16

Which World Should Be Represented in Representative Design?

Ulrich Hoffrage and Ralph Hertwig

University of California, Berkeley campus, early 1940s. A man and a young woman are strolling across the university campus. Suddenly the man stops and says, "Now." The woman also stops, raises her arm, points to the building in front of her, and says, "The window on the second floor." After contemplating the scene for a moment, she says "Three-and-a-half feet." The man records this and other numbers. They begin walking again. A few moments later, the man stops once again and the same interaction repeats itself.

University of Constance, vision laboratory, mid-1980s. A young man enters a dark room. The experimenter welcomes him and then asks him to sit down in front of a large cubicle, to put his chin on a rest (thus making it impossible to move his head), and to look into the cubicle through two small holes. Its interior is completely dark. Suddenly a white square appears, only to disappear a few seconds later. Then, a tone is heard, which is followed by another white square to the right of the previous one. The young man says, "The left one." This episode frequently repeats itself over the course of the next half hour.

Although both scenes represent the same genre, namely, psychological studies on the perception of object size, they share few common features. As we shall see shortly, their differences can be traced back to the fact that the Berkeley study was conceived by Egon Brunswik (1944; for a replication, see Dukes, 1951), whereas a contemporary experimental psychologist conducted the Constance study. In Brunswik's study, the participant (Miss Johanna R. Goldsmith, a graduate student in psychology) was interrupted at randomly chosen intervals during the course of her daily activities, in various outdoor and indoor situations (on the street, on campus, in the laboratory, at home, etc.). Once interrupted, she was asked to "indicate which linear extension happened to be most conspicuous to her at the moment" (Brunswik, 1944; cited after Hammond & Stewart, 2001, p. 70). Brunswik

summarized the sample of situations thus obtained as follows: Although it may "not be perfectly representative of 'life', there is no doubt that it is more representative and variegated with regard to size, distance, proportion, color, surrounding patterns of objects, and other characteristics than any laboratory design could hope to be" (Brunswik, 1944; cited in Hammond & Stewart, 2001, p. 70).

It seems fair to conclude that the study in the vision laboratory was not representative of "life" – indeed, this dimension was probably not even considered. Rather, the experimenter created highly constrained situations in which he had full control over variables, such as the size of the objects, their distance, and their colors. He also determined whether the participant's vision was binocular or monocular. (The participant was the first author, who was required to participate in a number of psychological studies as an undergraduate.) Whereas Brunswik aimed to sample situations that were representative of the actual demands of the environment in which the individual lives, the Constance study investigated situations that rarely occur outside the artificial world of the vision lab. For Brunswik, the key concern was being able to generalize the results to the person's natural habitat. Accordingly, he sampled situations directly from the person's habitat. In contrast, for many experimentalists – today as well as in Brunswik's time – the prime objective is to carefully disentangle possible variables that affect a person's perceptual and cognitive performance. Consequently, experimentalists create microworlds that may never occur in reality, but over which they exert full control.

The topic of this chapter is what we believe to be the constructive tension between these two paradigms in psychological experimentation. We begin with a short historical review of the notion of *representative design*, Brunswik's term for an experimental design that aims for a veridical representation of the environment in which organisms naturally perform. Then we ask whether representative design matters for the results obtained. Lastly, we identify a key conceptual difficulty of representative design, namely, the issue of how to define the reference class from which situations are sampled. We demonstrate how different reference classes may lead to different conclusions and discuss possible solutions for this problem.

BRUNSWIK'S CRITIQUE OF PSYCHOLOGISTS' WAY OF CONDUCTING BUSINESS

The methodological dictate that has ruled experimental practices since the birth of experimental psychology is *systematic design*. According to its rationale, experimenters select and isolate one or a few independent variables that are varied systematically, while holding all other variables constant or allowing them to vary randomly. Experimenters then observe the resulting changes in the dependent variable(s), thus hoping to identify

cause–effect relationships. Systematic design bets on internal validity, that is, on the sound demonstration that a causal relationship exists between two or more variables, rather than on external validity, which refers to the generalizability of the causal relationship beyond the experimental context. After all, so the logic goes, if internal validity is not guaranteed, no conclusions can be drawn with confidence about the effects of the independent variable. This logic was espoused by the founding fathers of experimental psychology, for instance, by Hermann Ebbinghaus, and has frequently been propagated in textbooks on experimental psychology. Take, for example, Woodworth's (1938) classic textbook *Experimental Psychology*, also known as the *Columbia Bible*, which promoted systematic design as *the* experimental tool (see Gillis & Schneider, 1966, for the historical roots of systematic design).

Brunswik opposed psychology's default experimental method (Kurz & Tweney, 1997). He questioned both the feasibility of disentangling variables and the realism of the stimuli created in doing so. In what he considered to be the simplest variant of systematic design, the one-variable design, Brunswik (1944, 1955, 1956) pointed to the fact that variables were "tied," thus making it impossible to determine the exact cause of an observed effect. To illustrate this, he referred to classic experiments in psychophysics. Using the Galton bar, experimenters presented lines of different lengths and requested observers to estimate their size. The distance between the experimental and the observed stimuli is held constant. Consequently, physical and retinal size are artificially "tied" because a small object will project a smaller retinal size and a large object will project a larger retinal size. Of course, such ties can be disentangled by using a more sophisticated variant of systematic design, namely, factorial design. Here, the levels of one variable (e.g., physical size) are combined with the levels of another variable (retinal size), exhausting all possible combinations. Brunswik criticized factorial design on the grounds that even if two or more variables are untied, each variable still remains tied to many other factors that may affect the organism's achievement. In his view, complete systematic isolation of one variable as the crucial factor would involve the combination of this variable with a "very large, and in fact indefinite, number of originally tied situational variables" (Brunswik, 1955, p. 197). He also argued that factorial design, because it aims to control for all variables that are not being investigated, destroys the natural covariation among variables. Therefore, factorial design inhibits the researcher's ability to examine the level of achievement that an organism reaches within its habitat.

Probabilistic Functionalism and Representative Design

Brunswik's methodological convictions were closely intertwined with his theoretical outlook (Hammond & Stewart, 2001; in particular, Kurz &

Hertwig, 2001). To fully appreciate Brunswik's criticism, it is helpful to view it in combination with the theoretical framework that he advocated. In his theory of *probabilistic functionalism*, Brunswik (1943, 1952) argued that the real world is an important consideration in experimental research because psychological processes are adapted in a Darwinian sense to the environments in which they evolve and function (Hammond, 1996). To the extent that psychology's objective is to study the adjustment of organisms to the environment in which they actually live, any test that is implemented to study adjustment should, according to Brunswik (1944), ensure "that the habitat of the individual, group, or species is represented with all of its variables, and that the specific values of these variables are kept in accordance with the frequencies in which they actually happen to be distributed" (cited in Hammond & Stewart, 2001, p. 69).

The ecology that an organism adapts to is not perfectly predictable for the organism (Brunswik, 1943). A particular distal stimulus, for instance, does not always imply the same specific proximal effects. Sometimes a specific proximal effect does not occur despite the presence of the distal stimulus. Similarly, particular proximal effects do not always imply a specific distal stimulus because sometimes the same effect may be caused by other distal stimuli. Proximal cues are therefore only probabilistic indicators of a distal variable. Brunswik (1940, 1952) proposed measuring the ecological (or predictive) validity of a cue by the correlation between the cue and the distal variable. The proximal cues are themselves interrelated, thus introducing redundancy (or intra-ecological correlations) into the environment. This redundancy, in turn, is the basis for *vicarious functioning* – a principle that Brunswik (1952) considered as being the foundation on which the adaptive system is built. Since a cue is not always present, an adaptive system has to rely on multiple cues that can be substituted for each other. Systematic design, with its policy of isolating and controlling variables, risks destroying the causal texture of the environment to which an organism has adapted (Brunswik, 1944), and thus, ultimately, the process of various functioning. To keep the process intact, Brunswik (1956) thought it necessary to sample experimental situations that are representative of a defined population of stimuli in terms of their *number*, *values*, *distributions*, *intercorrelations*, and *ecological validities* of their variable components. Otherwise, the obtained results are no longer representative of the organism's actual functioning in its habitat.

