the GPA that a college applicant will attain (Brunswick, 1943; Hammond, 1955). Our analysis of weighting biases applies to such cases, but it also extends to attributes for which no objective criterion is available, such as an individual's overall happiness or the probability that a particular patient will survive surgery. Normative standards for these attributes must be drawn from the constraints of ordinary language, and they are often imprecise. For example, the conventional meaning of *overall happiness* does not specify how much weight ought to be given to various life domains. However, it certainly does require that substantial weight be given to every important domain of life, and that no weight at all be given to the current weather, or to the recent consumption of a cookie. Similar rules of common sense apply to judgments of probability. For example, the statement "John is more likely to survive a week than a month" is a true statement in ordinary usage, which implies a rule that people would wish their judgments to follow. Accordingly, neglect of duration in assessments of survival probabilities would be properly described as a weighting bias, even if there is no way to establish a normative probability for individual cases (Kahneman & Tversky, 1996).

In some judgmental tasks, information that could serve to supplement or correct the heuristic is not neglected or underweighted, but simply lacking. If asked to judge the relative frequency of words beginning with K or R (Tversky and Kahneman, 1973) or to compare the population of a familiar foreign city with one that is unfamiliar (Gigerenzer and Goldstein, 1996), respondents have little recourse but to base such judgments on ease of retrieval or recognition. The necessary reliance on these heuristic attributes renders such judgements susceptible to biasing factors (e.g., the amount of media coverage). However, unlike weighting biases, such biases of insufficient information cannot be described as errors of judgment, because there is no way to avoid them.

### **Accessibility and Substitution**

The intent to judge a target attribute initiates a search for a reasonable value. Sometimes this search terminates almost immediately because the required value can be read from a stored memory (e.g., the question, "How tall are you?") or current experience ("How much do you like this cake?"). For other judgments, however, the target attribute does not come to mind immediately, but the search for it evokes activates the value of other attributes that are conceptually and associatively related (e.g., a question about overall happiness may retrieve the answer to a related question about satisfaction with a particular domain of life). Attribute substitution occurs when the target attribute is assessed by mapping the value of some another attribute on the target scale. This process will control judgment when three conditions are satisfied: (1) the target attribute is relatively inaccessible; (2) a semantically and associatively related is highly accessible; and (3) the substitution of the heuristic attribute in the judgment is not rejected by the critical operations of System 2.

Some attributes are permanent candidates for the heuristic role because they are routinely evaluated as part of perception and comprehension, and therefore always accessible (Tversky and Kahneman, 1983). These natural assessments include physical properties such as size and distance, and more abstract properties such as similarity (e.g., Tversky & Kahneman, 1983), cognitive fluency in perception and memory (e.g., Jacoby and Dallas, 1991; Schwarz & Vaughn, Chapter 5, this volume; Tversky & Kahneman, 1973), casual propensity (Kahneman & Varey, 1990; Heider, 1944; Michotle, 1963), surprisingness (Kahneman & Miller, 1986), affective valence (e.g., Bargh, 1997; Cacioppo, Priester, & Berntson, 1993; Kahneman, Ritov, & Schkade, 1999; Slovic, Finucane, Peters, & MacGregor, Chapter 23, this volume; Zajonc, 1980), and mood (Schwarz & Clore, 1983). Other attributes are accessible only if they have been recently evoked or primed (see, e.g., Bargh et al., 1986; Higgins & Brendl, 1995). The 'romantic satisfaction heuristic' for judging happiness illustrates the effect of temporary accessibility. The same mechanism of attribute substitution is involved, whether the heuristic attribute is accessible chronically or only temporarily.

There is sometimes more than one candidate for the role of heuristic attribute. For an example borrowed from Anderson (1991), consider the question, "Are more deaths caused by rattlesnakes or bees?" A respondent who read recently about someone who died from a snakebite or bee sting may use the relative availability of instances of the two categories as a heuristic. If no instances come to mind, the respondent might consult impressions of the "dangerousness" of the typical snake or bee, an application of representativeness. Indeed, it is quite possible that the question initiates both a search for instances and an assessment of dangerousness, and that a contest of accessibility determines the role of the two heuristics in the final response. As Anderson observed, it is not always possible to determine *a priori* which heuristic governs the response to a particular problem.

### **Cross-Dimensional Mapping**

The process of attribute substitution involves the mapping of the heuristic attribute of the judgment object onto the scale of the target attribute. Our notion of cross-dimensional mapping extends Stevens' (1975) concept of cross-modality matching. Stevens postulated that intensive attributes (e.g., brightness, loudness, the severity of crimes) can be mapped onto a common scale of sensory strength, allowing direct matching of intensity across modalities. Indeed, observers find it quite possible to match the loudness of sounds to the severity of crimes. Our conception allows other ways of comparing values across dimensions, such as matching relative positions (e.g., percentiles) in the frequency distributions or ranges of different attributes (Parducci, 1965). An impression of a student's position in the distribution of aptitude may be mapped directly onto a corresponding position in the distribution of academic achievement and then translated into a letter grade. Note that cross-dimensional matching is inherently nonregressive: A judgment or prediction is just as extreme as the

impression mapped onto it. Ganzach and Krantz (1990) applied the term *univariate matching* to a closely related notion.

Cross-dimensional mapping presents special problems when the scale of the target attribute has no upper bound. Kahneman, Ritov, and Schkade (1999) discussed two situations in which an attitude (or affective valuation) are mapped onto an unbounded scale of dollars: respondents in surveys may be required to indicate how much money they would contribute money for a cause, and jurors are sometimes required to specify an amount of punitive damages against a negligent firm. The mapping of attitudes onto dollars is a variant of direct scaling in psychophysics, where respondents assign numbers to indicate the intensity of sensations (Stevens, 1975). The normal practice of direct scaling is for the experimenter to provide a *modulus* – a specified number that is to be associated to a standard stimulus. For example, respondents may be asked to assign the number 10 to the loudness of a standard sound and judge the loudness of other sounds relative to that standard. Stevens (1975) observed that when the experimenter fails to provide a modulus, respondents spontaneously adopt one. The judgements of each respondent are therefore internally coherent but the overall level of these judgments reflects the individual's modulus. Because different respondents may pick moduli that differ greatly (sometimes varying by a factor of 100 or more), the variability in judgments of particular stimuli is dominated by arbitrary individual differences in the size of the modulus. A similar analysis applies to situations in which respondents are required to use the dollar scale to express affection for a species or outrage toward a defendant. Just as Stevens' observers had no principled way to assign a number to a moderately loud sound, survey participants and jurors have no principled way to scale affection or outrage onto dollars. The analogy of scaling without a modulus has been used to explain the notorious variability of dollar responses in surveys of willingness to pay and in jury awards (Kahneman, Ritov, & Schkade, 1999; Kahneman, Schkade, & Sunstein, 1998).

### The Affect Heuristics

The article that defined the heuristics and biases approach (Tversky and Kahneman, 1974) included anchoring and adjustment as a heuristic, along with representativeness and availability. However *Anchoring* does not fit the definition of judgment heuristic we have adopted here because it does not work through the substitution of one attribute for another, but by increasing the plausibility of a particular value of the target attribute (Chapman & Johnson, Chapter 6, this volume).

It has become evident that an *affect heuristic* (Slovic et al., Chapter 23, this volume) should replace anchoring in the list of major general-purpose heuristics. In hindsight, the failure to identify this heuristic earlier reflects the narrowly cognitive focus that characterized psychology for some decades. There is now compelling evidence for the proposition that every stimulus evokes an affective evaluation, and that this evaluation can occur outside of awareness (see reviews

by Zajonc, 1980, 1997; Bargh, 1997). *Affective valence* is a natural assessment, and therefore a candidate for substitution in the numerous situations in which an affectively loaded response is required. The affect heuristic fits the model of attribute substitution. Slovic and colleagues (Chapter 23, this volume) discuss how a basic affective reaction can be used as the heuristic attribute for a wide variety of more complex evaluations, such as the costs and benefit ratio of various technologies, the safe level of chemicals, or even the predicted economic performance of various industries. In the same vein, Kahneman and Ritov (1994) and Kahneman, Ritov, and Schkade (1999) proposed that an automatic affective valuation is the principal determinant of willingness to pay for public goods, and Kahneman, Schkade, and Sunstein (1998) interpreted jurors' assessments of punitive awards as a mapping of outrage onto a dollar scale of punishments.

The idea of a single affect heuristic should be treated as a useful oversimplification because good and bad come in many distinctive flavors. The semantic differential task illustrates both the basic unity and the diversity of valuation. Participants in this task rate objects and concepts on bipolar scales defined by pairs of adjectives, such as GOOD–BAD, KIND–CRUEL, LARGE–SMALL, STRONG–WEAK, WARM–COLD, and others. The main finding of this research is that adjectives such as KIND, PLEASANT, BEAUTIFUL, CUDDLY and SAFE are all highly correlated measures of a single evaluation factor, which has its highest loading on GOOD. However, the adjectives also retain their distinctive meanings: “Justice,” for example, is GOOD, but not especially KIND. Thus, “goodness,” “kindness,” “ugliness,” and “outrageousness” are best viewed as closely related but distinguishable evaluative attributes that can give rise to closely related but distinguishable heuristics.

