Journal of Experimental Psychology:
Learning, Memory, and Cognition
2008, Vol. 34, No. 1, 186–203

0278-7393/08/$12.00 DOI: 10.1037/0278-7393.34.1.186

The Ultimate Sampling Dilemma in Experience-Based Decision Making

Klaus Fiedler
University of Heidelberg

Computer simulations and 2 experiments demonstrate the ultimate sampling dilemma, which constitutes a serious obstacle to inductive inferences in a probabilistic world. Participants were asked to take the role of a manager who is to make purchasing decisions based on positive versus negative feedback about 3 providers in 2 different product domains. When information sampling (from a computerized database) was over, they had to make inferences about actual differences in the database from which the sample was drawn (e.g., about the actual superiority of different providers, or about the most likely origins of negatively valenced products). The ultimate sampling dilemma consists in a forced choice between 2 search strategies that both have their advantages and their drawbacks: natural sampling and deliberate sampling of information relevant to the inference task. Both strategies leave the sample unbiased for specific inferences but create errors or biases for other inferences.

Keywords: sampling, decision making, sampling bias, natural sampling, information search

Many everyday decision problems rely on direct environmental-learning experience. Teachers' grading decisions are informed by observations of students' performance in different disciplines. Personnel selection relies on applicants' reactions to various tasks and interview topics. Or, consumer choices reflect the information acquired about brands or providers in different product domains. There appears to be a simple and straightforward way of optimizing such experience-based decisions: If only the learning process relies on a sufficiently large sample of observations, it must be possible to discern the optimal decision through optimal data selection (Oaksford & Chater, 2003).

The learning task seems to have a clear-cut structure. For a consumer to make an optimal choice between alternative providers, it is only necessary to compare the quality feedback that is available for different providers in specific product domains. Granting that the feedback is reliable and accurately reflects the contingency between providers and product quality, figuring out the best provider, with the highest rate of positive evaluations, should be straightforward. The consumer's task should be easy to solve if only the differences between providers are significant enough and sufficient observations are available.

The aim of the present investigation is to contest this seemingly plausible sketch of simple experience-based decision making. In fact, finding a generally correct solution to such clearly structured problems is fraught with huge difficulties. It is virtually impossible, because every sample of observations about mundane decision problems entails the potential to be misleading under certain conditions. I refer to "the ultimate sampling dilemma" to highlight the fact that any reasonable sampling strategy that serves to optimize one decision produces a sampling bias with regard to other decisions informed by the same data.

Illustration of the Ultimate Sampling Dilemma

That judgments and decisions depend crucially on the samples of available information is not new. Sampling error and sampling bias have long been recognized as prominent topics in the methodology of the social sciences (Macrae, 1971), in epidemiology (Schlesselman, 1982), in econometrics (Manski, 1995), and in decision research in particular (Dawes, 1993; Denrell, 2005; Fiedler, 2000; Fiedler & Juslin, 2006; Juslin, Winman, & Hansson, 2007; Oaksford & Moussakowski, 2004; Stuart, Chater, & Brown, 2006). The so-called sampling approach has inspired a growing number of experiments demonstrating that although judgments and decisions are often remarkably sensitive to the data given in a stimulus sample, they may nevertheless be severely flawed due to biases inherent in the stimulus samples provided by the environment. The present research builds on the sampling approach. However, it goes beyond previous studies by showing that sampling errors and biases are even more fundamental than suggested before. They result not only from unreliable or inaccessible environments that obscure the real world or prevent decision makers from assessing relevant data, respectively. The ultimate sampling dilemma also emerges even in friendly environments that render all information available and do not constrain information search.

The dilemma can be illustrated with the following concrete sample task: Take the perspective of a manager, or entrepreneur, who, supposedly representing an "expert consumer," as it were, is motivated to be accurate and has access to all relevant data. The manager's task is to purchase electronic equipment of two kinds: computer technology and telephone devices. There are three providers offering hardware in both product domains. Let us assume all previous customers' positive or negative experience with all computers (C) and telephones (T) from all three providers—P_1, P_2,

The present research was supported by the Leibniz-Preis 2000 awarded by the Deutsche Forschungsgemeinschaft. Thanks to Nick Chater, Ralph Hertwig, Ulrich Hoffrage, Tobias Vogel, Peter Freytag, and Yaakov Kareev for their helpful and constructive comments on a draft of this article.

Correspondence concerning this article should be addressed to Klaus Fiedler, Psychologisches Institut, Universität Heidelberg, Hauptstrasse 47-51, 69117, Heidelberg, Germany. E-mail: kf@psychologie.uni-heidelberg.de

and P_3—are available in a large database, from which the manager can draw a sample of any size. On each trial, the information search can be, but need not be, constrained in any dimension. Thus, the decision maker may ask for an observation about a particular provider, P_1, or leave the selection of the next observation of any provider up to a random process. Or he/she may ask for an observation from product domain C and leave providers and evaluative outcome open. Or he/she may ask for an example of a negative observation about provider P_1, or about a positive aspect of provider P_3 observed in domain T. Or, last but not least, he/she may leave all three aspects unspecified and just ask for the next random draw from the entire database.

Let us assume that the true environmental distribution in the database of positive and negative entries referring to the three providers in the two product domains is the one given in the upper chart of Figure 1 (Ecology A)—and let us call it the "skewed world." This database contains twice as many positive as negative entries for all three providers and for both product domains. With regard to domains, the ratio of C to T entries is also 2:1, and with regard to providers, the ratio of entries on P_1, P_2, and P_3, respectively, is 4:2:1. Although the distribution is skewed in all three dimensions, all contingencies are zero; the 2:1 ratio of positive to negative entries remains constant across all providers and product domains. Would a manager who can gather as much data as desired, using any strategy, figure out these true parameters of the environment? Or would the manager come up with a biased picture of the world?

Natural Sampling

The answer depends crucially on whether the decision maker restricts the information search consistently to a strategy that has

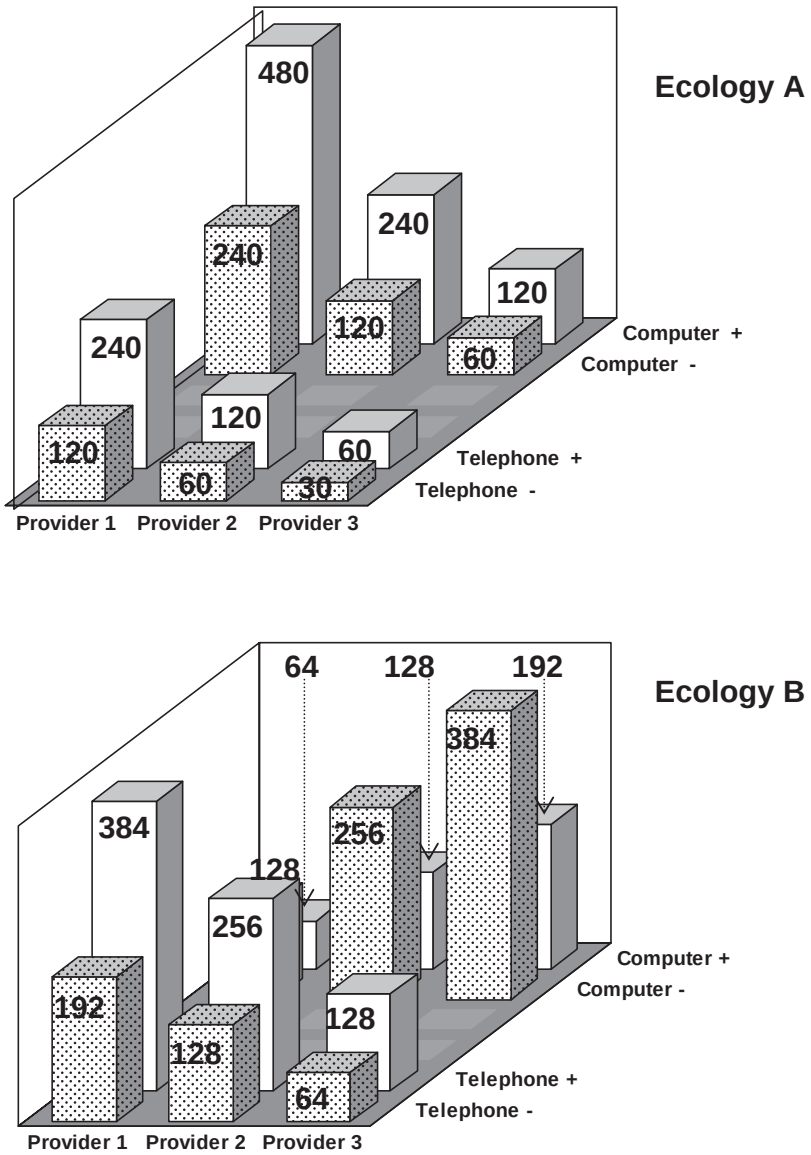

Figure 1. Two distinct sampling environments: the skewed ecology (A) and the spurious ecology (B).