Brunswik (1955) suggested three ways of achieving a representative design. The first and preferred way is by *random sampling* (also referred to as situational, representative, and natural sampling) of stimuli from a defined population of stimuli (reference class) to which the experimenter wishes to generalize the findings. This is one form of what we today call probability sampling, where each stimulus has an equal probability of being selected. Note that time sampling is not synonymous with random sampling

because intervals may be sampled systematically (for a review of this sampling method, see Czikszentmilhalyi & Larson, 1992). Brunswik's study that we described at the beginning used time sampling; for another application of this method see Hogarth's (2005) chapter in the present book. The second way of achieving a representative design is through what Brunswik (1955, 1956) called the *canvassing* of stimuli and what today is known as nonprobability sampling, namely, stratified, quota, proportionate, or accidental sampling (Baker, 1988). However, these procedures only provide a "primitive type of coverage of the ecology" (Brunswik, 1955, p. 204). Perhaps the most desirable, yet least feasible, way of achieving a representative design aims for a *complete coverage* of the entire population of stimuli.

The fundamental shift in the method of psychology that Brunswik (1943) has called for did not occur. His contemporaries were united in their wish to maintain the status quo (Postman, 1955). Despite the fact that his ideas were published in leading journals, they were largely ignored, misunderstood, and treated with skepticism and hostility (Feigl, 1955; Hilgard, 1955; Hull, 1943; Krech, 1955; Postman, 1955). His colleagues' hostile response to the notion of representative design is not without irony when one considers that sampling (albeit the sampling of subjects, not of objects) has been considered a *sine qua non* of psychological experimentation. In fact, Brunswik (1943, 1944) called attention to the "double-standard" in the practice of sampling in psychological research, pointing out that the entire problem of generalization was thrown "onto the responder rather than onto the situation" (Brunswik, 1955, p. 195). To quote Hammond (1998):

> Why, he wanted to know, is the logic we demand for generalization over the subject side ignored when we consider the input or environment side? Why is it that psychologists scrutinize subject sampling procedures carefully but cheerfully generalize their results – without any logical defense – to conditions outside those used in the laboratory. (p. 2)

It seems that many decades later, the same double-standard lives on, and the issue of generalization still defies solution by the conventional methods of experimental design. In their incisive critique of current social psychological experimentation, Wells & Windschitl (1999) point to the neglect of stimulus sampling as a "serious problem that plagues a surprising number of experiments" (p. 1115). Stimulus sampling, so they argued, is imperative whenever individual instances within categories (e.g., gender or race) vary from one another in ways that affect the dependent variable. For example, relying on only one or two male confederates to test the hypothesis that men are more courteous to women than to men "can confound the unique characteristics of the selected stimulus with the category," and "what might be portrayed as a category effect could in fact be due to the unique characteristics of the stimulus selected to represent

that category" (p. 1116). Indeed, in his own studies on social perception, Brunswik advocated the importance of presenting respondents with a representative sample of "person-objects" (a phrase referring to the person to be judged) (Brunswik, 1945, 1956; Brunswik & Reiter, 1937; for work in social perception in the tradition of Brunswik, see Funder, 1995).

In conclusion, Brunswik stressed two major shortcomings of the systematic design that he hoped to remedy with the representative design: First, systematic design does not allow researchers to elicit and study the process of vicarious functioning – the defining mark of an adaptive system. Second, as a consequece, experimenters who rely on systematic design cannot generalize their findings to the organisms' natural ecology. Next, we discuss the extent to which sampling experimental objects is as crucial for the results obtained as Brunswik believed.

DOES SAMPLING OF EXPERIMENTAL STIMULI MATTER?

Although Brunswik (1956) spelled out the theoretical rationale for representative design, he could not point to evidence indicating that the way stimuli are sampled affects the results obtained. In the hope that some fifty years later such evidence would be available, Mandeep Dhami and ourselves compared the effects of systematic and representative designs in various lines of investigation, namely, policy-capturing research and research on overconfidence and on the hindsight bias (Dhami, Hertwig, & Hoffrage, 2004). In the following, we briefly summarize our main findings.

Do Judgment Policies Differ in Representative versus Systematic Designs?

The key goal of many studies in the tradition of *policy-capturing* research has been to pin down how people process (e.g., combine or weight) cues to judge real-world problems – for instance, the degree to which patients have a mental health problem, the amount of bail to be set on a number of cases, the quality of shopping centers, or whether it was safe for a car to cross an intersection. To this end, participants are typically required to make decisions on a set of either real or hypothetical cases each with a corresponding set of cues. A person's judgment policy is then inferred, traditionally, using a multiple linear regression analysis and is characterized, for instance, in terms of the number and weight of cues used to make judgments. In addition, achievement is frequently measured in terms of the correlation between a person's judgments and the criterion values and by comparing the person's policy with a model of the task (for overviews, see Cooksey, 1996; Stewart, 1988).

Many researchers who aim at capturing policies have adopted the Brunswikian approach to study judgment and decision making. They have

expressed a commitment to the method of representative design, which, they argue, differentiates them from other researchers in cognitive psychology, in general, and judgment and decision making, in particular (e.g., see Brehmer, 1979; Cooksey, 1996; Hastie & Hammond, 1991, p. 498). In light of their explicit commitment to the method of representative design, we were surprised to find that a large proportion of studies (those that relied on formal situational sampling; see Hammond, 1966) often failed to represent the ecological properties toward which generalizations were intended. For instance, researchers rarely combined cues to preserve their intercorrelations – an essential condition for the operation of vicarious functioning.

Possibly, the most rigorous test of the effect of representative design is a within-subjects comparison of policies captured for individuals under both a representative and an unrepresentative condition. Not surprisingly, given the frequent failure to implement representative design, we found only two published studies that aimed for such a test. The first study examined how livestock experts judge the breeding quality of pigs. In an unrepresentative condition, Phelps & Shanteau (1978) asked experts to respond to descriptions of hypothetical cases comprising a fractional factorial combination of eleven cues indicative of the breeding quality of pigs. Two months later, now partaking in a representative condition, the same experts rated the overall quality of eight pigs in photographs and provided ratings on the eleven cues. The experts' inferred policy differed across conditions: In the unrepresentative condition, experts used markedly more cues (i.e., nine to eleven) than in the representative condition (i.e., fewer than three) as suggested by the number of statistically significant cues in the captured policies. In the second rigorous test of the impact of sampling procedure on the captured policy, Moore & Holbrook (1990) studied the car-purchasing policies of MBA students and found a significant difference between the weights attached to two cues in individuals' policies captured using representative and unrepresentative stimuli.

In addition, Dhami et al. (2004) found a small set of other studies that compared policies captured under representative and unrepresentative conditions, albeit less stringently. Overall, the findings in this set are mixed (for details, see Dhami et al., 2004, Table 5). Whereas some researchers found differences in the policies captured using representative and unrepresentative cases (Ebbesen & Konecni, 1975, 1980; Ebbesen, Parker, & Konecni, 1977; Hammond & Stewart, 1974), others concluded that there are no differences (Braspenning & Sergeant, 1994; Olson, Dell'omo, & Jarley, 1992; Oppewal & Timmermans, 1999). When differences were observed, they occurred on multiple dimensions including those that Phelps & Shanteau (1978) and Moore & Holbrook (1990) identified, that is, the number and weight of cues.