### **System 2: The Supervision of Intuitive Judgments**

Our model assumes that an intuitive judgment is expressed overtly only if it is endorsed by System 2. The Stroop task illustrates this two-system structure. Observers who are instructed to report the color in which words are printed tend to stumble when the word is the name of another color (e.g., the word *BLUE* printed in green). The difficulty arises because the word is automatically read, and activates a response (“blue” in this case) that competes with the required response. Errors are rare in the Stroop test, indicating generally successful monitoring and control of the overt response, but the conflict produces delays and hesitations. The successful suppression of erroneous responses is effortful, and its efficacy is reduced by stress and distraction.

Gilbert (1989) described a correction model in which initial impulses are often wrong and normally overridden. He argues that people initially believe whatever they are told (e.g., “Whitefish love grapes”) and that it takes some time and mental effort to “unbelieve” such dubious statements. Here again, cognitive load disrupts the controlling operations of System 2, increasing the rate of errors and revealing aspects of intuitive thinking that are normally suppressed. In an ingenious extension of this approach, Bodenhausen (1990) exploited natural

temporal variability in alertness. He found that “morning people” were substantially more susceptible to a judgment bias (the conjunction fallacy) in the evening and that “evening people” were more likely to commit the fallacy in the morning.

Because System 2 is relatively slow, its operations can be disrupted by time pressure. Finucane et al. (2000) reported a study in which respondents judged the risks and benefits of various products and technologies (e.g., nuclear power, chemical plants, cellular phones). When participants were forced to respond within 5 seconds, the correlations between their judgments of risks and their judgments of benefits were strongly negative.

The negative correlations were much weaker (although still pronounced) when respondents were given more time to ponder a response. When time is short, the same affective evaluation apparently serves as a heuristic attribute for assessments of both benefits and risks. Respondents can move beyond this simple strategy, but they need more than 5 seconds to do so. As this example illustrates, judgment by heuristic often yields simplistic assessments, which System 2 sometimes corrects by bringing additional considerations to bear.

Schwarz and his colleagues have shown that attribute substitution can be prevented by alerting respondents to the possibility that their judgment could be contaminated by an irrelevant variable (Schwarz & Clore, 1983; Schwarz, 1996). For example, sunny or rainy weather typically affect reports of well-being, but Schwarz and Clore (1983) found that merely asking respondents about the weather just before the well-being question eliminates the effect – apparently by reminding respondents that their current mood (a candidate heuristic attribute) is influenced by a factor (current weather) that is obviously irrelevant to the requested target attribute (overall well-being). Schwarz (1996) also found that the weight of any aspect of life on judgments of happiness is actually reduced by asking people to describe their satisfaction with that particular aspect of life just before the global question. As these examples illustrate, the effects of a variable on judgment are normally increased by priming (a System 1 effect), but can be reduced by an explicit reminder that brings the self-critical operations of System 2 into play.

We suspect that System 2 endorsements of intuitive judgments are granted quite casually under normal circumstances. Consider the puzzle: “A bat and a ball cost \$1.10 in total. The bat costs \$1 more than the ball. How much does the ball cost?” Almost everyone we ask reports an initial tendency to answer “10 cents” because the sum \$1.10 separates naturally into \$1 and 10 cents, and 10 cents is about the right magnitude. Many people yield to this immediate impulse. The surprisingly high rate of errors in this easy problem illustrates how lightly System 2 monitors the output of System 1: people are not accustomed to thinking hard, and are often content to trust a plausible judgment that quickly comes to mind.

The ball and bat problem elicits many errors, although it is not really difficult and certainly not ambiguous. A moral of this example is that people

often make quick intuitive judgments to which they are not deeply committed. A related moral is that we should be suspicious of analyses that explain apparent errors by attributing to respondents a bizarre interpretation of the question. Consider someone who answers a question about happiness by reporting her satisfaction with her romantic life. The respondent is surely not committed to the absurdly narrow interpretation of *happiness* that her response seemingly implies. More likely, at the time of answering she thinks that she *is* reporting happiness: a judgment comes quickly to mind and is not obviously mistaken; end of story. Similarly, we propose that respondents who judge probability by representativeness do not seriously believe that the questions, "How likely is X to be a Y?" and "How much does X resemble the stereotype of Y?" are synonymous. People who make a casual intuitive judgment normally know little about how their judgment came about, and know even less about its logical entailments. Attempts to reconstruct the meaning of intuitive judgments by interviewing respondents (see e.g., Hertwig and Gigerenzer, 1999) are therefore unlikely to succeed because such probes require better introspective access and more coherent beliefs than people normally muster.

#### **Heuristics: Deliberate or Automatic?**

So far, we have described judgment by heuristic as an intuitive and unintentional process of attribute substitution, which we attribute to System 1. However, attribute substitution can also be a deliberate System 2 strategy, as when a voter decides to evaluate candidates solely by their stance on a particular issue. In other cases, a heuristic is both initiated spontaneously by System 1 and adopted deliberately adopted by System 2. The *recognition heuristic* proposed by Gigerenzer and his colleagues appears to fall in that class.

Experiments described by Gigerenzer and Goldstein (1996; see also Gigerenzer et al., 1999) show that respondents rely on feelings of familiarity and unfamiliarity to compare uncertain quantities, such as the relative size of two cities. For example, 78% of a sample of German students recognized San Diego as an American city, but only 4% recognized San Antonio, and every student who recognized San Diego but not San Antonio concluded (correctly) that San Diego is larger. Though far from perfect (the correlation between actual population and recognition was only 0.60 in that experiment), the recognition heuristic is surely a reasonable strategy for that task. Indeed, when students were given pairs of the 22 most populous cities in the United States or Germany, Americans slightly outperformed Germans when comparing the size of German cities, and Germans did slightly better than Americans when judging American cities (Gigerenzer & Goldstein, 1996).

Gigerenzer and his colleagues have described the recognition heuristic as a deliberate strategy, which in our terms is an operation of System 2. This description seems highly plausible. In addition, however, we have proposed that familiarity is an attribute that System 1 evaluates routinely, regardless of the current judgment goal. On this view, the recognition heuristic has an automatic

component, which could be studied by varying tasks and by measuring reaction times. Imagine a reaction-time study in which respondents on each trial see a question such as “Which city name is printed in larger font?” or “Which city name contains more vowels?,” immediately followed by a pair of cities that differ in familiarity. Research on conceptual Stroop effects (e.g., Keysar, 1989) suggests that the more familiar city name will be the favored answer to any question that is associatively related to prominence, size or quantity. On this hypothesis, errors will be systematic, and response times will be faster for compatible than for incompatible responses. An even more radical possibility, arising from the work of Gilbert (1989), Begg (see, e.g., Begg & Armour, 1991; Begg, Anas, & Farinacci, 1992), and Mandler (see, e.g., Mandler, Hamson, & Dorfman, 1990) is that there will be a bias favoring the familiar item as an answer to *any* question – perhaps even “Which city is *smaller*?” or “Which city has *fewer* dentists?” If either of these hypotheses is correct, the recognition heuristic belongs to the family of heuristics that we consider here. Like many other members of that family, the recognition heuristic for judging city sizes (i) draws on a “natural assessment” of recognition or familiarity, (ii) may be endorsed as a deliberate strategy, (iii) makes people look smart under some conditions, and (iv) will produce systematic errors and biases, because impressions of familiarity and recognition are systematically correlated with factors other than city size, such as number of mentions in the media.

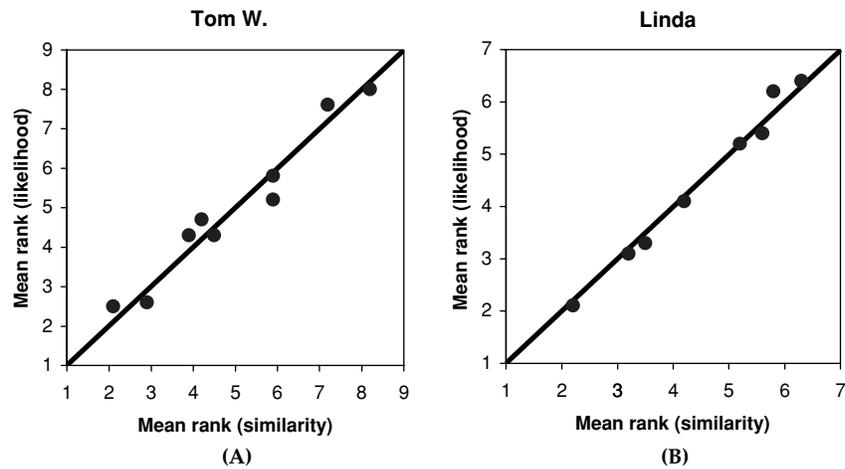
## IDENTIFYING A HEURISTIC

Hypotheses about judgment heuristics have most often been studied by examining weighting biases and deviations from normative rules. However, the hypothesis that one attribute is substituted for another in a judgment task – for example, representativeness for probability – can also be tested more directly. In the *heuristic elicitation design*, one group of respondents provides judgments of a target attribute for a set of objects and another group evaluates the hypothesized heuristic attribute for the same objects. The substitution hypothesis implies that the judgments of the two groups, when expressed in comparable units (e.g., percentiles), will be identical. This section examines several applications of heuristic elicitation.