been called *natural sampling* (Gigerenzer & Hoffrage, 1995). In natural sampling, one draws random events from the entire database, without ever restricting the base rates of providers, product domains, or evaluative outcomes. In other words, natural sampling means to refrain from all directed information search. If decision makers apply natural sampling all the time, the expected three-dimensional distribution in the sample will conserve the properties of the universe (as in Figure 1). However, the price for such representative, unbiased, natural sampling is that the information search cannot be tailored to the task at hand. When the base rate of observations about the provider of main interest, or the product domain of main interest, is very low, focusing on that rare provider and product domain is not allowed, nor is it possible to concentrate selectively on positive (+) and negative (–) events if the problem context calls for a focus on assets or deficits, respectively. Whenever the strategy deviates from natural sampling and concentrates on specific aspects more than others, to pursue a specific hypothesis or task goal, the resulting sample will not conserve the properties of the universe. In this case, when constrained samples are tailored for a specific purpose, the resulting information can only be trusted for the restrictive purpose for which it was tailored (e.g., a specific provider in a specific domain). As will become apparent soon, using a sample for a purpose for which it is not tailored can lead to seriously distorted and inaccurate decisions.

For instance, if a decision maker has deliberately gathered information about one particular provider, the resulting sample is conditionalized on providers. It may thus be used to infer p(+ / provider), the proportion of positive experiences given that provider. However, when the same sample is used to infer p(provider / +), that is, what proportion of positive outcomes is due to the focal provider, it may be extremely misleading.

Selective Sampling

To anticipate an important result, hardly any decision maker relies consistently on natural sampling, observing passively and giving away the chance to focus actively on specific providers, product domains, and outcomes. This divergence from natural sampling inevitably produces sampling biases. For example, consider what happens when the consumer faces the problem of diagnosing the origin of deficient products but deficient products are very rare. Very likely, the information search will focus on rare negative outcomes. As a result of selective sampling, the valence base rate will be turned from 2/3 positive (and 1/3 negative) entries in the universe into, say, 1/3 positive (and 2/3 negative) entries in the sample. Such oversampling of rare events will not distort the kind of diagnostic judgments that were the purpose of the selective search. That is, judgments of the origins of negative outcomes—such as judgments of p(provider P_1 / –) or p(domain C / –), or likelihood ratios such as p(P_1 / –) / p(P_2 / –)—will be unbiased. However, as vividly explained by Dawes (1993), as a consequence of valence-bound sampling, all sample-based judgments that use valence as the dependent variable—such as estimates of p(+ / P_1) or p(+ / P_1 in domain C) or p(–)—will be biased, for the sample proportion of positive and negative outcomes is biased toward the decision maker's search focus. As a general rule, to the extent that any variable constrains the search process, subsequent estimates of that variable are potentially biased. To that extent, sample estimates reflect the decision maker's search strategy rather than the true environmental parameter. The statistical implications of this conditional-sampling rule are well known and commonly recognized as a source of logical error (Dawes, 1993; Winship & Mare, 1992).

Why, then, do people not refrain from using a conditional information search? Why do they not exploit the advantage of natural sampling? An apparent answer is because the disadvantages of natural sampling may outweigh its advantages. First of all, natural sampling can be very expensive. If one is interested in a rare cell of a design—such as deficits (–) of provider P_3 in domain T in Figure 1—then natural sampling would require one to collect a huge number of observations, mostly about irrelevant cells, until a sufficient number of observations for the focal cell are found. This pragmatic problem may be further exacerbated for even more complex tasks. For instance, a real-world design might involve the Cartesian Product of Valence (maybe more than two levels) × Providers (maybe more than three), × Product Domains (maybe more than two) × Recency of Information (old vs. recent feedback) × Prize Level (high, medium, low) × Validity of Information Source, etc. Waiting for sufficient observations about the rarest cells in such a multicell design might easily become a never-ending assessment task.

Thus, if a decision problem calls for sufficient data about a rare event (e.g., analyzing deficits of P_3 in product domain T), a selective sampling strategy may be necessary, such as positive testing (Klayman & Ha, 1987; Oaksford & Chater, 1994), which means to actively search for those events that are in the focus of the task or hypothesis, however rare they are in the universe.

Unequal sample size, or the inaccessibility of evidence from rare cells, is but one problem of natural sampling. Another problem lies in the fact that human (and animal) learning depends on a sufficiently large sample of learning trials. Even when observation time is unrestricted and inexpensive, rare events are hard to learn and memorize, due to inhibition from more frequent neighboring events. There is ample evidence that when the same trend is observed in two categories (e.g., the same high positivity ratio for P_1 and P_3) but the number of observations is different (i.e., P_1 and P_3 differing by the ratio 4:1), then the positivity trend will be more readily discerned for the larger category P_1 than for P_3, due to less inhibition and more extensive learning experience for the former (see Fiedler, 1996, 2000; Fiedler & Walther, 2003). Thus, even when a reasonable number of items about a rare event have been observed after a long period of natural sampling, they may finally be lost through forgetting, or may be overshadowed by more frequent stimuli.

At a more fundamental level, the mere possibility of a truly natural, unconditional sampling may be questioned. In reality, unlike in the ideal world of statistics textbooks, information search is inevitably conditional on the decision maker's position in time and space, and his or her psychological distance from the decision target (Fiedler, 2007a). A consumer will hardly be able to sample information about all products, providers, markets, and product attributes nonselectively. Rather, some products or markets will be closer and others will be more remote; advertising structures render products and brands differentially available; positive evaluations of products are clearly more available in the advertising ecology than are negative evaluations; and consumers' sampling is mostly confined to the present time and to one's own country or regional market as opposed to the past and remote places. As all

information is not equally accessible, unconstrained sampling is principally impossible.

What makes the situation even worse is that consumers normally cannot perceive or know the sampling constraints imposed on newspapers, TV advertising, or the Internet. Thus, when encountering positivity rates of providers or product domains in the media, or when assessing the relative proportion of deficient products associated with providers, consumers do not know if and to what extent the media have been sampling naturally, and by what factor they have oversampled or undersampled specific aspects. Therefore, even when given a free choice, real decision makers can hardly ever realize the ideal of natural sampling.

Conversely, one may ask why decision makers do not abandon natural sampling and rely on selective samples tailored to the decision problem at hand. For instance, if the problem context calls for consumer judgments of the proportion p(+ / P_1 vs. P_2, P_3) of positive evaluations of provider P_1 in comparison to other providers (P_2 and P_3), then one is on safe ground when one samples an equal number of observations about all three providers in order to compare their positivity proportions. These proportions will be unbiased regardless of the provider base rates in the population. To repeat, sample-based judgments are unbiased as long as the sampling process is not contingent on the variable being judged (i.e., proportion of positive valence). Biases and distortions will only result when judging a variable on which sampling was contingent (e.g., when attributing positive or negative product evaluations to one of several providers—P_1 vs. P_2, P_3—based on a sample that greatly overrepresents P_1). Having understood this simple rule, the consumer might avoid all biases by tailoring the sample to the judgment problem at hand, and never misusing samples drawn for one purpose for another purpose.

Simple and straightforward as this solution may appear, it is hardly feasible for several reasons. First, as already noted, sampling constraints are typically unknown and extremely hard to diagnose. To repeat, consumers do not know whether published product information is conditional on providers, desirability, domains, the constraints of the advertising industry, or any mixture thereof. Second, even if the origin of a sample were known exceptionally (e.g., to an advertising expert who has all background knowledge of a product's marketing), the constraints may be completely different for other samples (e.g., for other products).

And third, the maxim to utilize for every judgment only those samples that were drawn with the independent variable in mind, and to ignore samples that were drawn with the dependent variable in mind, has untenable memory implications. The consumer would have to hold simultaneously many different representations of the same relation between providers, product domains, and valence—one for each sampling strategy that has been used (or imposed externally) for the collection of data. Some introspection and logical reflection tells us that our knowledge is presumably not organized by sampling strategies. It would be impossible to administrate such a multiply split memory system. For each *k*-tuple of variables (e.g., the triple of valence, providers, and domains), knowledge would have to be separately stored for each sampling strategy that is conceivable. Therefore, tailoring the sampling process to the specific judgment purpose is only possible in exceptional cases, in which decision makers have unlimited source memory capacities.