Because of the small number of studies, it is difficult to draw any definite conclusions regarding the effects of representative design in

policy-capturing studies. Fortunately, however, in Dhami et al. (2004), we could turn to two other lines of research that have investigated the impact of sampling experimental stimuli, namely, research on the overconfidence effect and the hindsight bias. What caused researchers' interest in this issue?

How Rational Do People Appear in Representative and Systematic Designs? The Case of Overconfidence and Hindsight Bias

In Brunswik's understanding of psychology as a functionally oriented science, one should study the adjustment of organisms to the environments in which they actually live and the resulting level of the organisms' achievements. Although many contemporary psychologists would not necessarily share Brunswik's commitment to a functional perspective, akin to his program, they aim to describe and explain people's cognitive and behavioral achievements. However, the empirical results are typically not couched in terms of adjustment and achievement, but, for instance, in terms of people's *rationality* or lack thereof. For an illustration, take the heuristics-and-biases program (e.g., Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1996), which is, on many accounts, the most influential research program within cognitive and social psychology over the past thirty years.

Since its inception in the early 1970s, the heuristics-and-biases program has produced a large and growing collection of findings demonstrating that human reasoning frequently departs from classic norms of rationality. These findings include insensitivity to sample size, base-rate neglect, misperceptions of chance, illusory correlations, overconfidence, and hindsight bias. Such "cognitive illusions" have been explained in terms of simple heuristics that people, being cognitively limited, need to rely on when they make inferences about an uncertain world (Tversky & Kahneman, 1974). Resting their judgment on the seemingly ubiquitous cognitive illusions, many researchers have arrived at a bleak assessment of human reasoning – as nothing more than "ludicrous," "indefensible," and "self-defeating" (see Krueger & Funder, 2004).

Like Brunswik, the heuristics-and-biases program has stressed that people need to function in an inherently uncertain world. In Brunswik's words, the "environment to which the organism must adjust presents itself as semi-erratic and that therefore all functional psychology is inherently probabilistic" (Brunswik, 1955, p. 193). If so, and this forms the core of representative design, then the statistical properties of the laboratory task need to represent the statistical properties of the ecology to which the results are to be generalized. Concerns about whether this has been true in studies of cognitive illusions have given rise to a Brunswikian perspective, first in research on the overconfidence effect and then in research on the hindsight bias.

The overconfidence effect plays a prominent role among the cognitive illusions catalogued by the heuristics-and-biases program (Kahneman & Tversky, 1996). It has received considerable attention both within psychology (for a recent review, see Hoffrage, 2004) and beyond [e.g., in economics – see Hertwig & Ortmann (2004); and in consumer decision making – see Alba & Hutchinson (2000)]. Studies in psychology that demonstrate the overconfidence effect typically present respondents with questions that test their general knowledge, such as "Which city has more inhabitants: Atlanta or Baltimore?" Participants are asked to select the correct option and then indicate their confidence that it is indeed correct. The frequently replicated finding is that among choices about which people say they are 100% confident, only about 80% are correct. Similarly, among choices about which people deem themselves 90% confident, only about 75% are correct, and so on. In quantitative terms, the overconfidence effect is usually defined as the difference between the mean confidence rating and the mean percentage correct across a series of such general knowledge questions.

Are people really as out of touch with the accuracy of their knowledge as the host of overconfidence studies suggests? Adopting a Brunswikian perspective, and thus paying special attention to how overconfidence researchers sampled general knowledge questions, Gigerenzer, Hoffrage, & Kleinbölting (1991) challenged this conclusion. According to their theory of *probabilistic mental models*, people respond to questions, such as which city has more residents, by constructing a mental model that contains probabilistic cues and their validities. For instance, when choosing between Atlanta and Baltimore, a person may retrieve the fact that Atlanta has one of the world's fifty busiest airports and that Baltimore does not, and that cities with such an airport tend to be larger than those without. Capitalizing on the knowledge of this cue, the person thus concludes that Atlanta has more inhabitants than Baltimore, and then states the cue's validity as her subjective confidence that this choice is correct.

At this point, the way in which experimenters sample questions becomes crucial. To illustrate this, let us assume that only one cue, the airport cue, can be retrieved, upon which basis the relative population size of U.S. cities is inferred. Among the fifty largest U.S. cities, the airport cue has an ecological validity of .6 (Soll, 1996). The ecological validity of a cue is defined as the percentage of correct choices rendered possible by this cue. If the participants' assessment of the validity of the cue approximates its ecological validity, then they should be well calibrated, that is, they do not overestimate the accuracy of their knowledge. This means that given a confidence category (e.g., 60%), the relative frequency of correct choices should equal this value (here, 60%). This, however, only holds true under two conditions: The first is that people's subjective cue validities approximate the ecological validities – an assumption that is consistent with a rich

literature demonstrating that people seem to be keenly sensitive to environmental frequencies (e.g., Hasher & Zacks, 1984). The second condition is that the experimenter samples questions such that the cues' validities in the experimental sample of questions are preserved.

Gigerenzer et al. (1991) suggested that overconfidence stems from the fact that the second condition typically is not preserved. Specifically, they argued that researchers do not sample general knowledge questions randomly but tend to overrepresent items in which cue-based inferences would lead to wrong choices. If so, then overconfidence would not reflect fallible reasoning processes but would be an artifact of the way the experimenter sampled the stimuli and ultimately misrepresented the cue-criterion relations in the ecology. Supporting this interpretation, Gigerenzer et al. (1991, Study 1) were able to show that people were well calibrated when questions included randomly sampled items from a defined reference class (here, German cities). Percentage correct and mean confidence amounted to 71.7% and 70.8%, respectively. They were also able to replicate the overconfidence effect in a nonrandomly selected set of items – here, percentage correct and mean confidence amounted to 52.9% and 66.7%, respectively; overconfidence was 13.8%. The same pattern of findings has been independently predicted and replicated by Peter Juslin and his colleagues (e.g., Juslin, 1993, 1994; Juslin & Olsson, 1997; Juslin, Olsson, & Björkman, 1997); for further effects that are consistent with a Brunswikian perspective, such as the confidence-frequency effect, see Gigerenzer et al. (1991).

One objection that has been raised against this Brunswikian interpretation of overconfidence is that the observed differences between the selected and representative set of questions are just another manifestation of the *hard–easy effect* (e.g., Griffin & Tversky, 1992). The hard–easy effect is the observation that overconfidence covaries with item difficulty: Difficult item sets (i.e., percentage of correct answers about 75% or lower) tend to produce overconfidence, whereas easy sets (i.e., percentage correct about 75% or higher) tend to produce underconfidence. At first glance, the hard–easy effect seems to be confounded with the sampling procedure. Specifically, sets consisting of representatively drawn items tend to be easier and confidence judgments tend to be well calibrated. In contrast, selected sets are difficult (as items often have been selected to be difficult, or even misleading) and tend to yield overconfidence.

The differential impact of item difficulty and sampling procedure, however, can be teased apart empirically: In a meta-analysis, Juslin, Winman, & Olsson (2000) conducted a review of ninety-five independent data sets with selected items and thirty-five sets with representatively sampled items. Across all selected and representative item sets, overconfidence was 9% and 1%, respectively (with 95% confidence intervals for each of the two sampling procedures of ±2%). Having statistically controlled for item

difficulty, the authors pointed out that this difference could not be explained by differences in percentage correct, as has been claimed by Griffin & Tversky (1992), based on three data points. Moreover, when they controlled for end effects of the confidence scale and linear dependence between percentage correct and the over-/underconfidence score (i.e., mean confidence minus percentage correct), the hard–easy effect nearly disappeared for the representative item sets.