### Eliciting Representativeness

Figure 2.1 displays the results of two experiments in which a measure of representativeness was elicited. These results were published long ago, but we repeat them here because they still provide the most direct evidence for both attribute substitution and the representativeness heuristic. For a more recent application of a similar design, see Bar-Hillel and Neter (1993, Chapter 3, this volume).

The object of judgment in the study from which Figure 2.1A is drawn (Kahneman & Tversky, 1973; p. 49 in Kahneman, Slovic and Tversky, 1982) was



**Figure 2.1.** (A) Plot of average ranks for nine outcomes for Tom W., ranked by probability and similarity to stereotypes of graduate students in various fields (from Kahneman & Tversky, 1973) (B) Plot of average ranks for eight outcomes for Linda, ranked by probability and representativeness (from Tversky & Kahneman, 1982, p. 94).

the following description of a fictitious graduate student, which was shown along with a list of nine fields of graduate specialization:

Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense. (p. 49)

Participants in a representativeness group ranked the nine fields of specialization by the degree to which Tom W. “resembles a typical graduate student.” Participants in the probability group ranked the nine fields according to the likelihood of Tom W.’s specializing in each. Figure 2.1A plots the mean judgments of the two groups. The correlation between representativeness and probability is nearly perfect (.97). No stronger support for attribute-substitution could be imagined. The interpretation of the relationship between the attributes rests on two assumptions, both of which seem plausible: that representativeness is more accessible than probability, and that there is no third attribute that could explain both judgments.

The Tom W. study was also intended to examine the effect of the base rates of outcomes on categorical prediction. For that purpose, respondents in a third group estimated the proportion of graduate students enrolled in each of the nine fields. By design, some outcomes were defined quite broadly, whereas others were defined more narrowly. As intended, estimates of base-rates

62    Daniel Kahneman and Shane Frederick

varied markedly across fields, ranging from 3% for Library Science to 20% for Humanities and Education. Also by design, the description of Tom W. included characteristics (e.g., introversion) that were intended to make him fit the stereotypes of the smaller fields (library science, computer science) better than the larger fields (humanities and social sciences). As intended, the correlation between the average judgments of representativeness and of base rates was strongly negative ( $-.65$ ).

The logic of probabilistic prediction in this task suggests that the ranking of outcomes by their probabilities should be intermediate between their rankings by representativeness and by base rate frequencies. Indeed, if the personality description is taken to be a poor source of information, probability judgments should stay quite close to the base-rates. The description of Tom W. was designed to allow considerable scope for judgments of probability to diverge from judgments of representativeness, as this logic requires. Figure 2.1 shows no such divergence. Thus, the results of the Tom W. study simultaneously demonstrate the substitution of representativeness for probability and the neglect of known (but not explicitly mentioned) base rates.

Figure 2.1B is drawn from an early study of the Linda problem, the best known and most controversial example in the representativeness literature (Tversky & Kahneman, 1982, p. 92), in which a woman named Linda was described as follows:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations.

As in the Tom W. study, separate groups of respondents were asked to rank a set of eight outcomes by representativeness and by probability. The results are shown in Fig. 2.1B. Again, the correlation between these rankings was almost perfect (.99).

Six of the eight outcomes that subjects were asked to rank were fillers (e.g., elementary school teacher, psychiatric social worker). The two critical outcomes were No. 6 (bank teller) and the so-called conjunction item No. 8 (bank teller and active in the feminist movement). Most subjects ranked the conjunction higher than its constituent, both in representativeness (85%) and probability (89%). The observed ranking of the two items is quite reasonable for judgments of similarity, but not for probability: Linda may resemble a feminist bank teller more than she resembles a bank teller, but she cannot be more likely to be a feminist bank teller than to be a bank teller. In this problem, reliance on representativeness yields probability judgments that violate a basic logical rule. As in the Tom W. study, the results make two points: they support the hypothesis of attribute substitution and also illustrate a predictable judgment error.

The entries plotted in Fig. 2.1 are averages of multiple judgments and the correlations are computed over a set of judgment objects. It should be noted that correlations between averages are generally much higher than corresponding

correlations within the data of individual respondents (Nickerson, 1995). Indeed, group results may even be unrepresentative, if they are dominated by a few individuals who produce more variance than others and have an atypical pattern of responses. Fortunately, this particular hypothesis is not applicable to the experiments of Fig. 2.1, in which all responses were ranks.

### Exploring the Outrage Heuristic

The results of Fig. 2.1 could be scored as a perfect hit for the hypothesis of attribute substitution in general and for the representativeness heuristic in particular. Next, we describe another study in the same design that yielded an instructive near-miss. One hypothesis of that study (Kahneman, Schkade, & Sunstein, 1998) couched in the language of the present treatment was that the setting of punitive damages in civil cases is mediated by an outrage heuristic. The heuristic elicitation procedure was used to test that hypothesis.

Participants drawn from a jury roll in Texas were shown vignettes of legal cases in which a plaintiff had suffered a personal injury while using a product. Respondents were told that the plaintiff had already been awarded compensatory damages, and that their next task as mock jurors was to determine whether punitive damages were also appropriate, and if so in what amount.

The study involved 28 variants of 10 basic scenarios. One of these scenarios concerned a child who had been burned when his pajamas caught fire as he was playing with matches. The pajamas were made of fabric that was not adequately fire-resistant, and the defendant firm had been aware of the problem. Each participant rated one version of each of the 10 scenarios. Two variables were manipulated experimentally as follows: For 4 of the 10 scenarios, 2 versions were constructed that differed in the severity of harm. In the high-harm version of the pajamas case, for example, the child was "severely burned over a significant portion of his body and required several weeks in the hospital and months of physical therapy." In the low-harm version, "the child's hands and arms were badly burned, and required professional medical treatment for several weeks." In addition, each of the 14 resulting vignettes was presented in two versions: one in which the defendant firm was large (annual profits in the range of \$100–200 million), and one in which it was of medium size (\$10–20 million). Each individual read vignettes involving firms of both sizes.

Respondents in a dollar punishment group were asked to indicate whether punitive damages were appropriate, and if so in what amount (the target attribute in this study). Respondents in the outrage group rated the outrageousness of the defendant's behavior (the hypothesized heuristic attribute). The mean outrageousness ratings and the median dollar awards were computed for each vignette. For the purpose of the present analysis, we also obtained (from 16 Princeton students) mean ratings of the severity of the harm suffered in each of the 14 vignettes. Lawsuits were not mentioned in these descriptions of harm.

Because the "pain" that a given financial penalty inflicts on a firm obviously varies with its annual profits, the relation between outrage and dollar awards

64 Daniel Kahneman and Shane Frederick

was evaluated separately for large and small firms. Because dollar responses are known to be a very noisy measure, there was reason to expect that the fit of dollar awards to rated outrageousness would not be as impressive as the relationship between probability ranks and similarity ranks in Figs. 2.1A, B. Even with this allowance for noisy data, the correlations between the median punitive damages in dollars and the means of outrageousness ratings were disappointingly low: .77 for large firms and .87 for small firms. The main reason for the low correlations was an unanticipated discrepancy between the two measures in their sensitivity to the harm suffered by the plaintiff. Harm had no effect on outrageousness ratings, but did strongly affect dollar awards. When we reconsidered these data for this chapter, we concluded that the instructions to “judge the outrageousness of the defendant’s *actions*” (italics ours) may have led many respondents to discount the harm resulting from these actions (Kahneman, Schkade, & Sunstein, 1998, offered a slightly different view). To test this interpretation, we defined a new variable for each case: the product of the average ratings of outrageousness and of harm. The data shown in Fig. 2.2 plot dollar punishments against this new variable. The correlations are still not as high as in Fig. 2.1, but they are now respectable: .90 for large firms and .94 for medium-sized firms.