Impact of Particular Sampling Schemes

Thus, the ultimate sampling dilemma is like sailing between Scylla and Charybdis. Natural sampling is potentially unbiased but expensive, insensitive to rare events, and in reality often not feasible. Selective sampling is almost inevitable, especially when focusing on rare events or pursuing specific hypotheses, but the resulting sampling biases will likely carry over to judgments of all variables that somehow contributed to the information search. Facing this dilemma, one has to admit that real-world decisions are likely to rely on information that entails sampling biases. Let us now elaborate on the consequences of these sampling biases for the decision process. So what will a manager do when facing the task depicted above? Granting that he/she will not refrain from conducting an active information search, what alternative search strategies could be used?

Output-Bound Sampling

One typical strategy is to make an information search contingent on certain outcomes. An individual motivated by the goal to avoid regret and not to make mistakes could mainly examine negative outcomes, biasing the information sample toward negative outcomes. All sample-based estimates of valence (either unconditional or conditional on specific levels of the other variables) will then tend to be too negative. If there are indeed differences between providers, sampling of an equal proportion (or any other constant ratio) of positive and negative events will obscure these differences. Lacking a priori knowledge of the true outcome proportion, the decision maker never knows what proportion to sample.

One might correct for the bias, in principle, if one has metacognitive insight into the sampling bias. However, as will soon be apparent, such metacognitive monitoring and control of sampling bias will hardly be successful. Decision makers will normally take the sample evidence at face value and base their judgments and decisions directly on the corresponding sample statistics (Juslin et al., 2007; Kareev, Arnon, & Horwitz-Zeliger, 2002). If the sample proportion of positive outcomes is downward biased due to selective attention to negative outcomes—say, if the sample proportion $p^*(+ / P_1 \& C) = 1/3 = .33$, whereas the invisible proportion in the universe is $p(+ / P_1 \& C) = 2/3 = .67$—then estimates will follow the visible sample proportion (Fiedler, Brinkmann, Betsch, & Wild, 2000; Juslin et al., 2007; Kareev & Fiedler, 2006), with little attempt to correct for bias.

In particular, an output-bound search by valence causes bias in judgments that use valence as a dependent variable, such as judgments of p(+ / P_1 & C). Backward judgments using valence as an independent variable—such as estimates of p(P_1 & C / –) or p(P_1 & C / +)—can be unbiased, although they strongly depend on whether samples are conditional on positive or negative valence. For instance, when judging, in the context of a liability affair, the likelihood with which different providers are responsible for deficits, the most prevalent provider P_1 and the most prevalent domain C will bear the strongest association with negative events. Had the task focused on diagnosing origins of positive outcomes, P_1 and C would have been most strongly associated with positive information.

Input-Bound Sampling

As the valence of outcomes can be considered the logical dependent variable of the problem, an "experimental sampling strategy" consists in assessing valence as a function of providers and domains, sampling an equal number of observations from all 3×2 cells of the design. Although an experimental design is commonly considered optimal, it is not without sampling constraints (Brunswik, 1955; Dhami, Hertwig, & Hoffrage, 2004; Hoffrage & Hertwig, 2006). To illustrate, when an information search leaves valence open and constrains providers and domains to be orthogonal—a typical experimental strategy—the resulting estimates of valence (conditional or unconditional) are indeed unbiased. However, when the need arises to estimate the likelihood that a certain provider or domain caused a negative outcome, then all differences between providers and domains are leveled off through the orthogonal design.

There are other insights prevented by experimental sampling. Consider, for example, the environment in the bottom chart of Figure 1, which represents a case of Simpson's paradox. When pooling across product domains, one finds that the positivity rate is higher for P_1 than P_2 than P_3. However, when product domains are taken into account, it turns out that the positivity rate is markedly higher in domain T than C and that the seeming advantage of P_1 is merely due to P_1's mostly providing products from the superior domain, T. As the mediating impact of domains is partialed out by comparing providers separately within both domains, P_1 is no longer dominant. The proportion of positive to negative outcomes is constant across providers within both domains (i.e., 2:1 = 384:192 = 256:128 = 128:64 for T vs. 1:2 for C) Such spurious correlations, or mediational effects (Baron & Kenney, 1986), go undetected if the correlation between providers and domains is eliminated in an orthogonal design (Fiedler, 2000).

Mixed Strategies

In reality, an information search is characterized by mixed strategies, anticipating the need to estimate different aspects of the same decision problems. Decision makers sometimes exhibit natural sampling, sometimes fix only the provider, only the product domain, or only the evaluative outcome, and on still other trials they consider specific cells of the design. Seemingly, such a mixture might result in a balanced representation of the decision problem from all vantage points. In fact, however, the resulting sample is so complexly contaminated with bias that it is practically impossible to reconstruct the original reality. This is especially so when sampling is not under the decision maker's own control but imposed by an information ecology that does not reveal its implicit constraints.

The ultimate sampling dilemma can thus be summarized as follows. When all information pertaining to a decision problem is freely available and the decision maker is motivated to solve the problem rationally, he/she faces a dilemma between two sampling schemes: either to refrain from all active information search and rely on natural sampling or to engage in deliberate information search, enjoying its advantages but obscuring the original environmental distribution. The former strategy will conserve the true data structure, but the costs and the time required to collect any, let alone reliable, information about rare events can be immense, and memory will be biased toward more frequent event combinations. The latter strategy can be tailored to fit the focus of the task at hand, but the resulting information sample will distort other judgments for which the sample was not tailored. Any mixture of these two opposite extremes will result in an incalculable combination of both problems.

Metacognitive Myopia

Assuming complete rationality, to be sure, one might correct for any biases inherent in the sample. Estimates of any measure could be corrected downward or upward depending on the degree to which it has been oversampled or undersampled, respectively, using Bayesian calculus. However, the computational work required for such a Bayesian correction would exceed human capacity for most real problems, and the necessary statistics (base rates, likelihood ratios, conditional dependencies) are hardly ever known. For instance, to recompute the original proportion $p(+ / P_1)$ from an observed sample proportion $p^*(+ / P_1)$ of positive evaluations of provider P_1, one would have to know, for each individual observation, on what combination of factors (Providers $\times$ Domains $\times$ Valence) it was restricted. Separately for each combination of sampling constraints, the degree of over- or undersampling would have to be calculated, and the correction algorithm would have to be a weighted function of all correction factors computed for each combination of sampling constraints. It goes without saying that such a monstrous task is unlikely to be mastered. It is virtually impossible to solve because in reality we seldom know the sampling constraints, including all conditional dependencies, and we do not keep separate memories or archives for experiences that are subject to different sampling constraints.

It may be for this reason that decision makers have evolved what might be called metacognitive myopia (Fiedler, 2000; Fiedler, Freytag, & Unkelbach, 2007; Fiedler & Wänke, 2004). That is, they exhibit remarkable degrees of accuracy relative to the stimulus sample given, but they are short-sighted and sometimes almost blind for the origin and the history of the sampled data, and for obvious bias inherent in the way in which the stimulus data were generated (Juslin, Winman, & Olsson, 2000; Kareev et al., 2002; Winman & Juslin, 2006). For example, even when judges are explicitly told that information about some daily winners on the stock market is presented repeatedly and that repetitions should be discounted, their judgments are biased in favor of repeated shares (Unkelbach, Fiedler, & Freytag, 2007).

Plan of the Present Research

In the remainder of this article, I first present a simulation study that illustrates the strength and scope of the biases resulting from different sampling strategies. Complementing the verbal explanation of sampling biases and the algebraic tools that exist for the quantitative calculation of bias (Cuddeback, Wilson, Orme, & Combs-Orme, 2004; Winship & Mare, 1992; Yarnold & Soltysik, 2005), the goal of the simulation is to convey a systematic picture of the gradients of bias across the entire space of possible sampling strategies. A later section will then be devoted to experimental evidence on the performance of decision makers exposed to the ultimate sampling dilemma. Both simulations and experiments will keep within the same task setting that was used in the intro-

duction—buying computers and telephone equipment offered by three providers based on stored positive and negative feedback—within the skewed world ecology depicted in Figure 1.

Biases Resulting From Different Sampling Strategies: A Simulation Study

Method and Design

Sampling biases were simulated as a function of sampling strategies, with reference to two types of judgment tasks. One task calls for inferences of the conditional probabilities of positive outcome, given different combinations of providers and domains—p(+ / P_1, C), p(+ / P_2, C), . . ., p(+ / P_3, T)—and allowing for relative evaluations of providers and domains. Let these inferences be called *causal* or *forward* inferences, because providers and domains can be conceived as causing evaluations. The second task calls for *diagnostic* or *backward* inferences, based on reverse conditional probabilities—p(P_1 / –), p(P_2 / –),. . ., p(C / +),. . . p(P_3, T / –). Here, the respective sample statistics are used to diagnose the origin (in P_1, P_2, P_3 × C, T) of positive or negative outcomes.