The impact of the item-sampling procedure is not restricted to the overconfidence effect. It also matters, for instance, for the hindsight bias – the tendency to falsely believe, after the fact, that one would have correctly predicted the outcome of an event. [For a recent collection of papers on the hindsight bias, see Hoffrage & Pohl (2003).]

Akin to research on overconfidence, Winman (1997) presented participants with general knowledge questions, such as "Which of these two countries has a higher mean life expectancy, Egypt or Bulgaria?" Before they responded to the question, participants were told the correct answer, Bulgaria, and then were asked to identify the option they would have chosen had they not been told the correct answer. Winman (1997) presented participants with selected and representative sets of questions. In the latter, the countries involved were drawn randomly from a specified reference class. The differences in hindsight bias were striking: In the selected set, 42% of the items elicited the hindsight bias, whereas in the representative set only 29% did. Moreover, Winman observed that in the representative sample only three out of twenty participants reached a higher degree of accuracy in hindsight than in foresight (thus indicating the hindsight bias). In the selected sample, in contrast, fourteen out of twenty participants fell prey to the hindsight bias (in terms of higher hindsight accuracy); for more detailed analyses and additional data regarding the impact of sampling of items on the hindsight bias, see Winman and Juslin's (2005) chapter in the present volume.

In conclusion, in Brunswik's view, psychology is the study of the adjustment of organisms to environments and the resulting degree of achievement. A prominent research program in contemporary psychology, the heuristics-and-biases program, also focuses on achievement, measured in terms of the degree to which the cognitive system is able to reason in accordance with the laws of statistics and probability theory. For Brunswik, the experimental stimuli used to measure the organisms' degree of achievement need to be selected so that the sample is representative of the actual demands that the environment makes upon the organism. Recent research on the overconfidence effect and the hindsight bias has demonstrated that people's cognitive achievement, here in terms of the veridical assessment of their knowledge, depends strongly on how the experimental stimuli are sampled. Thus, experimenters' conclusions about how much or little people achieve in their uncertain inferences, and how rational or irrational

they are, appear to depend also on how experimenters select the stimuli that constitute the reality of the laboratory.

REPRESENTATIVE DESIGN AND SIZE OF THE REFERENCE CLASS

Sampling of experimental stimuli matters. Overconfidence, or the lack thereof, for instance, appears to be a function of whether experimental stimuli are randomly drawn from a specified reference class. In the following, we examine the boundary condition of this observation. Specifically, we examine whether the overconfidence phenomenon is robust across different sizes of the reference class. According to the theory of probabilistic mental models (Gigerenzer et al., 1991), people are well adapted to their natural environments, and they are able to estimate cue validities with a reasonable degree of accuracy. But what if cue validities change as a function of different sizes of the reference class, and if so, which reference class is the one to which people are adapted?

Cue validities can depend on the size of the reference class. To appreciate this, consider the hypothetical environment of objects in Figure 16.1. In this environment, there are six objects with their cue values on one dichotomous cue: Objects A and C have a cue value of "1," and Objects B, D, E, and F have a cue value of "0". Cue values are coded such that in a comparison between two objects with values of "1" and "0," the object with the value of "1" is more likely to be larger. The validity of the cue (as computed in the complete paired comparison) is 87.5% (i.e., seven out of eight inferences are correct). How would this validity be affected if we reduced the size of

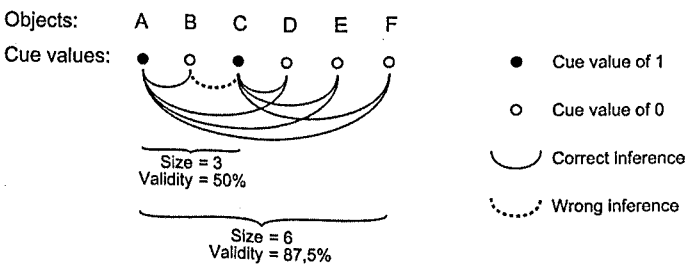

FIGURE 16.1. Six fictitious objects, ordered according to a numerical criterion, with A being the largest and F the smallest. Black circles represent a cue value of "1" (e.g., a particular feature is present), and white circles represent a cue value of "0" (e.g., a particular feature is absent). The validity of this cue depends on the size of the reference class, that is, on the number of objects larger than a specified threshold. Two objects between which the cue does not discriminate (and thus does not allow for an inference) are not connected; such a pair does not enter the computation of cue validity.

TABLE 16.1. *Validities of Twelve Cues in the Set of German Cities (Table Taken from Hoffrage, 1995)*

	Threshold			
	100,000	150,000	200,000	300,000
Size of reference class[a]	65	42	31	16
Resulting number of pairs	2080	861	465	120
Cues and their validities				
University	70	75	69	64
Intercity station	77	75	76	0
Airport	89	80	86	78
Soccer team in national league	90	82	81	65
Industrial belt	57	51	42	44
Government	84	72	76	73
License plate	92	90	84	65
Zip code	80	93	83	85
Symbol in road map	99	97	99	100
Court ("Oberlandesgericht")	80	79	86	83
National pension offices	76	71	62	87
Area (square kilometers)	82	87	85	89

[a]The size of the reference class is the number of cities with more inhabitants than the threshold.

the reference class from which objects are sampled to a subset of the largest n objects? (Whenever we use the term *size of the reference class*, here and in the following, we refer to the number of objects with a criterion value that is higher than a specific threshold.) It turns out that the validity depends on the threshold; specifically, for the reference classes of size 6, 5, 4, 3, and 2, the validities are 7/8, 5/6, 3/4, 1/2, and 1, respectively (see Figure 16.1). Thus, except for the reference class of size 2, validity decreases as the size of the reference class decreases.

Figure 16.1 illustrates a hypothetical environment, but what about real-world environments? Clearly, in many environments the issue of the size of the reference class will not matter. For instance, the reference class of all Nobel laureates, or nations that partook in the last Olympic Games, is well defined. Other reference classes, however, have fuzzy borders. Take the class of German cities. Though there is a strict bureaucratic definition of what counts as a city, people's subjective reference classes may take on very different sizes. If so, how would different sizes impact on cue validity? Gigerenzer et al. (1991) used all German cities with more than 100,000 inhabitants (as of 1988). Although 100,000 is a salient number, other thresholds might have been used. Indeed, as Table 16.1 shows, the cue validities in this environment depend on this threshold, that is, on the minimum number of inhabitants a city must have to be included in the set.

Across four different thresholds, cue validities varied widely: For one of the twelve cues, the validity dropped from 77% to 0%; for the others, the average absolute difference between the validities among all cities with more than 100,000 inhabitants and those among all cities with more than 300,000 was 10.3% (when we designed and published the studies reported in Gigerenzer et al., 1991, we were not aware of this dependency). The average Pearson and Spearman correlations between the cue validities (first computed across the twelve cues within a given pair of thresholds, and then averaged across the six possible pairs of thresholds that can be constructed from the four thresholds displayed in Table 16.1) were .66 and .62, respectively.

The observation that cue validities depend on the specific size of the reference class as determined by the experimenter can matter for the interpretation of empirical results and for the selection of experimental materials. In the following, we illustrate these points in the contexts of overconfidence research (Study 1) and of a computer simulation of different inference heuristics (Study 2).

Study 1: Over-/Underconfidence Depends on the Size of the Reference Class

The first study was designed both as a test of the recognition heuristics and to investigate the effect of different sizes of the reference class on overconfidence. The recognition heuristic is an inference heuristic that can be applied to infer, for instance, which of two cities has more inhabitants. Its policy is as follows: "If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion" (Goldstein & Gigerenzer, 2002, p. 76).