We do not intend to suggest that respondents separately assessed outrageousness and harm and then computed their product; rather, we propose that the feeling of outrage is ordinarily sensitive to both the recklessness of the culprit and the suffering of the victim, but that the instructions to judge the outrageousness of the defendant’s *actions* encouraged respondents to report something other than their immediate emotional response. In our view, the judgment of outrageousness is psychologically more complex than the emotion

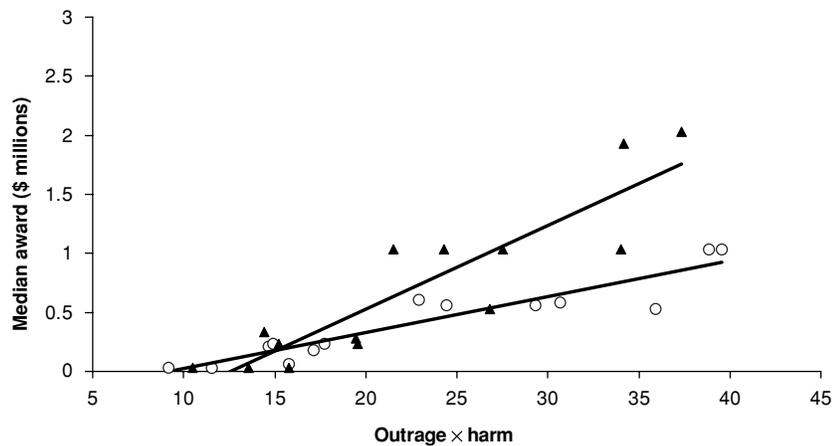


Figure 2.2. Median punitive awards (in dollars) for 14 cases, plotted against the product of average ratings of outrageousness and of severity of harm for each case, for large firms (filled triangles) and for medium-size firms (circles).

of outrage. This interpretation is *post hoc*, but testable in at least two ways: (1) The heuristic elicitation procedure could be repeated, with instructions that simply require a report of the respondent's anger or indignation rather than an evaluation of the defendant's behavior, or (2) the original heuristic elicitation procedure could be replicated under time pressure or cognitive load – manipulations that would interfere with discounting and thereby make outrageousness sensitive to harm.

Even when the heuristic elicitation procedure does demonstrate perfect correspondence (as it did for representativeness), the naming of a heuristic remains a judgment call. The analogy of naming factors in factor analysis is instructive: in both cases the appropriate label is rarely unique, although competing candidates are bound to have much in common. The *outrage heuristic* could just as well have been named the *indignation heuristic*, or perhaps just the *anger heuristic*. As is also the case in factor analysis, the label that is chosen for a heuristic may take on a life of its own in subsequent theorizing. We speculate that the course of research on heuristics could have been somewhat different if the *representativeness heuristic* had been named more simply the *similarity heuristic*.

### THE REPRESENTATIVENESS CONTROVERSY

The experiments summarized in Fig. 2.1 provided direct evidence for the representativeness heuristic and two concomitant biases: neglect of base-rates and conjunction errors. In the terminology introduced by Tversky and Kahneman (1983), the design of these experiments was “subtle”: adequate information was available for participants to avoid the error, but no effort was made to call their attention to that information. For example, participants in the Tom W. experiment had general knowledge of the relative base-rates of the various fields of specialization, but these base-rates were not explicitly mentioned in the problem. Similarly, both critical items in the Linda experiment were included in the list of outcomes, but they were separated by a filler so that respondents would not feel compelled to compare them. In the anthropomorphic language used here, System 2 was given a chance to correct the judgment, but was not prompted to do so.

In view of the confusing controversy that followed, it is perhaps unfortunate that the articles documenting base-rate neglect and conjunction errors did not stop with subtle tests. Each article also contained an experimental flourish – a demonstration in which the error occurred in spite of a manipulation that called participants' attention to the critical variable. The engineer–lawyer problem (Kahneman & Tversky, 1973) included special instructions to ensure that respondents would notice the base-rates of the outcomes. The brief personality descriptions shown to respondents were reported to have been drawn from a set containing descriptions of 30 lawyers and 70 engineers (or vice versa), and respondents were asked, “What is the probability that this description belongs to one of the 30 lawyers in the sample of 100?” To the authors' surprise, base

66 Daniel Kahneman and Shane Frederick

rates were largely neglected in the responses, despite their salience in the instructions. Similarly, the authors were later shocked to discover that more than 80% of undergraduates committed a conjunction error even when asked point blank whether Linda was more likely to be “a bank teller” or “a bank teller who is active in the feminist movement” (Tversky & Kahneman, 1983). The novelty of these additional direct or “transparent” tests was the finding that respondents continued to show the biases associated with representativeness even in the presence of strong cues pointing to the normative response. The errors that people make in transparent judgment problems are analogous to observers’ failure to allow for ambient haze in estimating distances: a correct response is within reach, but not chosen, and the failure involves an unexpected weakness of the corrective operations of System 2.

Discussions of the heuristics and biases approach have focused almost exclusively on the direct conjunction fallacy and on the engineer–lawyer problems. These are also the only studies that have been extensively replicated with varying parameters. The amount of critical attention is remarkable, because the studies were not, in fact, essential to the authors’ central claim. In the terms of the present treatment, that claim was that intuitive prediction is an operation of System 1, which is susceptible both to base-rate neglect and conjunction fallacies. There was no intent to deny the possibility of System 2 interventions that would modify or override intuitive predictions. Thus, the articles in which these studies appeared would have been substantially the same, although far less provocative, if respondents had overcome base-rate neglect and conjunction errors in transparent tests.

To appreciate why the strong forms of base-rate neglect and of the conjunction fallacy sparked so much controversy, it is useful to distinguish two conceptions of human rationality (Kahneman, 2000b). *Coherence rationality* is the strict conception, which requires the agent’s entire system of beliefs and preferences to be internally consistent, and immune to effects of framing and context. For example, an individual’s probability  $P$  (“Linda is a bank teller”) should be the sum of the probabilities  $P$  (“Linda is a bank teller and is a feminist”), and  $P$  (“Linda is a bank teller and not a feminist”). A subtle test of coherence rationality could be conducted by asking individuals to assess these three probabilities on separate occasions under circumstances that minimize recall. Coherence can also be tested in a between-groups design. Assuming random assignment, the sum of the average probabilities assigned to the two component events should equal the average judged probability of “Linda is a bank teller.” If this prediction fails, then at least some individuals are incoherent. Demonstrations of incoherence present a significant challenge to important models of decision theory and economics, which attribute to agents a very strict form of rationality (Tversky & Kahneman, 1986). Failures of perfect coherence are less provocative to psychologists, who have a more realistic view of human capabilities.

A more lenient concept, *reasoning rationality*, only requires an ability to reason correctly about the information currently at hand, without demanding perfect

consistency among beliefs that are not simultaneously evoked. The best known violation of reasoning rationality is the famous “four-card” problem (Wason, 1960). The failure of intelligent adults to reason their way through this problem is surprising because the problem is “easy,” in the sense of being easily understood once explained. What everyone learns, when first told that intelligent people fail to solve the four-card problem, is that one’s expectations about human reasoning abilities had not been adequately calibrated. There is, of course, no well-defined metric of reasoning rationality, but whatever metric one uses, the Wason problem calls for a downward adjustment. The surprising results of the Linda and engineer–lawyer problems led Tversky and Kahneman to a similar realization: The reasoning of their subjects was less proficient than they had anticipated. Many readers of the work shared this conclusion, but many others strongly resisted it.

The implicit challenge to reasoning rationality was met by numerous attempts to dismiss the findings of the engineer–lawyer and the Linda studies as artifacts of ambiguous language, confusing instructions, conversational norms, or inappropriate normative standards. Doubts have been raised about the proper interpretation of almost every word in the conjunction problem, including *bank teller*, *probability*, and even *and* (see, e.g., Dulany & Hilton, 1991; Hilton & Slugoski, 2001). These claims are not discussed in detail here. We suspect that most of them have some validity, and that they identified mechanisms that may have made the results in the engineer–lawyer and Linda studies exceptionally strong. However, we note a significant weakness shared by all these critical discussions: They provide no explanation of the essentially perfect consistency of the judgments observed in direct tests of the conjunction rule and in three other types of experiments: (1) subtle comparisons; (2) between-Ss comparisons; and, most importantly, (3) judgments of representativeness (see also Bar-Hillel and Neter, 1993, Chapter 3, this volume). Interpretations of the conjunction fallacy as an artifact implicitly dismiss the results of Fig. 2.1B as a coincidence (for an exception, see Ayton, 1998). The story of the engineer–lawyer problem is similar. Here again, multiple demonstrations in which base rate information was used (see Koehler, 1996, for a review) invite the inference that there is no general problem of base-rate neglect. Again, the data of prediction by representativeness in Fig. 2.1A (and related results reported by Kahneman & Tversky, 1973) were ignored.