In accordance with Figure 1 (upper chart), a database included 480 positive instances about provider P_1 in domain C; 240 positive instances of P_2 in C; 120 positive instances of P_3 in C, and so forth. Each simulated sample consisted of $n = 100$ observations drawn from the database within the constraints imposed by the different strategies.

Sampling strategies were manipulated to cover all reasonable possibilities. Decision makers should be completely free on each information-search trial to engage in natural sampling or to restrict the information search in one, two, or all three dimensions. Thus, the decision maker may just ask for the next piece of evidence, leaving open whether the valence is positive or negative; whether the provider is P_1, P_2, or P_3, and whether the domain is C or T. On such a natural-sampling trial, the computer will randomly draw one item from the whole database, with each item in the universe having the same probability of being drawn. Alternatively, the decision maker might want to see an observation about P_1, leaving open domain and valence, or ask for a positive item from the C domain, or a negative item about P_3 in the T domain, or solicit any other combination of {+, –, open} × { P_1, P_2, P_3, open} × {C, T, open}. The computer makes a random draw from the specified subset of all items in the population (e.g., from all positive P_1 items when positive and P_1 are asked for).

Each simulated judgment experiment consists of $N = 100$ individual judges, of which each one is exposed to $n = 100$ observations, drawn with a certain sampling strategy. Altogether, $343 = 7 \times 7 \times 7$ strategies are included, completing a 7 (restrictions on providers) × 7 (restrictions on domains) × 7 (restrictions on valence) design. Each specific restriction entails a specific assumption as to how often, across the $n = 100$ trials, sampling is unconstrained or constrained to a deliberately chosen variable level. With regard to provider restrictions, the seven manipulated levels are:

1. Unrestricted on all 100 trials (natural sampling)
2. 40 unrestricted, $30P_1$, $15P_2$, $15P_3$
3. 40 unrestricted, $20P_1$, $20P_2$, $20P_3$
4. 40 unrestricted, $15P_1$, $15P_2$, $30P_3$
5. 0 unrestricted, $50P_1$, $25P_2$, $25P_3$
6. 0 unrestricted, $33P_1$, $34P_2$, $33P_3$
7. 0 unrestricted, $25P_1$, $25P_2$, $50P_3$

Thus, across the seven levels, the proportion of unrestricted, natural sampling decreases from 100 (level 1) to 40 (levels 2–4) to 0 (levels 5–7), and within these three blocks, the enforced proportions of items drawn for the three providers change.

With regard to the sampling restrictions for both domains and providers, which are dichotomous, the following seven levels were manipulated:

1. 100 unrestricted
2. 40 unrestricted, 45C, 15T for domains; 40, 45+, 15– for valence
3. 40 unrestricted, 30C, 30T for domains; 40, 30+, 30– for valence
4. 40 unrestricted, 15C, 45T for domains; 40, 15+, 45– for valence
5. 0 unrestricted, 75C, 25T for domains; 0, 75+, 25– for valence
6. 0 unrestricted, 50C, 50T for domains; 0, 50+, 50– for valence
7. 0 unrestricted, 25C, 75T for domains; 0, 25+, 75– for valence

Based on the sample data generated by these strategies, the two judgment tasks were simulated by computing two sets of indicators (for causal and diagnostic inferences): p(+ / provider, domain): forward inferences of the likelihood of positive valence given all combinations of three providers and two domains; and p(provider, domain / –): backward inferences of the origins, in all combinations of providers and domains, of negative outcomes.

Results and Discussion

Recall that the population distribution of the skewed world in Figure 1 includes more P_1 than P_2 than P_3 data, more C than T data, and more positive than negative data. However, all pairwise correlations are zero; the ratio of positive to negative is the same (2:1) at all levels of providers and domains, just as the ratio of C to T is constant (2:1) across providers and valence, and the provider proportions are invariant across domains and valence. Thus, in reality the correct value of forward inferences, p(+ / providers, domain) is always 0.67 (see top row of Table 1). The correct backward inferences to the three providers, both from positive and negative valence, summing over domains, are always 0.57, 0.29, 0.14 (reflecting the 4:2:1 ratio). The correct backward inference to domains C and T, given any provider or valence, is

Table 1
Simulation of Forward (Causal) Inferences

Domain C, T	Provider P_1, P_2, P_3	Valence +, −	p(+/C, P_1)	p(+/C, P_2)	p(+/C, P_3)	p(+/T, P_1)	p(+/T, P_2)	p(+/T, P_3)
			Correct population values					
			0.67	0.67	0.67	0.67	0.67	0.67
			Natural sampling					
			0.66	0.67	0.65	0.67	0.68	0.65
			Output-bound sampling					
		45, 15	0.72	0.70	0.74	0.71	0.72	0.72
		30, 30	0.56	0.59	0.58	0.55	0.55	0.61
		15, 45	0.41	0.43	0.42	0.41	0.42	0.46
		75, 25	0.75	0.75	0.74	0.76	0.74	0.74
		50, 50	0.49	0.49	0.51	0.50	0.53	0.50
		25, 75	0.24	0.25	0.25	0.26	0.24	0.24
			Input-bound sampling					
	50, 25, 25		0.67	0.67	0.67	0.67	0.67	0.67
	33, 34, 33		0.67	0.68	0.68	0.67	0.66	0.70
	25, 25, 55		0.66	0.66	0.67	0.69	0.67	0.67
75, 25			0.66	0.68	0.67	0.66	0.68	0.66
50, 50			0.66	0.65	0.66	0.67	0.65	0.66
25, 75			0.67	0.67	0.67	0.67	0.67	0.67
			Joint input and output sampling					
50, 50	33, 34, 33	50, 50	0.50	0.39	0.50	0.57	0.56	0.46
	33, 34, 33	50, 50	0.43	0.53	0.54	0.42	0.53	0.56
50, 50		50, 50	0.62	0.62	0.60	0.38	0.37	0.39
50, 50	33, 34, 33		0.67	0.65	0.67	0.67	0.68	0.66
25, 75	25, 25, 50	25, 75	0.00	0.22	0.00	0.33	0.19	0.34
25, 75	25, 25, 50	75, 25	0.86	1.00	0.60	0.56	0.82	0.83
	50, 25, 25	75, 25	0.74	0.76	0.76	0.75	0.75	0.76
	50, 25, 25	25, 75	0.18	0.36	0.28	0.19	0.36	0.27
	25, 25, 50	75, 25	0.71	0.61	0.83	0.74	0.58	0.85
	25, 25, 50	25, 75	0.24	0.20	0.28	0.24	0.20	0.28

Note. C = computers; T = telephones; P = providers; p = proportion.

always 0.67 versus 0.33 (reflecting the 2:1 ratio). Deviations from these normative values in the top row of Tables 1 and 2 indicate sampling errors or biases.

Natural sampling. Consider first the simulation results for a purely natural sampling strategy (i.e., unrestricted sampling in all three dimensions on all 100 trials). As Tables 1 and 2 reveal, the average sample estimates resulting from this strategy are quite accurate for all forward and backward judgment tasks. Unrestricted sampling from a population yields unbiased estimates—an elementary statistics lesson. However, the drawback of this seemingly ideal strategy lies in the paucity of information obtained about the more infrequent event combinations. The mean number of observations (out of 100) sampled for the four rarest event combinations is less than 4; for eight event classes the mean number is less than 7.

Output-bound sampling. The next block in Tables 1 and 2 shows the impact of output-bound sampling. To the extent that decision makers themselves determine the proportion of positive versus negative outcomes, not surprisingly, forward inferences of p(+ / providers, domains) are biased toward the self-determined valence rates. For example, when search is unrestricted regarding providers and domains, but the rate of positive outcomes (across all 100 trials) is set in advance to be high (i.e., 75+, 25–), medium (50, 50), or low (25, 75), the sample estimates of p(+ / providers, domains) reflect exactly these predetermined values (cf. Table 1).

When decision makers might leave valence unrestricted on 40 trials and restrict the valence of the outcome on only the remaining 60 trials (e.g., gathering 75% negative outcomes when the aim is to diagnose origins of deficits), as evident from Table 1, the result of the mixture of natural and constrained output sampling resembles the completely restricted sampling, for obvious reasons. Mixing up 40% natural sampling (i.e., 67% positive) with 60% trials that impose only 25% positive (and 75% negative) yields an overall positivity rate of only about 42% (cf. Table 1), well below the original population value of 67%.