Austrian students (from the University of Salzburg, $N = 60$) were asked to decide for 100 pairs of U.S. cities which of two cities has more residents (Hoffrage, 1995, Study 4). For ten participants, these pairs were created by combining cities that were *randomly* sampled from the set of all ($n = 72$) cities with more than 200,000 inhabitants (as of 1988, henceforth the *large* set); for another ten participants, the cities were randomly sampled from the set of those ($n = 32$) cities with more than 400,000 inhabitants (henceforth the *small* set). After making their decisions and judging their confidence, participants were asked to state the following for each city:

(i) whether they had any knowledge about the city beyond mere name recognition (henceforth denoted by *K*, for knowledge);
(ii) whether they recognized the city's name, but had no more knowledge about it (*R*, for recognition); or
(iii) whether they had never heard of the city (*U*, for unknown).

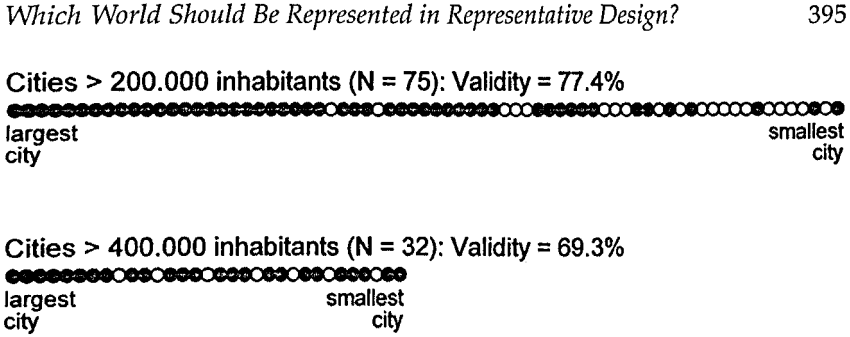

FIGURE 16.2. Austrian participants' recognition values of U.S. cities. Thirty participants provided recognition judgments about the largest seventy-five cities (top) and another thirty participants about the largest thirty-two cities (bottom). Black circles represent cities that at least half of the participants recognized (regardless of whether they also had some knowledge beyond mere name recognition); white circles represent cities that less than half of the participants recognized. The validity of the recognition cue was first computed for each participant separately and then averaged across participants.

Because the experimenter did not control which objects (of the large set and of the small set) were paired together, these two groups were tested in an approximation to a representative design. For the other participants, the experimenter systematically manipulated how the cities were paired. These forty participants first stated their recognition knowledge for each city: twenty participants for all cities from the large set and another twenty for all cities from the small set. For half of each group of twenty, the comparisons were constructed such that the recognition heuristic often discriminated; for the other half they were constructed such that the heuristic rarely discriminated.[1]

Figure 16.2 represents the knowledge of the average participant about the American cities. Specifically, the black circles represent cities that at least half of the participants recognized irrespective of whether they also had some knowledge beyond mere name recognition; the white circles represent the cities that less than half of the participants recognized. As one can see, the more inhabitants a city has, the higher the likelihood is that the name is recognized. The validity of the recognition knowledge is computed among all possible combinations of two cities where one is

[1] For each participant who was given the comparisons such that the recognition heuristic often discriminated, the types of comparisons (and their frequencies) were the following (types in which the recognition heuristic discriminated are in italics): *K–U (25)*, *R–U (30)*, K–R (30), K–K (5), R–R (5), and U–U (5). For participants who were given the comparisons such that the recognition heuristic rarely discriminated, the composition of pairs was as follows: *K–U (5)*, *R–U (20)*, K–R (20), K–K (5), R–R (30), and U–U (20). Note that in addition to the cases in which the recognition heuristic discriminated (K–U and R–U), knowledge about cue values could also lead to discrimination (potentially in K–R and K–K comparisons).

TABLE 16.2. *Mean Confidence, Percentage of Correct Inferences, and Over-/Underconfidence as a Function of the Size of the Reference Class and the Discrimination Rate of the Recognition Cue*

Discrimination Rate of Recognition Cue	Mean Confidence	Percent Correct	Over-/Under-Confidence
Large set (all cities > 200,000)			
Chance	70.7	72.3	−1.6
High	67.4	69.4	−2.0
Low	64.2	63.9	0.3
Small set (all cities > 400,000)			
Chance	73.6	61.5	12.1
High	74.9	65.5	9.4
Low	68.1	58.2	9.9

recognized and the other is not. The validity is the percentage of pairs in this set for which the recognized city has more inhabitants than the one that is not recognized (i.e., black–white combinations where the black circle is to the left). This validity was computed for each individual participant. Averaged across all participants who received the large set, the validity was 77.4%. For the small set, in contrast, this value was only 69.3%.

Did the participants' percentage of correct inferences show the same relationship? Yes. For each of the three discrimination conditions (cities randomly sampled irrespective of their recognition values, high discrimination rate, low discrimination rate), not only the validity of the recognition cue (Figure 16.2) but also the percentage of correct inferences (Table 16.2) was higher for the large set.

What about confidence and, thus, overconfidence? Were participants aware that the validity of the recognition cue (and probably also of the other cues they used) depends on the size of the reference class and did they adjust their confidences accordingly? In fact, mean confidences were different for the two sizes of the reference class. However, the difference points in the opposite direction: For each discrimination rate, confidences were higher for the small set (Table 16.2). That is, as the proportion of recognized cities increased, the validity of the recognition cue and the percentage correct decreased, whereas mean confidence increased. Taking mean confidence and percentage correct together, we can see that overconfidence disappeared for the large set, whereas there was substantial overconfidence for the small set (Table 16.2). Note that the effect of the size of the reference class on overconfidence was most pronounced in the chance set where the discrimination rate of the recognition cue had not been manipulated. However, also note that the effect could still be observed in the two conditions where the discrimination rate had been controlled for,

suggesting that the effect in the chance condition was not due to only different frequency distributions of comparison types (e.g., a higher percentage of cases in which both cities were recognized in the small set compared to the corresponding percentage for the large set). For another example of different degrees of over-/underconfidence attributed to different reference classes, see Juslin, Olsson, & Winman (1998).

A closer look at the data (not shown in Table 16.2) revealed that once the type of comparison (K–U, K–R, etc.) was controlled, the participants in the same size of reference class condition did not differ with respect to mean percentage correct, mean confidence, and overconfidence. In other words, the effect of the discrimination rate (which was either quasi-experimentally observed or experimentally manipulated) could be fully accounted for by the different frequency distributions of the types of comparisons. Thus, the results support the hypothesis that participants used the recognition heuristic: Manipulating how often recognition discriminates between two cities affects overall performance on the group level.[2]

In summary, whether people appear overconfident in a study is a function of not only the sampling procedure (randomly versus selected) but also the size of the reference class from which the experimental stimuli are randomly drawn. In the next study, we turn to a daunting problem any experimenter faces when aiming to determine which of several cognitive policies people use. In doing so, one often realizes that it is difficult to discriminate among different policies because they frequently predict the same behavior. To illustrate this problem, let us return to the recognition heuristic.