The demonstrations that under some conditions people avoid the conjunction fallacy in direct tests, or use explicit base-rate information, led some scholars to the blanket conclusion that judgment biases are artificial and fragile, and that there is no need for judgment heuristics to explain them. This position was promoted most vigorously by Gigerenzer (1991). Kahneman and Tversky (1996) argued in response that the heuristics and biases position does not preclude the possibility of people performing flawlessly in particular variants of the Linda and of the lawyer–engineer problems. Because laypeople readily acknowledge the validity of the conjunction rule and the relevance of base-rate information,

the fact that they sometimes obey these principles is neither a surprise nor an argument against the role of representativeness in routine intuitive prediction. However, the study of conditions under which errors are avoided can help us understand the capabilities and limitations of System 2. We develop this argument further in the next section.

### **Making Biases Disappear: A Task for System 2**

Much has been learned over the years about variables and experimental procedures that reduce or eliminate the biases associated with representativeness. We next discuss conditions under which errors of intuition are successfully overcome, and some circumstances under which intuitions may not be evoked at all.

*Statistical Sophistication.* The performance of statistically sophisticated groups of respondents in different versions of the Linda problem illustrates the effects of both expertise and research design (Tversky and Kahneman, 1983). Statistical expertise provided no advantage in the eight-item version, in which the critical items were separated by a filler and were presumably considered separately. In the two item-version, in contrast, respondents were effectively compelled to compare “bank teller” to “bank teller and is active in the feminist movement”. The incidence of conjunction errors dropped dramatically for the statistically sophisticated in this condition, but remained essentially unchanged among the statistically naïve. Most of the experts followed logic rather than intuition when they recognized that one of the categories contained the other. In the absence of a prompt to compare the items, however, the statistically sophisticated made their predictions in the same way as everyone else does – by representativeness. As Stephen Jay Gould (1991, p. 469) noted, knowledge of the truth does not dislodge the *feeling* that Linda is a feminist bank teller: “I know [the right answer], yet a little homunculus in my head continues to jump up and down, shouting at me – ‘but she can’t just be a bank teller; read the description.’”

*Intelligence.* Stanovich and West (Chapter 24, this volume) and Stanovich (1999) observed a generally negative correlation between conventional measures of intelligence and susceptibility to judgment biases. They used transparent versions of the problems, which provide adequate cues to the correct answer and therefore provide a test of reasoning rationality. Not surprisingly, intelligent people are more likely to possess the relevant logical rules and also to recognize the applicability of these rules in particular situations. In the terms of the present analysis, high-IQ respondents benefit from relatively efficient System 2 operations that enable them to overcome erroneous intuitions when adequate information is available. When a problem is too difficult for everyone, however, the correlation is likely to reverse because the more intelligent respondents are more likely to agree on a plausible error than to respond randomly (Kahneman, 2000b).

*Frequency Format.* Relative frequencies (e.g., 1 in 10) are more vividly represented and more easily understood than equivalent probabilities (.10) or

percentages (10%). For example, the emotional impact of statements of risk is enhanced by the frequency format: "1 person in 1000 will die" is more frightening than a probability of .001 (Slovic et al., Chapter 23, this volume). The frequency representation also makes it easier to visualize partitions of sets and detect that one set is contained in another. As a consequence, the conjunction fallacy is generally avoided in direct tests, in which the frequency format makes it easy to recognize that feminist bank tellers are a subset of bank tellers (Gigerenzer & Hoffrage, 1995; Tversky & Kahneman, 1983). For similar reasons, some base-rate problems are more easily solved when couched in frequencies than in probabilities or percentages (Cosmides & Tooby, 1996). However, there is little support for the more general claims about the evolutionary adaptation of the mind to deal with frequencies (Evans et al., 2000). Furthermore, the ranking of outcomes by predicted relative frequency is very similar to the ranking of the same outcomes by representativeness (Mellers, Hertwig, & Kahneman, 2001). We conclude that the frequency format affects the corrective operations of System 2, not the intuitive operations of System 1; the language of frequencies improves respondents' ability to impose the logic of set inclusion on their considered judgments, but does not reduce the role of representativeness in their intuitions.

*Manipulations of Attention.* The weight of neglected variables can be increased by drawing attention to them, and experimenters have devised many ingenious ways to do so. Schwarz et al. (1991) found that respondents pay more attention to base-rate information when they are instructed to think as statisticians rather than clinical psychologists. Krosnick, Li, and Lehman (1990), exploited conversational conventions about the sequencing of information and confirmed that the impact of base-rate information was enhanced by presenting that information *after* the personality description rather than before it. Attention to the base-rate is also enhanced when participants observe the drawing of descriptions from an urn (Gigerenzer, Hell, & Blank, 1988), perhaps because watching the draw induces conscious expectations that reflect the known proportions of possible outcomes. The conjunction fallacy can also be reduced or eliminated by manipulations that increase the accessibility of the relevant rule, including some linguistic variations (Macchi, 1995), and practice with logical problems (Agnoli, 1991; Agnoli & Krantz, 1989).

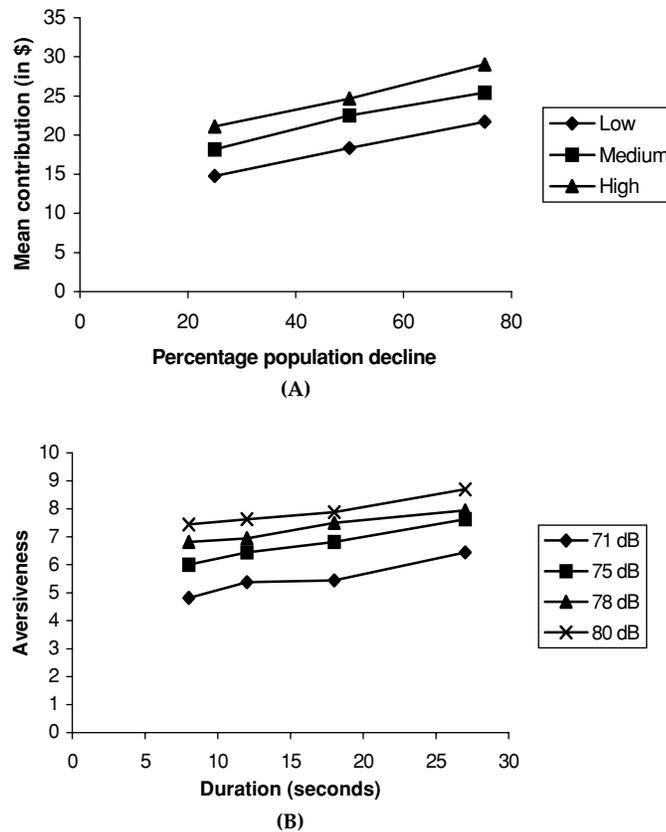
The interpretation of these attentional effects is straightforward. We assume that most participants in judgment studies know, at least vaguely, that the base-rate is relevant and that the conjunction rule is valid (Kahneman & Tversky, 1982). Whether they apply this knowledge to override an intuitive judgment depends on their cognitive skills (education, intelligence) and on formulations that make the applicability of a rule apparent (frequency format) or a relevant factor more salient (manipulations of attention). We assume that intuitions are less sensitive to these factors, and that the appearance or disappearance of biases mainly reflects variations in the efficacy of corrective operations. This conclusion would be circular, of course, if the corrective operations were both

inferred from the observation of correct performance and used to explain that performance. Fortunately, the circularity can be avoided – because the role of System 2 can be verified for example, by using manipulations of time pressure, cognitive load, or mood to interfere with its operations.

*Within-Subjects Factorial Designs.* The relative virtues of between-subjects and within-subjects designs in studies of judgment are a highly contentious issue. Factorial designs have their dismissive critics (e.g., Poulton, 1989) and their vigorous defenders (e.g., Birnbaum, 1999). We do not attempt to adjudicate this controversy here. Our narrower point is that between-subjects designs are more appropriate for the study of heuristics of judgment. The following arguments favor this conclusion:

- Factorial designs are transparent. Participants are likely to identify the variables that are manipulated – especially if there are many trials and especially in a fully factorial design, in which the same stimulus attributes are repeated in varying combinations. The message that the design conveys to the participants is that the experimenter expects to find effects of every factor that is manipulated (Bar-Hillel & Fischhoff, 1981; Schwarz, 1996).
- Studies that apply a factorial design to judgment tasks commonly involve schematic and impoverished stimuli. The tasks are also highly repetitive. These features encourage participants to adopt simple mechanical rules that will allow them to respond quickly, without forming an individuated impression of each stimulus. For example, Ordóñez and Benson (1997) required respondents to judge the attractiveness of gambles on a 100-point scale. They found that under time pressure many respondents computed or estimated the expected values of the gambles and used the results as attractiveness ratings (e.g., a rating of 15 for a 52% chance to win \$31.50).
- Factorial designs often yield judgments that are linear combinations of the manipulated variables. This is a central conclusion of a massive research effort conducted by Anderson and colleagues (see Anderson, 1996), who observed that people often average or add where they should multiply.