Thus, output-bound sampling, even when applied to only a subset of trials, leads to marked biases in forward inference tasks. As expected, to the extent that sampling is restricted (e.g., partially predetermined) in one dimension, inferences concerning that dimension are biased. Inferences in the other direction (i.e., from the restricted dimension to other dimensions) may be unbiased, provided the search strategy leaves the dimension to be inferred

Table 2
Simulation of Backward (Diagnostic) Inferences

Domain C, T	Provider P_1, P_2, P_3	Valence +, −	p(C, P_1/−)	p(C, P_2/−)	p(C, P_3/−)	p(T, P_1/−)	p(T, P_2/−)	p(T, P_3/−)
			Correct population values					
			0.39	0.19	0.10	0.18	0.09	0.05
			Natural sampling					
			0.38	0.19	0.10	0.19	0.10	0.05
			Output-bound sampling					
		45, 15	0.37	0.20	0.09	0.20	0.09	0.05
		30, 30	0.38	0.18	0.10	0.21	0.09	0.04
		15, 45	0.38	0.18	0.09	0.20	0.10	0.04
		75, 25	0.38	0.19	0.11	0.18	0.10	0.05
		50, 50	0.39	0.19	0.10	0.19	0.09	0.05
		25, 75	0.39	0.20	0.09	0.18	0.10	0.05
			Input-bound sampling					
	50, 25, 25		0.38	0.19	0.10	0.19	0.10	0.05
	33, 34, 33		0.34	0.16	0.17	0.17	0.09	0.07
	25, 25, 55		0.23	0.23	0.22	0.11	0.11	0.11
75, 25			0.30	0.13	0.07	0.28	0.14	0.08
50, 50			0.14	0.07	0.04	0.42	0.22	0.11
25, 75			0.38	0.19	0.10	0.19	0.10	0.05
			Joint input and output sampling					
50, 50	33, 34, 33	50, 50	0.12	0.22	0.20	0.18	0.14	0.14
	33, 34, 33	50, 50	0.25	0.22	0.20	0.13	0.10	0.10
50, 50		50, 50	0.22	0.11	0.05	0.35	0.18	0.09
50, 50	33, 34, 33		0.18	0.14	0.19	0.15	0.21	0.14
25, 75	25, 25, 50	25, 75	0.05	0.09	0.16	0.19	0.17	0.33
25, 75	25, 25, 50	75, 25	0.04	0.00	0.24	0.32	0.16	0.24
	50, 25, 25	75, 25	0.35	0.16	0.16	0.17	0.08	0.08
	50, 25, 25	25, 75	0.36	0.14	0.16	0.18	0.07	0.08
	25, 25, 50	75, 25	0.19	0.26	0.22	0.09	0.14	0.10
	25, 25, 50	25, 75	0.17	0.18	0.32	0.08	0.09	0.16

Note. C = computers; T = telephones; P = providers; p = proportion.

unrestricted. Thus, backward inferences from restricted valence to providers and domains closely resemble the correct rates (cf. Table 2).

Input-bound sampling. By the same token, input-bound search (i.e., total or partial restrictions imposed on the proportions of providers or domains) leads to biases in backward inferences, just as output-bound search obscures forward inferences. For example, oversampling P_3 cases and undersampling P_1 cases leads to inflated backward inferences of the likelihood that P_3 rather than P_1 was the origin of a negative (or else, a positive) outcome.

Selective input–output sampling. Note that all sampling strategies considered here merely impose constraints on base rates, rather than selective contingencies. It goes without saying that a strategy that looks for many positive outcomes in domain C but mostly looks at negative outcomes in domain T will result in an illusory contingency between domains and valence. Although such motivated, self-deceptive sampling may not be uncommon in reality (Lee, Herr, Kardes, & Kim, 1999; Schulz-Hardt, Frey, Lüthgens, & Moscovici, 2000; Wang & Lee, 2006), I exclude these blatant cases from consideration.

Summary. Thus, simulations confirm that natural sampling is, of course, unbiased but may not be feasible when the problem focus is on a rare event, for which few observations are available. When this major disadvantage of natural sampling is avoided through an active information search, the resulting samples are biased in those dimensions that have governed the information search process. Selective focusing on positive or negative outcomes (i.e., output-bound sampling), while informing unbiased backward estimates p(provider, domain / –), causes biases in forward evaluative judgments of p(+ / provider, domain). Selective focusing on particular providers and domains (i.e., input-bound sampling) yields unbiased forward judgments p(+ / provider, domain) but biased backward judgments of p(provider, domain / –). While the biases resulting from pure strategies focusing on a single dimension can be anticipated intuitively, the systematic simulations reveal that mixed strategies (e.g., input-bound sampling on some trials and output-bound sampling or multiply constrained sampling on other trials) can also lead to strong biases in both forward and backward inferences (see bottom blocks in Tables 1 and 2).

Experimental Investigation of the Ultimate Sampling Dilemma

The simulations presented so far provide an overview of the strength and the boundary conditions of the biases resulting from a systematic variation of sampling strategies. Let us now go one step further and investigate the sampling dilemma experimentally, using human participants rather than computer algorithms. It can be expected that when presented with the same task as the simulation program, decision makers will face the same strategy selection problem. They might engage in purely natural sampling, never constraining their search on any dimension or focusing on specific problem aspects. Such a strategy would, of course, leave their samples unbiased. However, again, undirected search would be very uneconomic; extremely large samples would be needed to fill the most infrequent cells with a reasonable number of observations. Memory capacity would be overwhelmed, and motivation would be exhausted. Therefore, rather than using natural sampling, decision makers can be expected to actively focus on task-relevant information. However, the price for such a focused search is that the resulting samples can be trusted for only some judgments but not others. Only estimates of those variables that have not influenced the sampling process will be unbiased.

Several previous studies have documented judgment biases that reflect hard-to-control sampling biases imposed by nontransparent environments (Fiedler, Brinkmann, Betsch, & Wild, 2000; Fiedler, Walther, Freytag, & Plessner, 2002; Juslin et al., 2007). However, prior studies did not tackle sampling effects when the information search is completely under the judges' control. Particularly, no prior research has addressed the ultimate sampling dilemma, that is, the trade-off between natural sampling and selective, focused sampling strategies. How often will participants spontaneously engage in natural sampling under ideal conditions, and how often will they actively constrain their sample? If they constrain the information search, will they do it consistently, or change their strategy from trial to trial, producing complexly mixed samples that are multiply biased?

Empirical answers were sought in two experiments. In Experiment 1, the information search was fully controlled by the participants. Different search strategies were solicited, though, through manipulations of the task focus, or hypothesis to be tested. In Experiment 2, natural sampling was enforced. Both experiments together provide empirical evidence on how people attempt to handle the ultimate sampling dilemma.

Predictions

The main predictions derive from the analysis of the ultimate sampling dilemma and from the simulation results. Regarding Experiment 1, we expected that natural sampling would occur very rarely when participants can freely choose their strategy. Instead, pragmatic time limits and memory constraints should force participants to tune their information search to those aspects that are the focus of task instructions, taking into account that (some of their) judgments reflect severe sampling biases. Moreover, we expected that judges would not confine the use of sampled information to only those tasks for which it was tailored. Rather, they should also base their judgments on sample proportions when judging those variables on which the sampling process has been conditionalized, thus reflecting metacognitive myopia (cf. Fiedler et al., 2000; Fiedler & Wänke, 2004).

Thus, they should readily rely on the same samples for forward and backward judgments, regardless of whether sampling had been contingent on the independent variables (provider, domain) or the dependent variable (valence). As a consequence, output-bound sampling (i.e., obscuring the valence base rates) should result in biased forward judgments of p(+ / providers, domains). Similarly, input-bound sampling (i.e., constraining the information search to specific providers or domains) should produce biases in backward diagnoses of negative outcomes, that is, in ratings of p(provider, domains / –). Mixed-sampling constraints (i.e., obscuring the base rates of two or more variables) should produce biases in either direction. When natural sampling is enforced in Experiment 2, the typical biases resulting from selective sampling should be eliminated, but new problems should arise. Sampling errors and regression effects (Fiedler, 1996; Furby, 1973; See, Fox & Rottenstreich, 2006) should render judgments about the least frequent design combinations extremely inaccurate.