Study 2: Policy Capturing and the Size of the Reference Class

If a person has to decide which of two cities is larger, and if the only information at hand is whether or not the person recognizes one of the cities, then that person can do little better than rely on partial ignorance, choosing recognized cities over unrecognized ones. Both the aforementioned study as well as Goldstein & Gigerenzer's (2002) studies show that people appear to rely on this judgment policy – a policy that works well when recognition

[2] The data allow for an even stronger test. Remember that for each city the recognition value was elicited, either before or after the comparisons were performed. In the vast majority of cases, participants decided in favor of the city they knew more about. In particular, for comparisons of the types K–U, K–R, and R–U , the percentages of decisions that were made in favor of the city with a higher recognition value were 91.1%, 79.3%, and 79.9%, respectively. Moreover, participants obtained a markedly better performance when they chose the city suggested by the recognition heuristic: The percentages of correct choices were 83.6%, 74.9%, and 65.3%, respectively. In contrast, if they decided in favor of the city they knew less about, the performance for each comparison type was even below chance level, namely, 21.2%, 48.5%, and 39.2%, respectively.

is correlated with the criterion that needs to be inferred. To explain how an association between recognition and a criterion may develop, Goldstein & Gigerenzer (2002) proposed that there are "mediators" in the environment that reflect the criterion and, at the same time, are accessible to the decision maker. For example, an American citizen may have no direct information about the population size of a German city, say, Mannheim. However, Mannheim's population size may be reflected in how often Mannheim is mentioned in U.S. daily newspapers. Frequency of mentions, in turn, is correlated with recognition: The more frequently the name of a city appears in the newspaper, the more likely it is that a reader will encounter its name. In this sense, the newspaper can serve as a mediator between recognition and the criterion (here, population size). In line with this view, Goldstein & Gigerenzer (2002) found that the ecological correlation – that is, the correlation between how often the names of German cities (with more than 100,000 residents) are mentioned in a major U.S. newspaper, the *Chicago Tribune*, and their actual populations – was .82.

As with any heuristics, the recognition heuristic is a judgment policy of limited scope: It cannot be applied when both cities are either recognized or not recognized. Another heuristic, the *fluency heuristic*, however, is applicable both when the recognition heuristic is applicable and when both cities are recognized and the recognition heuristic is thus not applicable. Like the recognition heuristic, it relies on only one reason: the fluency with which the objects are reprocessed. Fluency has been shown to function as a cue across a range of judgments (e.g., Begg, Anas, & Farinacci, 1992; Jacoby et al., 1989). In the context of a two-alternative choice, such as the city-size task, Schooler & Hertwig (2005) defined the heuristic task as follows: *If one of two objects is more fluently reprocessed, then infer that this recognized object has the higher value with respect to the criterion.*

To study the performance of both the recognition heuristic and the fluency heuristic, Schooler & Hertwig (2005) implemented them within a well-known cognitive architecture, namely, ACT-R (e.g., Anderson & Lebierre, 1998). Here, we will not be concerned with the details of this implementation (for such details, see Schooler & Hertwig, 2005). The bottom line is that fluency is a function of a city's activation, and activation within ACT-R is a function of two factors: the objects' environmental frequencies, such as mentions in the newspaper, and recency of occurrence (e.g., when a city was mentioned in the newspaper).[3] Using the

[3] Within ACT-R, the system cannot inspect activation levels directly; that is, it cannot simply read off the activation of a record. However, Schooler & Hertwig (2005) proposed that by taking advantage of the one-to-one mapping between activation and retrieval time, the speed of retrieval could be used as a proxy for activation. Rather than assuming that the system can discriminate between minute differences in any two retrieval times, however, they allowed for limits on the system's ability to do this. Specifically, if the retrieval times of the two alternatives were within a just noticeable difference of each other, then the system guessed.

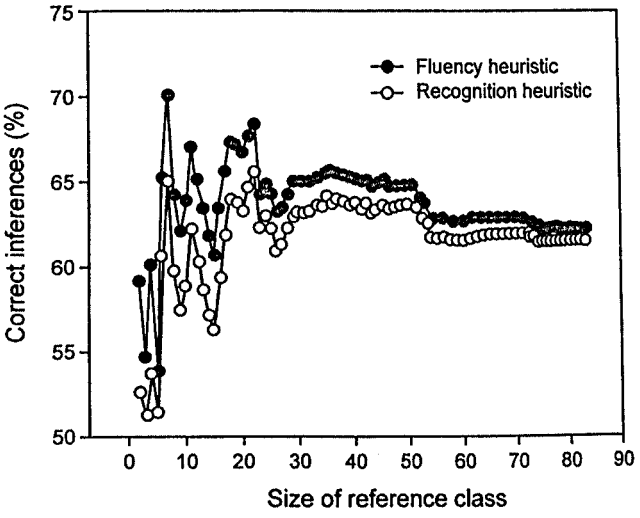

FIGURE 16.3. Percentage of correct inferences made by the recognition heuristic and the fluency heuristic as a function of the size of the reference class. Adapted from Schooler & Hertwig (2005).

environmental frequencies (i.e., mentions in the *Chicago Tribune*) and the recognition rate that Goldstein & Gigerenzer (2002) observed for the largest eighty-three German cities in a sample of students from the University of Chicago, Schooler & Hertwig (2005) calculated the cities' activations. This was carried out to ensure that performance based on the recognition rate of the model and empirical recognition was in agreement. Having thus ensured this correspondence, they could now analyze the performance of the recognition heuristic and the fluency heuristic, respectively.

Recall the policies of the two heuristics in question: For all possible pairs of cities within a reference class, the recognition heuristic chose the recognized city when only one of the cities in a given pair was recognized, and otherwise it guessed. In contrast, the fluency heuristic chose the city that was more fluently reprocessed as long as both cities were above the recognition threshold; otherwise it guessed. Figure 16.3 shows the performance of the two heuristics as a function of an increasing size of the reference class – the dimension that concerns us here. When the reference class and thus the pair comparison only include the ten largest German cities, then the fluency heuristic clearly outperforms the recognition heuristic. Whereas the fluency heuristic scores 70.1% correct inferences, the recognition heuristic scores 62%. This picture, however, changes as the reference class becomes more encompassing. The performance edge drops to a 5% difference when the twenty largest cities are included (69% versus 64.6% correct inferences), and it shrinks to a 1% advantage when all eighty-three largest German cities are included.

Why does the performance accuracy level of the heuristics converge with more encompassing reference size? The relative performance edge the fluency heuristic has over the recognition heuristic stems from the frequency of pairs in which both objects are recognized. It is only here that the two heuristics arrive at different judgments – when one object is recognized and the other is not, they behave identically, and when neither object is recognized, both heuristics need to guess. The likelihood, however, that both cities are recognized is particularly high for the very large cities (whose names appeared relatively frequently in the *Chicago Tribune*) and shrinks as the size of the reference class increases.

This finding highlights the possible performance dependence of the two heuristics on the reference class upon which they are tested. Had their performance been tested by relying only on the ten or twenty largest cities, then the fluency heuristic would have been judged to be markedly more accurate than the recognition heuristic. Indeed it is – but merely in the small set of large to very large German cities. Across a less restricted set of German cities (e.g., all cities with more than 100,000 residents), the strategies' performances are almost indistinguishable and there is no clear winner. This outcome thus demonstrates that the performance of heuristics, as well as ultimately the experimenter's conclusion, also depends on the size of the reference class.

Trying to distinguish between heuristics can be a daunting endeavor. To appreciate this, consider two experimenters who try to discover which of the two heuristics – the recognition heuristic or the fluency heuristic – people tend to use. To do so, the first experimenter uses all eighty-three German cities (with more than 100,000 residents), thus composing 3,403 pair comparisons. In this case, to find out whether a person's judgment policy is akin to the recognition heuristic or the fluency heuristic, respectively, will be extremely difficult, because in the majority of the 3,403 comparisons, both heuristics will arrive at the same predictions. The proportion of discriminatory cases will be minuscule, and the performance level of users of the recognition heuristic and users of the fluency heuristic will be hardly distinguishable. The second experimenter, in contrast, uses only the ten largest German cities, thus composing only 45 comparisons. Among them, the proportion (though not the absolute number) of discriminatory cases will be much larger, and the level of accuracy reached by users of the recognition heuristic and the fluency heuristic, respectively, will be markedly different.