In summary, the factorial design is not appropriate for testing hypotheses about biases of neglect, because it effectively guarantees that no manipulated factor is neglected. Figure 2.3 illustrates this claim by several examples of an additive extension effect discussed further in the next section. The experiments summarized in the different panels share three important features: (1) In each case, the quantitative variable plotted on the abscissa was completely neglected in similar experiments conducted in a between-subjects or subtle design; (2) in each case, the quantitative variable combines additively with other information; (3) in each case, a compelling normative argument can be made for a quasi-multiplicative rule in which the lines shown in Fig. 2.3 should fan out. For example, Fig. 2.3C presents a study of categorical prediction (Novemsky & Kronzon, 1999) in which the respondent judged the relative likelihood that a person was a member of one occupation rather than another (e.g., computer



**Figure 2.3.** (A) Willingness to pay to restore damage to species that differ in popularity as a function of the damage they have suffered (from Kahneman, Ritov, & Schkade, 1999); (B) Global evaluations of aversive sounds of different loudness as a function of duration for subjects selected for their high sensitivity to duration (from Schreiber & Kahneman, 2000); (C) Ratings of probability for predictions that differ in representativeness as a function of base-rate frequency (from Novemsky & Kronzon, 1999); (D) Global evaluations of episodes of painful pressure that differ in temporal profile as a function of duration (Ariely, 1998).

programmer vs. flight attendant) on the basis of short personality sketches (e.g., “shy, serious, organized, and sarcastic”) and one of three specified base rates (10%, 50%, or 90%). Representativeness and base-rate were varied factorially within subjects. The effect of base-rate is clearly significant in this design (see also Birnbaum and Mellers, 1983). Furthermore, the effects of representativeness and base-rate are strictly additive. As Anderson (1996) argued, averaging (a special case of additive combination) is the most obvious way to combine the effects of two variables that are recognized as relevant, e.g., “she

72 Daniel Kahneman and Shane Frederick

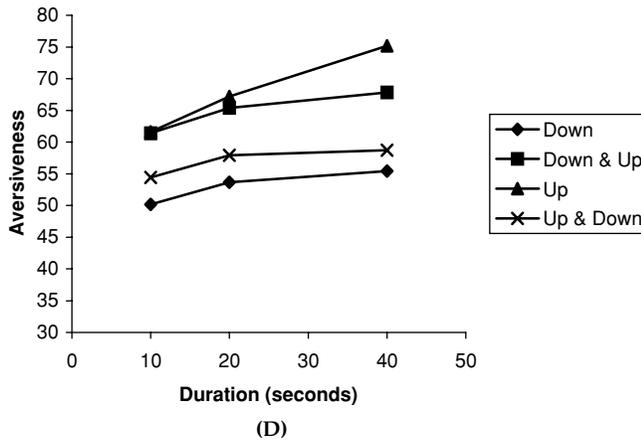
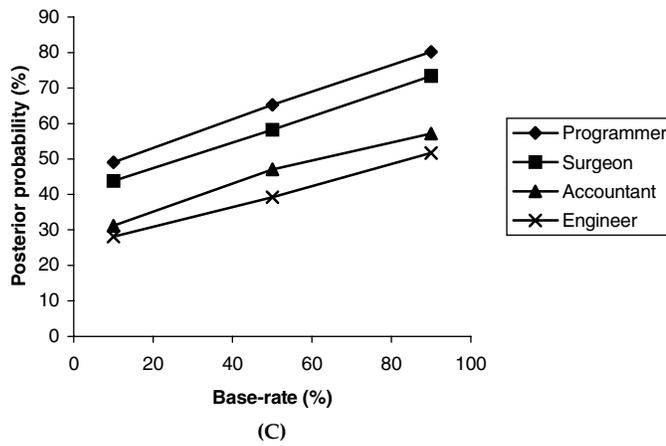


Figure 2.3. (continued)

looks like a bank teller, but the base-rate is low.” Additivity is not normatively appropriate in this case – any Bayes-like combination would produce curves that initially fan out from the origin and converge again at high values. Similar considerations apply to the other three panels of Fig. 2.3 (discussed later).

Between-subjects and factorial designs often yield different results in studies of intuitive judgment. Why should we believe one design rather than the other? The main argument against the factorial design is its poor ecological validity. Rapidly successive encounters with objects of rigidly controlled structure are unique to the laboratory, and the solutions that they evoke are not likely to be typical. Direct comparisons among concepts that differ in only one variable – such as *bank teller* and *feminist bank tellers* – also provide a powerful hint and a highly unusual opportunity to overcome intuitions. The between-subjects

design in contrast, mimics the haphazard encounters in which most judgments are made and is more likely to evoke the casually intuitive mode of judgment that governs much of mental life in routine situations (e.g., Langer, 1978).

### **PROTOTYPE HEURISTICS AND THE NEGLECT OF EXTENSION**

In this section, we offer a common account of three superficially dissimilar judgmental tasks: (1) categorical prediction (e.g., "In a set of 30 lawyers and 70 engineers, what is the probability that someone described as 'charming, talkative, clever, and cynical' is one of the lawyers?"); (2) summary evaluations of past events (e.g., "Overall, how aversive was it to be exposed for 30 minutes to your neighbor's car alarm?"); and (3) economic valuations of public goods (e.g., "What is the most you would be willing to pay to prevent 200,000 migrating birds from drowning in uncovered oil ponds?"). We propose that a generalization of the representativeness heuristic accounts for the remarkably similar biases that are observed in these diverse tasks.

The original analysis of categorical prediction by representativeness (Kahneman & Tversky 1973; Tversky & Kahneman, 1983) invoked two assumptions in which the word *representative* was used in different ways: (1) a prototype (a *representative exemplar*) is used to represent categories (e.g. bank tellers) in the prediction task; (2) the probability that the individual belongs to a category is judged by the degree to which the individual resembles (is *representative of*) the category stereotype. Thus, categorical prediction by representativeness involves two separate acts of substitution – the substitution of a prototypical exemplar for a category, and the substitution of the heuristic attribute of similarity for the target attribute of probability. Perhaps because they share a label, the two processes have not been distinguished in discussions of the representativeness heuristic. We separate them here by describing *prototype heuristics*, in which a prototype is substituted for its category, but in which *representativeness* is not necessarily the heuristic attribute.

The target attributes to which prototype heuristics are applied are extensional. An *extensional attribute* pertains to an aggregated property of a set or category for which an extension is specified – the probability that a set of lawyers includes Jack; the overall unpleasantness of a set of moments of hearing a car alarm; and the personal value of saving a certain number of birds from drowning in oil ponds. Normative judgments of extensional attributes are governed by a general principle of *conditional adding*, which dictates that each element of the set adds to the overall judgment an amount that depends on the elements already included. In simple cases, conditional adding is just regular adding – the total weight of a collection of chairs is the sum of their individual weights. In other cases, each element of the set contributes to the overall judgment, but the combination rule is not simple addition and is typically subadditive. For example, the economic value of protecting  $X$  birds should be increasing in  $X$ ,

but the value of saving 2,000 birds is for most people less than twice as large as the value of saving 1,000 birds.

The logic of categorical prediction entails that the probability of membership in a category should vary with its relative size, or base-rate. In prediction by representativeness, however, the representation of outcomes by prototypical exemplars effectively discards base-rates, because the prototype of a category (e.g., lawyers) contains no information about the size of its membership. Next, we show that phenomena analogous to the neglect of base-rate are observed in other prototype heuristics: the monetary value attached to a public good is often insensitive to its *scope* and the global evaluations of a temporally extended experience is often insensitive to its *duration*. These various instantiations of *extension neglect* (neglect of base rates, scope, and duration) have been discussed in separate literatures, but all can be explained by the two-part process that defines prototype heuristics: (1) a category is represented by a prototypical exemplar, and (2) a (nonextensional) property of the prototype is then used as a heuristic attribute to evaluate an extensional target attribute of the category. As might be expected from the earlier discussion of base-rate neglect, extension neglect in all its form is most likely to be observed in between-subjects experiments. Within-subject factorial designs consistently yield the *additive extension effect* illustrated in Fig. 2.3.

#### **Scope Neglect in Willingness to Pay**

The contingent valuation method (CVM) was developed by resource economists (see Mitchell & Carson, 1989) as a tool for assessing the value of public goods for purposes of litigation or cost-benefit analysis. Participants in contingent valuation (CV) surveys are asked to indicate their willingness to pay (WTP) for specified public goods, and their responses are used to estimate the total amount that the community would pay to obtain these goods. The economists who design contingent valuation surveys interpret WTP as a valid measure of economic value and assume that statements of WTP conform to the extensional logic of consumer theory. The relevant logic has been described by a critic of CVM (Diamond, 1996), who illustrates the conditional adding rule by the following example: in the absence of income effects, WTP for saving  $X$  birds should equal WTP for saving  $(X-k)$  birds, plus WTP to save  $k$  birds, where the last value is contingent on the costless prior provision of safety for  $(X-k)$  birds (Diamond, 1996).