Experiment 1

Participants were asked to take the role of a leading manager whose task is to purchase hardware equipment for an organization. The cover story said there were three providers—MediaCom, EMG, and Hi-Tech (in the following denoted P_1, P_2, and P_3, respectively)—offering products in two domains (computers and telecommunication) and that former customers' positive and negative experiences were stored in an exhaustive electronic database. Participants were free to gather as many observations from the database as they considered appropriate. The task focus was manipulated to induce forward inferences versus (diagnostic) backward inferences. Either forward comparative evaluations of the positivity of providers, p(+ / providers, domains), or backward diagnostic judgments of the providers responsible for negative outcomes, p(providers / –), were called for. Another manipulation, provider focus, consisted of the instruction to compare one specific provider, either P_1 or P_3, with the others. When the focus was on the most frequent provider, P_1 (see Figure 1), input-bound information sampling should accentuate the skewed provider base rates. When the focus was on the rarest provider, P_3, the skew of provider base rates should be reduced, eliminated, or even reversed.

The first prediction, to repeat, was that natural sampling should be rare. Most participants should resort to selective sampling, producing some mix of input-bound and output-bound sampling. Second, a task focus on positive evaluation, p(+ / providers, domains), should induce predominantly input-bound sampling (by providers and domains), and if output-bound search occurs, the focus should be on positive outcomes. In contrast, a task focus on diagnosing deficits, p(providers, domains / –), should encourage output-bound samples biased toward negative outcomes and, in the case of input-bound sampling, enhanced interest in the focal provider. And third, depending on the degree of input-bound and output-bound sampling—which can vary between instruction conditions and between individual judges—biases should carry over to backward and to forward judgments, respectively. More positive forward judgments are predicted when output-bound sampling concentrates on positive outcomes, encouraged by a task focus on

p(+ / providers, domains) rather than negative outcomes, given a focus on p(providers / –). The strength of these biases should correlate across judges with the strength of sampling biases. Backward (diagnostic) judgments of the origins of negative outcomes, p(providers / –), should tend to attribute the origin to the most frequently encountered provider. Thus, the main target of diagnostic judgments should be P_1 by default, and under P_1 focus instructions in particular, but not (to the same degree) under P_3 focus instructions. More generally, a focus on P_1 as the provider of main interest should increase the skew of the database and thereby strengthen the association between P_1 and other prevalent trends in the sample. A focus on P_1 should strengthen attribution of negative outcomes to P_1 in backward diagnostic judgments. A P_1 focus may also strengthen the learned association of P_1 and positive outcomes in forward evaluations. These tendencies should be attenuated or reversed when the focus is on P_3.

Method

Participants and design. Fifty-six male and female students of the University of Heidelberg participated either for course credit or for payment. They were randomly assigned to one of four groups representing all combinations of provider focus (on P_1 vs. P_3) and task focus (positive evaluation vs. diagnosing deficits).

Materials and procedure. Participants arrived alone or in groups of two to six. They were seated in front of separate computers that administered instructions, stimulus presentation, and dependent measures. Instructions consisted of the cover story—to play the role of a manager whose task is to find out the best provider for purchasing computers or telephone hardware, based on former customers' positive and negative reactions concerning all three providers in both product domains. Then, in the specific part of the instructions, two aspects were manipulated—task focus and provider focus—between four experimental groups:

1. In the positive evaluation, P_1 focus condition, judges were asked to make "forward evaluative inferences of the positivity of provider P_1 in comparison to other providers."

2. In the positive evaluation, P_3 focus condition, judges were asked to make "forward evaluative inferences of the positivity of provider P_3 in comparison to other providers."

3. In the diagnosing deficits, P_1 focus condition, judges had to make "backward inferences about causes of negative outcomes originating in P_1 compared to other providers."

4. In the diagnosing deficits, P_3 focus condition, judges had to make "backward inferences about causes of negative outcomes originating in P_3 compared to other providers."

The translated text in the Appendix shows that participants were explicitly instructed to make inferences about the evaluation of providers in the whole database, or inferences about the origins of deficits in the database, as distinguished from the sample. It was then explained at length that participants could sample as many observations as they liked, from the database in which the reactions of former customers had been stored. They were free to use any search strategy they wished. On every trial they could call for either an item drawn at random from the database (fully unconstrained) or an item about provider P_x drawn at random from all P_x entries in any domain or valence category, or any positive reaction from the computer domain, about any provider, or any other combination constrained in 0, 1, 2, or all 3 dimensions. A $2 \times 3 \times 2$ cube was presented graphically on the screen, with the rows labeled "Computers" and "Telecommunication"; the columns labeled "MediaCom," "EMG," and "Hi-Tech"; and the foreground and background slices labeled "Positive" and "Negative." Below the cube, the response keys that could be used to constrain sampling in any subset of the three dimensions were marked in three rows (i.e., the Y and U keys in the upper row to select domain C or T, respectively; the G, H, and J keys in the middle row to select provider P_1, P_2, or P_3, respectively; and the B and N keys to call for a positive or negative outcome, respectively). They could fix any value on any dimension or leave a dimension open. The graphical display supported the instructions such that when a certain value on a dimension was fixed, the other values disappeared (e.g., when domain C was chosen, only the upper row of the cube remained; when P_1 was chosen, the other columns were removed from the display, etc.).

After the participant indicated his or her constraints, the computer randomly selected one out of all items in the database that met the constraints chosen. The database was the same population distribution as in the simulation above (see Figure 1). If the item drawn was positive, the computer selected a positive comment from a pool of 240, such as, "If one needs maintenance, somebody is immediately available." If it was negative, the comment was selected from a pool of negative ones such as, "If one needs maintenance, nobody is available." This item was presented for 3 s on the screen and then inserted in the cube position that corresponded to the domain, the provider, and the valence. Participants knew that they could terminate the information search at any time by pressing the Escape key.

Dependent measures. The main dependent measures were percentage inferences from the sample observations to the database. Participants were reminded of the distinction between the sample they had drawn and the overall database, and they were then asked to infer the percentage of positive entries in the database concerning each provider: "What is your estimate of the proportion of positive information entries in the entire database (across product domains) for the provider MediaCom / EMG / Hi-Tech, over all information stored about this provider?" They were then asked to make, in addition to these forward inferences, backward inferences of the proportion of deficits that were due to each provider: "Now consider exclusively negative information. Please estimate what percentage of all negative information in the entire database originates in the provider MediaCom / EMG / Hi-Tech." (The same backward inferences were solicited for the origins of positive outcomes, with similar results, not reported here).

At the end of the session, the three-dimensional cube appeared again on the screen, just as during the stimulus presentation, and judges were asked to estimate (in cardinal frequencies) how many observations they had sampled from each cell of the $2 \times 3 \times 2$ scheme. Although the judges were possibly influenced by the preceding inferences, these sample estimates were included if only to ensure that judges were aware of the distinction between the sample and the population.

Table 3
Characteristics of Spontaneously Gathered Samples in Experiment 1

Variable	Task and provider focus				
	Forward positive P_1 focus	Forward positive P_3 focus	Backward negative P_1 focus	Backward negative P_3 focus	Overall
Mean sample size	52.00	62.71	33.86	56.57	51.29
SD	33.76	41.65	27.24	38.75	36.43
Natural sampling proportion of trials	.06	.10	.19	.08	.11
p(domain unspecified)	.18	.20	.49	.49	.34
p(domain specified)	.82	.80	.51	.51	.66
p(C called for)	.45	.43	.25	.28	.35
p(T called for)	.37	.37	.26	.23	.31
p(C in sample)	.576	.559	.566	.607	.577
p(provider unspecified)	.26	.27	.51	.55	.40
p(provider specified)	.74	.73	.49	.45	.60
p(valence unspecified)	.17	.44	.28	.24	.28
p(valence specified)	.83	.56	.72	.76	.72
p(+ called for)	.51	.33	.24	.15	.31
p(− called for)	.32	.23	.48	.61	.41
p(+ in sample)	.64	.63	.41	.30	.495

Note. P = provider; p = proportion; C = computer; T = telephone.

Results and Discussion

Basic sample data. Consider first some basic descriptive data about the samples drawn in the present task situation. The average size of the self-determined samples across all conditions was 51.29. As Table 3 shows, sample size was somewhat larger when the focus was on the rare provider P_3 rather than P_1, reflecting the need to sample longer where environmental supply for the focal provider was low.

As expected, natural sampling rarely occurred. The proportion of trials in which the average participant engaged in an unconstrained search was 0.11 across all conditions, ranging between 0.08 and 0.19 under specific instructions (cf. Table 3). No differences between conditions were significant (all Fs < 2). Only one participant engaged in natural sampling consistently, across all trials, and another one did for 98% of the trials. All other participants chose natural sampling for less than 50% of their trials.