This thought experiment suggests that the sampling procedure from the reference class is not the only thing that matters for the results obtained. The size of the reference class may also have drastic implications for, for instance, the observed accuracy level of heuristics and the likelihood with which an experimenter is able to distinguish among users of different heuristics.

DISCUSSION

We have discussed several conceptual and methodological implications of different sampling procedures in psychological experiments. In the first section, we briefly reviewed the historical and methodological foundations of representative design. In the second section, we introduced three lines of research – on policy capturing, overconfidence, and hindsight bias – in which empirical evidence has been accumulated indicating that the sampling procedure matters for the results obtained. Specifically, it matters whether the experimental stimuli are randomly sampled from a defined reference class, or whether the experimenter samples the objects systematically, thus causing the frequency distribution and informational structures in the sample and the population to diverge. After having shown that the *how* of sampling experimental stimuli matters, in the third section we turned to an issue that has rarely been addressed in the literature on representative design: Apart from the sampling procedure, the size of the reference class from which stimuli are sampled also matters for the results obtained. Clearly, both parties in an experiment – the participants and experimenters – cannot help but to settle on a reference class. The question is whether they will settle on the same one. We conclude with a discussion of the issues raised.

Selection of the Reference Class: A Time-Honored and Ubiquitous Problem

The problem of choosing the adequate reference class is neither trivial nor is it new. It is, for instance, fundamental to the frequentistic interpretation of probabilities (for the historical routes and interpretations of probabilities, see Gigerenzer et al., 1989). Conceptualizing probabilities in terms of relative frequencies requires specifying the reference class within which we count the objects or the events in question. Logic demands that the specification of the reference class precedes the counting of designated objects within it. Or, in the words of the great probability theorist Richard von Mises (1957), "We shall not speak of probability until a collective has been defined" (p. 18).

Take, for illustration, the problem of calculating the probability of a person dying within, say, the next ten years. Insurance companies need to estimate such probabilities to calculate a person's premium. But which of the person's innumerable properties – age, sex, profession, health, income, eating habits, family life, to name only a few – should be used to construct a reference population? Each of these and many other properties (as well as combinations thereof) could be used to define the reference class, and in all likelihood, many of the resulting reference classes would yield different statistics and thus different estimations for mortality risks, leaving open

the question of which is the correct one (von Mises, 1957; for an example of different types of reference classes for probabilities, see Gigerenzer et al., 2005). To the best of our knowledge, neither probability theorists nor insurance companies have yet been able to provide a solution to this problem.

Selection of the Reference Class in Psychological Theory and Experimental Practice

The theory of probabilistic mental models predicts that representative sampling of experimental stimuli from a specified reference class results in well-calibrated confidence judgments (Gigerenzer et al., 1991). In this chapter, we have shown that this prediction may overlook an important additional condition for good calibration. Specifically, we have illustrated that cue validities can vary markedly as a function of the inclusiveness of the reference class. To the extent that confidence judgments rest on cue validities, they are only well calibrated when participants are sensitive to the inclusiveness of the reference class. Arguably, this is not a very plausible expectation. Thus, for overconfidence to be eliminated in psychological experiments, two conditions need to be met: (1) representative sampling and (2) representative sampling from *the* reference class from which people's knowledge of cues and cue validities stem. Therein, of course, lurks the problem: Which reference class is chosen, and what is its size? Moreover, how can the experimenter know the "right" reference class and its size?

Frankly, we do not have answers to these questions. Yet, let us suggest some directions in which one may search for answers. First and foremost, let us stress that we do not believe that the problem of the reference class is unique to the theory of probabilistic mental models. Rather, the problem of the "right" reference class permeates all of psychology. In principle, any psychological theory has to provide an answer as to which reference class of objects, stimuli, and situations it is meant to apply. For instance, any theory of classification and object perception ought to be explicit about the reference class of objects and properties in the world and in people's minds to which it applies. In reality, to the best of our knowledge, hardly any psychological theory tackles the reference class issue. Perhaps, the reason is that there seems to be no good answer – for purely logical reasons. For each reference class in relation to which a sampling process is representative, there exists a superordinate or a subordinate reference class relative to which the outcome of the sampling process is simply biased. For instance, sampling representatively from the 83 largest German cities results in a sample that is unrepresentative with respect to the 500 largest German cities or to the 10 largest German cities.

Does this mean that representative design is a chimera – a creature that exists in the fantasy of some experimenters, but for all practical purposes is nonviable? We do not think so. We believe that there are possible pragmatic routes toward a "good enough solution." Under some circumstances, experimenters may circumvent the problems that result from fuzzy reference classes either by selecting one that is small, finite, and complete (e.g., all African states) or by creating microworlds (e.g., Fiedler et al., 2002) over which experimenters have full control. They thus can determine participants' exposure to these worlds and make sure that the intended reference class and the participants' reference class converge.

Another route to determining the size of the reference class is to explore its boundaries empirically, for instance, by analyzing environmental frequencies. Anderson & Schooler (1991) examined a number of environmental sources (e.g., the *New York Times* and electronic mail) to show that the probability that a memory for a particular piece of information will be needed shows reliable relationships to frequency, recency, and patterns of prior exposure. Such an analysis of environmental statistics could also be conducted in the context of overconfidence research. For the German city task, for instance, it may show that people are much more likely to encounter larger cities, such as *Frankfurt* and *Stuttgart* (ranked 5 and 8), than smaller cities, such as *Leinefeld* and *Zeulenroda* (ranked 345 and 403). To the extent that environmental frequencies are also indicative of how often an object is the subject of an inference, we suggest that decision makers rarely need to make comparisons among items drawn from the lower tail of a criterion distribution.

In the context of social cognition, an empirical approach to determining the "right" size of *social* reference classes may capitalize on empirical studies indicating that the size of people's "sympathy" group is in the order of 10–15 people (i.e., number of friends and relatives whom they contact at least once a month), that the typical size of the network of acquaintances is around 135–150, and that the number of people whose faces one can attach names to is around 1,500–2,000 (see Dunbar, 1996; Hill & Dunbar, 2003). On the basis of these and similar quantities, researchers investigating social cognition infer plausible estimates of the size of people's social reference classes.

Yet another way to determine the "right" size of people's reference classes is to transfer the task of sampling experimental stimuli from the experimenter to the participants. In his aforementioned investigation of size constancy, Brunswik (1944) took exactly this approach (see also Hogarth, 2005, this volume). Specifically, he randomly interrupted the perceiver in her flow of daily activities at randomly chosen intervals, thereby letting the perceiver, the environment, and chance determine which environmental stimuli are designated to become experimental ones.

EPILOGUE

Participants in psychological experiments typically respond to stimuli selected by the experimenter. Each of these stimuli in and of itself constitutes reality. However, is this slice of reality in the experiment a sample of the reality to which generalization is intended? According to Brunwik, the answer is often "no." Owing to psychologists' high esteem of internal validity, experimental studies, according to Brunswik, often end up investigating phenomena at the fringe of reality. To remedy this situation, he advocated an entirely different approach to the sampling of stimuli: an approach in which experimenters apply the same standards to the sampling of stimuli that they espouse when they sample participants. We believe Brunswik's criticism of experimental practices and his proposal for reform are as timely today as they were back then. Although Brunswik's representative design is by no means convenient, nor without its own problems, it provides us with a valuable if preliminary framework to conceptualize stimulus sampling in experiments. Clearly, as experimenters we cannot help but sample stimuli, and it is up to us to strive for even better ways of doing so.

References

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research, 27*, 123–156.

Anderson, J. R., & Lebierre, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396–408.

Baker, T. L. (1988). *Doing social research*. New York: McGraw-Hill.