Strict adherence to Bayes' rule may be an excessively demanding standard for intuitive predictions; similarly, it would be too much to ask for WTP responses that strictly conform to the "add-up rule." In both cases, however, it seems reasonable to expect *some* sensitivity to extension – to the base rate of outcomes in categorical prediction and to the scope of the good in WTP. In fact, several studies have documented nearly complete neglect of scope in CV surveys. The best-known demonstration of scope neglect is an experiment by Desvougues et al. (1993), who used the scenario of migratory birds that drown

in oil ponds. The number of birds said to die each year was varied across groups. The WTP responses were completely insensitive to this variable, as the mean WTPs for saving 2,000, 20,000, or 200,000 birds were \$80, \$78, and \$88, respectively.

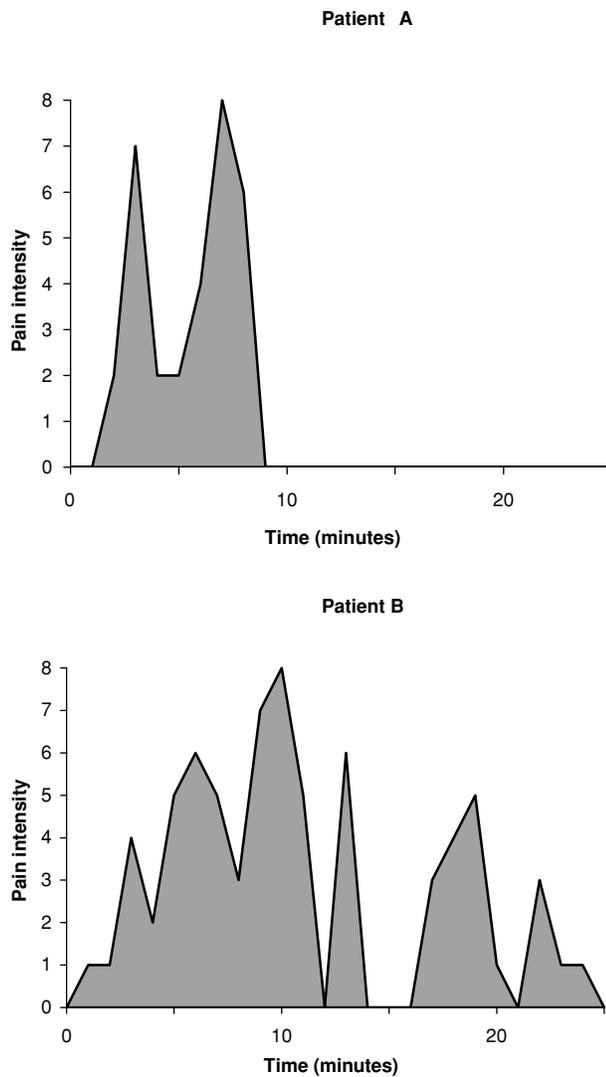
A straightforward interpretation of this result involves the two acts of substitution that characterize prototype heuristics. The deaths of numerous birds are first represented by a prototypical instance, perhaps an image of a bird soaked in oil and drowning. The prototype automatically evokes an affective response, and the intensity of that emotion is then mapped onto the dollar scale – substituting the readily accessible heuristic attribute of affective intensity for the more complex target attribute of economic value. Other examples of radical insensitivity to scope lend themselves to a similar interpretation. Among others, Kahneman and Knetsch (see Kahneman, 1986) found that Toronto residents were willing to pay almost as much to clean up polluted lakes in a small region of Ontario as to clean up all the polluted lakes in Ontario, and McFadden and Leonard (1993) reported that residents in four western states were willing to pay only 28% more to protect 57 wilderness area than to protect a single area (for more discussion of scope insensitivity, see Frederick & Fischhoff, 1998).

The similarity between WTP statements and categorical predictions is not limited to such demonstrations of almost complete extension neglect. The two responses also yield similar results when extension and prototype information are varied factorially within subjects. Panel (a) of Figure 2.3 shows the results of a study of WTP for programs that prevented different levels of damage to species of varying popularity (Ritov and Kahneman, unpublished observations; cited in Kahneman, Ritov & Schkade, 1999). As in the case of base-rate (Fig. 2.3, panel c), extensional information (levels of damage) combines additively with nonextensional information. This rule of combination is unreasonable; in any plausible theory of value the lines would fan out.

Finally, the role of the emotion evoked by a prototypical instance was also examined directly in the same experiment, using the heuristic elicitation paradigm introduced earlier. Some respondents were asked to imagine that they saw a television program documenting the effect of adverse ecological circumstances on individual members of different species. The respondents indicated, for each species, how much concern they expected to feel while watching such a documentary. The correlation between this measure of affect and willingness to pay, computed across species, was .97.

### **Duration Neglect in the Evaluation of Experiences**

We next discuss experimental studies of the global evaluation of experiences that extend over some time, such as a pleasant or a horrific film clip (Fredrickson & Kahneman, 1993), a prolonged unpleasant noise (Schreiber & Kahneman, 2000), pressure from a vise (Ariely, 1998), or a painful medical procedure (Redelmeier & Kahneman, 1996). Participants in these studies provided



**Figure 2.4.** Pain intensity reported by two colonoscopy patients.

a continuous or intermittent report of hedonic or affective state, using a designated scale of momentary affect (Fig. 2.4). When the episode had ended, they indicated a global evaluation of “the *total* pain or discomfort” associated with the entire episode.

We first examine the normative rules that apply to this task. The global evaluation of a temporally extended outcome is an extensional attribute, which is governed by a distinctive logic. The most obvious rule is temporal monotonicity: there is a compelling intuition that adding an extra period of pain to an episode

of discomfort can only make it worse overall. Thus, there are two ways of making a bad episode worse – making the discomfort more intense or prolonging it. It must therefore be possible to trade off intensity against duration. Formal analyses have identified conditions under which the total utility of an episode is equal to the temporal integral of a suitably transformed measure of the instantaneous utility associated with each moment (Kahneman, 2000d; Kahneman, Wakker, & Sarin, 1997).

Next, we turn to the psychology. Fredrickson and Kahneman (1993) proposed a “snapshot model” for the retrospective evaluation of episodes, which again involves two acts of substitution: first, the episode is represented by a prototypical moment; next, the affective value attached to the representative moment is substituted for the extensional target attribute of global evaluation. The snapshot model was tested in an experiment in which participants provided continuous ratings of their affect while watching plotless films that varied in duration and affective value (e.g., fish swimming in coral reefs; pigs being beaten to death with clubs), and later reported global evaluations of their experiences. The central finding was that the retrospective evaluations of these observers were predicted with substantial accuracy by a simple average of the Peak Affect recorded during a film and the End Affect reported as the film was about to end. This has been called the *Peak/End Rule*. However, the correlation between retrospective evaluations and the duration of the films was negligible, a finding that Fredrickson and Kahneman labeled *duration neglect*. The resemblance of duration neglect to the neglect of scope and base-rate is striking, and unlikely to be accidental. In the present analysis, all three are manifestations of extension neglect, caused by the use of a prototype heuristic.

The Peak/End Rule and duration neglect have both been confirmed on multiple occasions. Figure 2.4 presents raw data from a study reported by Redelmeier and Kahneman (1996), in which patients undergoing colonoscopy reported their current level of pain every 60 seconds throughout the procedure. Here again, an average of Peak/End pain quite accurately predicted subsequent global evaluations and choices. The duration of the procedure varied considerably among patients (from 4 to 69 minutes), but these differences were not reflected in subsequent global evaluations in accord with duration neglect. The implications of these psychological rules of evaluation are paradoxical. In Fig. 2.4, for example, it appears evident that patient B had a worse colonoscopy than patient A (assuming that they used the scale similarly). However, it is also apparent that the Peak/End average was worse for patient A, whose procedure ended at a moment of relatively intense pain. The Peak/End rule prediction for these two profiles is that A would evaluate the procedure more negatively than B, and would be more likely to prefer to undergo a barium enema rather than a repeat colonoscopy. The prediction was correct for these two individuals, and confirmed by the data of a large group of patients.

The effects of substantial variations of duration remained small (though statistically robust) even in studies conducted in a factorial design. Figure 2.3D is

78    Daniel Kahneman and Shane Frederick

drawn from a study of responses to ischemic pain (Ariely, 1998) in which duration varied by a factor of 4. The Peak/End average accounted for 98% of the systematic variance of global evaluations in that study and 88% of the variance in a similar factorial study of responses to loud unpleasant sounds (Schreiber & Kahneman, 2000, panel 3b). Contrary to the normative standard for an extensional attribute, the effects of duration and other determinants of evaluation were additive (see panels b and d in Figure 3).