Instead, virtually everybody constrained the information search in one or more dimensions on a large part of all trials. The average prevalence of trials constraining search to a specific domain was .66 across all conditions, .81 for (forward) positive evaluation as compared with .51 for (backward) diagnosing deficits. The corresponding task-focus main effect was significant, $F(1, 52) = 13.75$, $p = .001$. Similarly, the proportion of trials in which one specific provider was fixed was higher for positive evaluation (.74), which is a forward task, than for the backward task, diagnosing deficits (.53), $F(1, 52) = 5.66$, $p = .05$ (overall average = .63). Together, these two findings provide a successful manipulation check. Apparently, forward-evaluation instructions induced more experimental strategies (i.e., search conditionalized on the independent variables, domains, and providers) than did backward-diagnosing instructions.

Whereas the tendency to conditionalize the search on providers and domains (i.e., input-bound sampling) is reminiscent of experimental strategies, the strong output-bound sampling tendency to call for either positive or negative outcomes is more surprising. On average, the proportion of trials in which participants restricted the outcome (to positive or negative) was .718 across all conditions. Curiously, the output-bound search (cf. Table 3) was most elevated in the forward-evaluation / P_1 focus condition (.83), but the differences between experimental groups were not significant.

Unfortunately, due to a mistake in the computer program, when sampling was contingent on a provider, the specific provider chosen was not registered. This precluded a systematic analysis of provider base rates in the sample beyond the correctly assessed fact that a provider was rarely left unspecified (i.e., only in .40 of the trials).

Sampling biases. Thus far, we have confirmed the basic premises that natural sampling occurs very rarely and that decision makers instead restrict their information search in one or more dimensions, being sensitive to the manipulations of task focus and provider focus. In the next step, we can now examine the resulting sampling biases, that is, whether the sample base rates deviated from the original population base rates. Recall that the original distribution was skewed in all three dimensions (i.e., 2:1 base rate ratios for domains and valence, and the 4:2:1 ratio for providers). This skew was clearly reduced in the samples acquired, reflecting regression toward more equal base rates (cf. Table 3). The proportion of observations drawn from the more frequent C domain was .577, due to input-bound sampling, as compared with an original base rate of .667. Likewise, the proportion of positive items decreased from .667 in the population to .495 in the sample, due to output-bound sampling.

Estimates of sample frequencies. The data registration failure for the providers chosen precludes an analogous check for this dimension, but the subjective estimates of sample frequencies afford a substitute here. The estimated frequency of the focal provider shrink from .571 in the population to .361 in the sample.

Table 4
Mean Estimates of Joint Sample Frequencies by Conditions in Experiment 1

	Computers						Telecommunication					
	P_1		P_2		P_3		P_1		P_2		P_3	
Variable	+	−	+	−	+	−	+	−	+	−	+	−
Population	.254	.127	.127	.063	.063	.032	.127	.063	.063	.032	.032	.016
Forward positive P_1 focus	.127	.097	.077	.070	.087	.072	.097	.062	.094	.067	.091	.058
Forward positive P_3 focus	.119	.060	.094	.074	.100	.069	.098	.062	.108	.057	.105	.055
Backward negative P_1 focus	.059	.173	.068	.119	.061	.074	.060	.112	.063	.098	.056	.058
Backward negative P_3 focus	.094	.088	.061	.101	.095	.123	.078	.067	.064	.071	.071	.087
Total	.100	.105	.075	.091	.085	.085	.083	.076	.082	.073	.081	.060

Note. P = provider.

As evident from Table 4, estimates of the observed frequencies of all 12 event combinations were generally regressive; that is, actually existing frequency differences were underestimated. However, it is also apparent that the average participant correctly found out the ordinal differences between domains, providers, and valence levels and that the focus manipulations exerted the intended influence. Thus, when the focus was on P_1 rather than P_3, the higher prevalence of P_1 data was more apparent. And the high prevalence of positivity was more evident when the task focus was on evaluating positivity rather than on diagnosing deficits. All these differences proved significant in a Domain × Provider × Task Focus × Provider Focus analysis of variance (ANOVA) that is not reported here to save space.

More importantly, the sample-frequency estimates allow for a first check on the major predictions derived from the simulation study, pertaining to biases in forward evaluations due to output-bound sampling and biases in backward diagnoses due to input-bound sampling. As judges varied in the proportion of positive data they solicited, the sampling bias resulting from output-bound sampling is evident in a correlation as high as $r = .867$, $p < .001$, between the individual proportion of positive items sampled, across all domains and providers, and the estimated positivity proportion (i.e., the sum of all six positive frequency estimates divided by the sum of all 12 estimates). With regard to input-bound sampling, which could only be examined for domains, the proportions of items chosen from the C domain was similarly correlated ($r = .760$, $p < .001$) with the pooled estimated proportion of C items in the sample.

Judgment biases. Having confirmed that the restricted information search actually produced biased samples, and that judges recognized these biases in the samples, we now turn to the question of major interest, namely, whether biased samples actually led to biases in the final population inferences. Let us first consider forward inferences of the positivity of the three providers, p(+ / P_1, P_2, P_3), as assessed in three direct ratings. Recall that output-bound sampling of positive outcomes and, consequently, positively biased population inferences were predicted when the task focus was on positive rather than negative information.

Table 5 provides the pertinent means as a function of focus conditions. Apparently, both predictions are clearly borne out. Positivity rates in the samples were higher for positive evaluation (.64 and .63 for P_1 and P_3 focus, respectively) than for diagnosing deficits (.41 and .30, respectively). Accordingly, the average rated percentages of positive information in the database (i.e., the average of all domain–provider combinations given in Table 5) was higher under the former task focus ($M = 48.56$) than the latter ($M = 38.10$). This difference proved to be significant as a task-focus main effect, $F(1, 52) = 9.33$, $p < .01$, in an ANOVA-treating task focus and provider focus as between-subjects factors and a contrast between ratings of P_1 and average ratings of P_2 and P_3 as the dependent variable. As predicted, then, the deliberate, selective sampling of positive information did not prevent judges from basing their population inferences on the biased positivity proportions resulting from their own output-bound sampling. Across all judges, the summed positivity rating

Table 5
Mean Population Inferences (in Percentages) by Conditions in Experiment 1

	Forward inferences p(+/providers)			Backward inferences p(providers/−)		
Variable	P_1	P_2	P_3	P_1	P_2	P_3
Objective percentage	66.67	66.67	66.67	57.14	28.57	14.29
Forward positive P_1 focus	54.14	37.71	40.86	40.43	30.86	35.14
Forward positive P_3 focus	53.71	51.57	53.36	40.93	36.29	39.86
Backward negative P_1 focus	35.71	40.64	41.50	44.71	32.00	26.71
Backward negative P_3 focus	39.64	36.07	35.00	24.81	33.88	40.24

Note. p = proportion; P = provider.

(over all providers) correlated significantly with the individual bias toward positive information ($r = .413$, $p < .01$).

The Provider Focus × Task Focus interaction was also significant, $F(1, 51) = 15.11$, $p < .001$. As predicted, the task focus effect was mainly due to judges who focused on provider P_1. Due to the highest density of information associated with P_1, this provider was most strongly associated with the predominant valence.

Also as expected, the impact of input-bound sampling biases is manifested in backward attributions of negative outcomes to provider P_1, as compared with the other two providers, P_2 and P_3. From the means in Table 5 it is evident that negative outcomes were generally attributed to P_1, who was most frequent in the database, except for the backward P_3 focus condition, in which P_3 was associated with a focus on negative observations. The deviant result for this group was manifested in a three-way Provider Contrast × Task Type × Provider Focus interaction, $F(1, 52) = 7.81$, $p < .01$, as well as a two-way Task Type × Provider Focus interaction, $F(1, 52) = 12.09$, $p < .01$. Altogether, these findings corroborate the double assumption that the preference for restricted rather than natural sampling induces distinct sampling biases, which in turn carry over to analogous judgment biases. These findings are also reflective of metacognitive myopia, that is, judges' failure or inability to control and correct for self-generated sampling biases.

Experiment 2

In Experiment 1, when search strategies were completely free, participants rarely chose natural sampling. They rather constrained the information search to specific domains, providers, and valence levels. As a consequence, the resulting samples exhibited distinct biases, and the final judgments were biased accordingly. One might conjecture that the entire problem merely lies in the reluctance to apply natural sampling. However, the ultimate sampling dilemma suggests good reasons to suspect that natural sampling will also lead to inaccuracy.

A second experiment, which replicated Experiment 1 in all respects except that natural sampling was enforced, was therefore conducted. All participants were exposed to an unconstrained random sample of observations drawn from the same population as in Experiment 1. The predictions were straightforward. On aggregate, across all judges, the resulting sample should provide an unbiased picture of the universe. However, at the level of individual judges, samples should be impoverished with respect to rarest events, leading to inaccurate and highly regressive judgments. Moreover, as the degree of regression (i.e., the underestimation of positivity) increases with decreasing sample size, new judgment biases should come in through the back door, via differential regression. The same high degree of positivity should be judged to be lower for infrequent than for frequent providers.