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121*, 446–458.

Braspenning, J., & Sergeant, J. (1994). General practitioners' decision making for mental health problems: Outcomes and ecological validity. *Journal Clinical Epidemiology, 47*, 1365–1372.

Brehmer, B. (1979). Preliminaries to a psychology of inference. *Scandinavian Journal of Psychology, 20*, 193–210.

Brunswik, E. (1940). Thing constancy as measured by correlation coefficients. *Psychological Review, 47*, 69–78.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50*, 255–272.

Brunswik, E. (1944). Distal focussing of perception: Size constancy in a representative sample of situations. *Psychological Monographs, 56*, 1–49.

Brunswik, E. (1945). Social perception of traits from photographs. *Psychological Bulletin, 42*, 535.

Brunswik, E. (1952). The conceptual framework of psychology. In *International Encyclopedia of Unified Science* (Vol. 1, No. 10, pp. iv–102). Chicago: University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193–217.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

Brunswik, E., & Reiter, L. (1937). Eindrucks-Charaktere schematisierter Gesichter. *Zeitschrift für Psychologie, 142*, 67–134.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Diego: Academic Press.

Czikszentmihalyi, M., & Larson, R. (1992). Validity and reliability of the experience sampling method. In M. W. de Vries (Ed.), *The experience of psychopathology: Investigating mental disorders in their natural settings* (pp. 43–57). New York: Cambridge University Press.

Dhami, M., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959–988.

Dukes, W. F. (1951). Ecological representativeness in studying perceptual size-constancy in childhood. *American Journal of Psychology, 64*, 87–93.

Dunbar, R. (1996). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard University Press.

Ebbesen, E. B., & Konecni, V. J. (1975). Decision making and information integration in the courts: The setting of bail. *Journal of Personality and Social Psychology, 32*, 805–821.

Ebbesen, E. B., & Konecni, V. J. (1980). On the external validity of decision-making research: What do we know about decisions in the real world? In T. S. Wallsten (Ed.), *Cognitive processes in choice and decision behavior* (pp. 21–45). Hillsdale, NJ: Lawrence Erlbaum.

Ebbesen, E. B., Parker, S., & Konecni, V. J. (1977). Laboratory and field analyses of decisions involving risk. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 576–589.

Feigl, H. (1955). Functionalism, psychological theory, and the uniting sciences: Some discussion remarks. *Psychological Review, 62*, 232–235.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom – A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88*, 527–561.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652–670.

Gigerenzer, G., Hertwig, R., van der Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). "A 30% chance of rain tomorrow:" How does the public understand probabilistic weather forecasts? *Risk Analysis, 25*, 623–629.

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

Gigerenzer, G., Switjink, Z., Porter, T., Daston, L., Beatty J., & Krüger, L. (1989). *The empire of chance*. Cambridge, UK: Cambridge University Press.

Gillis, J., & Schneider, C. (1966). The historical preconditions of representative design. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 204–236). New York: Holt, Rinehart and Winston.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 75–90.

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

Hammond, K. R. (1966). Probabilistic functionalism: Egon Brunswik's integration of the history, theory, and method of psychology. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 15–80). New York: Holt, Rinehart and Winston.

Hammond, K. R. (1996). Upon reflection. *Thinking & Reasoning, 2*, 239–248.

Hammond, K. R. (1998). Representative design. http://www.brunswik.org/notes/essay3.html.

Hammond, K. R., & Stewart, T. R. (1974). *The interaction between design and discovery in the study of human judgment.* Report No. 152. University of Colorado Institute of Behavioral Science.

Hammond, K. R, & Stewart, T. R. (Eds.) (2001). *The essential Brunswik: Beginnings, explications, applications.* Oxford: Oxford University Press.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*, 1372–1388.

Hastie, R., & Hammond, K. R. (1991). Rational analysis and the lens model. *Behavioral and Brain Sciences, 14*, 498.

Hilgard, E. R. (1955). Discussion of probabilistic functionalism. *Psychological Review, 62*, 226–228.

Hill, R. A., & Dunbar, R. I. M. (2003). Social network size in humans. *Human Nature, 14*, 53–72.

Hertwig, R., & Ortmann, A. (2004). The cognitive illusion controversy: A methodological debate in disguise that matters to economists. In R. Zwick & A. Rapoport (Eds.), *Experimental business research* (pp. 361–378). Boston: Kluwer.

Hoffrage, U. (1995). *Zur Angemessenheit subjektiver Sicherheits-Urteile. Eine Exploration der Theorie der probabilistischen mentalen Modelle [The adequacy of subjective confidence judgments: Studies concerning the theory of probabilistic mental models].* Doctoral thesis, University of Salzburg.

Hoffrage, U. (2004). Overconfidence. In R. F. Pohl (Ed.), *Cognitive illusions: A Handbook on fallacies and biases in thinking, judgement and memory* (pp. 235–254). Hove, UK: Psychology Press.

Hoffrage, U., & Pohl, R. F. (Eds.) (2003). Hindsight Bias (Special Issue). *Memory, 11*, 329–504.

Hogarth, R. (2006). Is confidence in decisions related to feedback? Evidence from random samples of real-world behavior. This volume.

Hull, C. L. (1943). The problem of intervening variables in molar behavior theory. *Psychological Review, 50*, 273–291.

Jacoby, L. L., Kelley, C. M., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56*, 326–338.

Juslin, P. (1993). An explanation of the "hard-easy effect" in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology, 5*, 55–71.

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

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344–366.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making, 10,* 189–209.

Juslin, P., Olsson, H., & Winman, A. (1998). The calibration issue: Theoretical comments on Suantak, Bolger, and Ferrell (1996). *Organizational Behavior & Human Decision Processes, 73,* 3–26.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107,* 384–396.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases* (pp. 493–508). Cambridge, UK: Cambridge University Press.

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

Krech, D. (1955). Discussion: Theory and reductionism. *Psychological Review, 62,* 229–231.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27,* 313–327.

Kurz, E. M., & Hertwig, R. (2001). To know an experimenter. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 180–186). New York: Oxford University Press.

Kurz, E. M., & Tweney, R. D. (1997). The heretical psychology of Egon Brunswik. In W. G. Bringmann, H. E. Lueck, R. Miller, & C. E. Early (Eds.), *A pictorial history of psychology* (pp. 221–232). Carol Stream, IL: Quintessence.

Moore, W. L., & Holbrook, M. B. (1990). Conjoint analysis on objects with environmentally correlated attributes: The questionable importance of representative design. *Journal of Consumer Research, 16,* 490–497.

Olson, G. A., Dell'omo, G. G., & Jarley, P. (1992). A comparison of interest arbitrator decision-making in experimental and field settings. *Industrial and Labor Relations Review, 45,* 711–723.

Oppewal, H., & Timmermans, H. (1999). Modeling consumer perception of public space in shopping centers. *Environment and Behavior, 31,* 45–65.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance, 21,* 209–219.

Postman, L. (1955). The probability approach and nomothetic theory. *Psychological Review, 62,* 218–225.

Schooler, L., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112,* 610–628.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65,* 117–137.

Stewart, T. (1988). Judgment analysis: Procedures. In B. Brehmer & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 41–74). North-Holland: Elsevier.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

von Mises, R. (1957). *Probability, statistics, and truth.* New York: Dover.
Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115–1125.
Winman, A. (1997). The importance of item selection in "knew-it-all-long" studies of general knowledge. *Scandinavian Journal of Psychology, 38,* 63–72.
Winman, A., & Juslin, P. (2006). "I'm m/n confident that I'm correct": Confidence in foresight and hindsight as a sampling probability. This volume.
Woodworth, R. (1938). *Experimental psychology.* New York: Holt.