The participants in these studies were well aware of the relative duration of their experiences and did not consciously decide to ignore duration in their evaluations. As Fredrickson and Kahneman (1993, p. 54) noted, duration neglect is an attentional phenomenon:

[D]uration neglect does not imply that duration information is lost, nor that people believe that duration is unimportant. . . . people may be aware of duration and consider it important in the abstract [but] what comes most readily to mind in evaluating episodes are the salient moments of those episodes and the affect associated with those moments. Duration neglect might be overcome, we suppose, by drawing attention more explicitly to the attribute of time.

This comment applies equally well to other instances of extension neglect: the neglect of base-rate in categorical prediction, the neglect of scope in willingness to pay, the neglect of sample size in evaluations of evidence (Griffin & Tversky, 1992; Tversky & Kahneman, 1971), and the neglect of probability of success in evaluating a program of species preservation (DeKay & McClelland, 1995). More generally, inattention plays a similar role in any situation in which the intuitive judgments generated by System 1 violate rules that would be accepted as valid by the more deliberate reasoning that we associate with System 2. As we noted earlier, the responsibility for these judgmental mishaps is properly shared by the two systems: System 1 produces the initial error, and System 2 fails to correct it, although it could.

### **Violations of Dominance**

The conjunction fallacy observed in the Linda problem is an example of a dominance violation in judgment: Linda must be at least as likely to be a bank teller as to be a feminist bank teller, but people believe the opposite. Insensitivity to extension (in this case, base-rate) effectively guarantees the existence of such dominance violations. For another illustration, consider the question, "How many murders were there last year in [Detroit/Michigan]?" Although there cannot be more murders in Detroit than in Michigan because Michigan contains Detroit, the word *Detroit* evokes a more violent image than the word *Michigan* (except, of course, for people who immediately think of Detroit when Michigan is mentioned). If people use an impression of violence as a heuristic and neglect geographic extension, their estimates of murders in the city may exceed their estimates for the state. In a large sample of University of Arizona students, this hypothesis was confirmed – the median estimate of the number of murders was 200 for Detroit, and 100 for Michigan.

Violations of dominance akin to the conjunction fallacy have been observed in several other experiments, involving both indirect (between-subjects) and direct tests. In a clinical experiment reported by Redelmeier, Katz, and Kahneman (2001), half of a large group of patients ( $N = 682$ ) undergoing a colonoscopy were randomly assigned to a condition that made the actual experience strictly worse. Unbeknownst to the patient, the physician deliberately delayed the removal of the colonoscope for approximately 1 minute beyond the normal time. For many patients, the mild discomfort of that extra period was an improvement relative to the pain than they had just experienced. For these patients, of course, prolonging the procedure reduced the Peak/End average of discomfort. As expected, retrospective evaluations were less negative in the experimental group. Remarkably, a 5-year follow-up showed that participants in that group were also significantly more likely to comply with recommendations to undergo a repeat colonoscopy (Redelmeier, Katz, & Kahneman, 2001).

In an experiment that is directly analogous to the demonstrations of the conjunction fallacy, Kahneman et al. (1993) exposed participants to two cold-pressor experiences, one with each hand: a short episode (immersion of one hand in  $14^{\circ}\text{C}$  water for 60 seconds), and a long episode (the short episode plus an additional 30 seconds during which the water was gradually warmed to  $15^{\circ}\text{C}$ ). The participants indicated the intensity of their pain throughout the experience. When they were later asked which of the two experiences they preferred to repeat, a substantial majority chose the long trial. These choices violate dominance, because after 60 seconds in cold water most people prefer the immediate experience of a warm towel to 30 extra seconds of slowly diminishing pain. In a replication, Schreiber and Kahneman (2000, Exp. 2) exposed participants to pairs of unpleasant noises in immediate succession. The participants listened to both sounds and chose one to be repeated at the end of the session. The short noise lasted 8 seconds at 77 db. The long noise consisted of the short noise plus an extra period (of up to 24 sec) at 66 db (less aversive, but still unpleasant and certainly worse than silence). Here again, the longer noise was preferred most of the time, and this unlikely preference persisted over a series of five choices.

The violations of dominance in these direct tests are particularly surprising because the situation is completely transparent. The participants in the experiments could easily retrieve the durations of the two experiences between which they had to choose, but the results suggest that they simply ignored duration. A simple explanation is that the results reflect "choosing by liking" (see Frederick, Chapter 30, this volume). The participants in the experiments simply followed the normal strategy of choice: "When choosing between two familiar options, consult your retrospective evaluations and choose the one that you like most (or dislike least)." *Liking* and *disliking* are products of System 1 that do not conform to the rules of extensional logic. System 2 could have intervened, but in these experiments, it generally did not. Kahneman et al. (1993) described a participant in their study who chose to repeat the long cold-pressor experience. Soon after the choice was recorded, the participant was asked which of the two

experiences was longer. As he correctly identified the long trial, the participant was heard to mutter, "The choice I made doesn't seem to make much sense." *Choosing by liking* is a form of mindlessness (Langer, 1978) that illustrates the casual governance of System 2.

Like the demonstrations of the conjunction fallacy in direct tests (discussed earlier), violations of temporal monotonicity in choices should be viewed as an expendable flourish. Because the two aversive experiences occurred within a few minutes of each other, and because respondents could recall accurately the duration of the two events, System 2 had enough information to override choosing by liking. Its failure to do so is analogous to the failure of people to appreciate the set inclusion rule in direct tests of the conjunction fallacy. In both cases, the violations of dominance tell us nothing new about System 1; they only illustrate an unexpected weakness of System 2. Just as the theory of intuitive categorical prediction would have remained intact if the conjunction fallacy had not "worked" in a direct test, the model of evaluation by moments would have survived even if violations of dominance had been eliminated in highly transparent situations. The same methodologic issues arise in both contexts. Between-subjects experiments or subtle tests are most appropriate for studying the basic intuitive evaluations of System 1, and also most likely to reveal complete extension neglect. Factorial designs in which extension is manipulated practically guarantee an effect of this variable, and almost guarantee that it will be additive, as in Figures 3b and 3d (Ariely, 1998; Ariely, Kahneman, & Loewenstein, 2000; Schreiber and Kahneman, 2000). Finally, although direct choices sometimes yield systematic violations of dominance, these violations can be avoided by manipulations that prompt System 2 to take control.

In our view, the similarity of the results obtained in diverse contexts is a compelling argument for a unified interpretation, and a significant challenge to critiques that pertain only to selected subsets of this body of evidence. A number of commentators have offered competing interpretations of base-rate neglect (Cosmides & Tooby, 1996; Koehler, 1996), insensitivity to scope in WTP (Kopp, 1992), and duration neglect (Ariely & Loewenstein, 2000). However, these interpretations are generally specific to a particular task and would not carry over to analogous findings in other domains. Similarly, the various attempts to explain the conjunction fallacy as an artifact do not explain analogous violations of dominance in the cold-pressor experiment. The account we offer is, in contrast, equally applicable to all three contexts and possibly others as well (see also Kahneman, Ritov, & Schkade, 1999). We attribute extension neglect and violations of dominance to a lazy System 2 and a prototype heuristic that combines two processes of System 1: the representation of categories by prototypes and the substitution of a nonextensional heuristic attribute for an extensional target attribute. We also propose that people have some appreciation of the role of extension in the various judgment tasks. Consequently, they incorporate extension in their judgments when their attention is drawn to this factor – most reliably in factorial experiments and sometimes (although not always) in

direct tests. The challenge for competing interpretations is to provide a unified account of the diverse phenomena that have been considered in this section.

### **FINAL COMMENTS**

The goal of the heuristics and biases program in its early days was to understand intuitive judgment under uncertainty, not develop a unified theory of it. *Judgment heuristics* were described as a collection of disparate cognitive procedures that are bound together by their common function in a particular domain. In contrast, we argue here that heuristics share a common process of attribute substitution and are not limited to questions about uncertain events. Our treatment is otherwise firmly anchored in previous work. The substitution of one question for another, the representation of categories by prototypes, the view of erroneous intuitions as easy to override but almost impossible to eradicate – all these ideas are quite old (Kahneman, Slovic, & Tversky, 1982). We show here that the same ideas apply to a diverse class of difficult judgments, including retrospective evaluations of colonoscopies and decisions about saving birds from drowning in oil.

Ironically, the original research program might have been less interesting, and less influential, if the scope of its central concept had been appreciated and made explicit at the outset. The attention that the program attracted was due in no small part to an expository style in which illustrative examples embedded in the text turn readers into observers. A more general treatment could not have exploited this style because broad categories, in the classic Roschian sense, do not have ‘good’ exemplars. Just as a robin is a good bird but a mediocre animal, the Linda problem is a compelling illustration of a conjunction fallacy, but a less obvious example of a violation of dominance. The restriction of the original treatment of judgment heuristics to the domain of uncertainty was therefore as fortunate as it was arbitrary. We believe that the interpretation of representativeness and of other heuristics that we have offered here is both more complete and more general than earlier versions. But the original concepts, at least for us, had a charm that this more elaborate analysis cannot match.