Method

Participants and design. Thirty-nine male and female students of the University of Heidelberg participated. They were randomly assigned to the same four instruction conditions (resulting from orthogonal crossing of task focus and provider focus) as in Experiment 1. Because natural sampling is fully random, neither task focus nor provider focus could affect the sampling stage. The possibility cannot be excluded, though, that the two treatments might still influence selective memory and attention to task aspects and providers during the final judgment stage.

Materials and procedure. The same computer program was used as in Experiment 1, including all instructions (except for a few necessary changes) and dependent measures. Rather than being allowed to constrain the information search deliberately on every trial, participants were told that every sampled item represented a random draw from the entire database. It was explicitly added that "every entry in the database, regardless of its reference to a specific provider and domain, has the same chance of being drawn." Stimuli appeared at a constant rate of 3 s per observation. Information search was terminated as the participant pressed the Escape key.

Results and Discussion

Basic sample data. The average participant sampled 35.21 observations ($SD = 20.90$). As expected, natural sampling resulted in impoverished data about the less frequent cells of the Product Domains × Providers × Valence Levels design. Although the overall distribution of sampled observations (across all participants) closely resembled the population distribution (cf. Table 6), the average individual sample included less than two observations in five out of the 12 cells. Inferences from these infrequent data would be extremely unreliable, even with doubled sample size. This is also evident from the large number of judges (out of 39) who based their estimates on observed frequencies smaller than, or equal to, 0, 1, or 2 (cf. Table 6).

Sample estimates. Nevertheless, the advantage of natural sampling is apparent in the absence of biases at the level of average judgments (across all participants) of the 12 Domains × Providers × Valence combinations (cf. Table 6). The ordinal relations are correctly reproduced, as the average judge correctly estimated having sampled more positive ($M = 14.83$) than negative ($M = 11.16$) observations, more data for domain C ($M = 14.03$) than for domain T ($M = 11.95$), and decreasing frequencies from P_1 to P_2 to P_3 ($M = 14.46$, 13.68, and 10.85, respectively). However, in spite of the absence of crude biases at the group level, sample frequency estimates were highly regressive, yielding ratios much smaller than the actual ratios of 2:1 or even 4:1. When frequency estimates were transformed into proportions to render them comparable to population proportions (cf. Table 6), large frequencies were clearly underestimated, whereas small frequencies were overestimated.

Note, however, that an unsystematic regression error can turn into a systematic bias when the strength and direction of regression varies between judgment targets. This is apparent from an analysis of inaccuracy scores, defined as signed differences between subjective estimates (transformed to proportions) and objective proportions (cf. Table 6). Regression effects are manifested in negative difference scores for the more frequent levels on the domains, providers, and valence factors but positive scores for infrequent levels (reflecting regression). In a repeated-measures ANOVA (including all 39 participants), biases resulting from differential regression were apparent in strong main effects on all three factors. The more frequent domain, C, was underestimated compared to T, $F(1, 38) = 39.44$, $p < .001$. The most frequent provider, P_1, was underestimated relative to the other domains, $F(2, 76) = 53.76$, $p < .001$, and the rate of positive (vs. negative) observations was underestimated as well, $F(1, 38) = 18.35$, $p < .001$.

Table 6
Mean Estimates of Joint Sample Frequencies by Conditions in Experiment 2

Variable	Computers						Telecommunication					
	P_1		P_2		P_3		P_1		P_2		P_3	
	+	−	+	−	+	−	+	−	+	−	+	−
Population proportion	.254	.127	.127	.063	.063	.032	.127	.063	.063	.032	.032	.016
Sample proportion	.280	.107	.121	.065	.066	.035	.132	.057	.062	.028	.034	.014
Estimated proportion	.157	.093	.119	.079	.080	.050	.103	.072	.093	.062	.055	.038
Estimated proportion − population proportion	−.10	−.03	−.01	.02	.02	.02	−.02	.01	.03	.03	.02	.02
Mean *n* sampled	9.56	3.82	4.28	2.31	2.49	1.13	4.64	1.95	2.33	1.03	1.18	0.49
Number of *S*s with $n = 0$	0	2	0	6	5	11	1	7	7	14	15	24
Number of *S*s with $n \leq 1$	0	9	2	17	17	29	3	18	20	28	26	35
Number of *S*s with $n \leq 2$	1	19	11	24	24	35	9	31	24	36	33	39

Note. P = providers; *S* = subject.

Moreover, the strong regression tendency produced three two-way interactions. The tendency to underestimate positive and to overestimate negative observations increased from domain T to C, $F(1, 38) = 9.73$, $p < .01$, and from P_3 to P_2 to P_1, $F(2, 76) = 8.67$, $p < .001$; and the inaccuracy difference between providers was more apparent for domain C than for T, $F(2, 76) = 19.42$, $p < .001$. Thus, as different aspects were unequally affected by regression, the resulting subjective estimates were markedly biased in all three dimensions as well as in their interactions. Such a differential pattern of over- and underestimation can be expected to be typical of natural sampling in skewed environments.

Biased population inferences. Let us finally consider the crucial inferences about the population, as distinguished from the estimates of sample frequencies. It was expected that similar biases as in the sample proportions should arise for the population inferences, reflecting the differential regression resulting from extremely unequal cell frequencies. Thus, the same trends (e.g., the same 2:1 ratio of positive to negative outcomes) should be more apparent for frequently observed than for rarely observed events. In particular, in forward evaluations of providers' assets the prevalent positivity should be rated highest for the most frequent provider, P_1; intermediate for P_2; and lowest for the least frequent provider, P_3.

To capture this sort of bias, a weighted sum score was computed (cf. Table 7) by multiplying the positivity ratings of P_1, P_2, and P_3 by coefficients +1, 0, and –1, respectively. The higher (i.e., the more positive) this score, the stronger the expected bias to overestimate P_1 and to underestimate P_3 in forward ratings of p(+ / providers). For backward diagnostic inferences, an analogous weighted score of p(providers / –) reflects the tendency to attribute deficits to the frequent (P_1) rather than infrequent (P_3) source. It was further expected that infrequency effects might come to interact with the provider focus manipulation. Higher P_1 than P_3 judgments—both in terms of more positive forward inferences and in terms of more negative backward diagnoses—should be accentuated when an attention focus on P_1 reinforces the actually existing density differences.

Consider first the forward inferences of p(+ / providers). As expected, the composite scores tended to be positive, $M = 14.05$, $t(21) = 2.87$, $p < .01$, when the provider focus was on P_1, but not when the focus was on P_3, $M = –2.82$, $t(16) = –0.365$. Thus, when the focus was consistent with the prevalence of providers, inferred positivity decreased from P_1 to P_2 and P_3, although the actual positivity rate in the population was constant. With regard to backward inferences of p(providers / –), the tendency to attribute negative information to frequent rather than infrequent providers reveals a similar bias to attribute deficits to the most prevalent provider, but only when the focus was on P_1, $M = 13.77$, $t(21) = 2.40$, $p < .05$, rather than on P_3, $M = –9.88$, $t(16) = –1.33$. (Ironically, a reverse bias to praise P_1 could be shown for backward inferences of the origins of positive outcomes.)

When backward inferences were analyzed as a function of the two between-subjects factors—task focus and provider focus (see Table 7)—the bias score tended to be strongest when the provider

Table 7
Mean Population Inferences (in Percentages) by Conditions in Experiment 2

Variable	Forward inferences $(+1)p(+/P_1) + 0p(+/P_2) + (-1)p(+/P_3)$	Backward inferences $(+1)p(P_1/-) + 0p(P_2/-) + (-1)p(P_3/-)$
Population	0.00	42.86
Rating overall	6.69	3.46
Forward positive P_1 focus	18.25	20.08
Forward positive P_3 focus	−1.67	−1.67
Backward negative P_1 focus	9.00	6.20
Backward negative P_3 focus	−4.13	−19.13

Note. p = proportion; P = provider.

focus was actually on P_1 rather than on P_3, yielding a significant provider focus main effect, $F(1, 35) = 6.61$, $p < .05$. The forward-inference bias favoring P_1 also tended to be most pronounced when the focus was on P_1 rather than P_3, although the provider focus main effect was not quite significant, $F(1, 35) = 3.04$. All other effects were nil ($F < 1$).

Altogether, Experiment 2 confirms that although natural samples are unbiased at first, the final judgments run into new problems. Information about rare events is impoverished. Moreover, regression error can be very strong, and differential regression can produce a new class of biases that can distort the relative impression of frequent and infrequent sources.

General Discussion

Summary of Results

The present inquiry into the ultimate sampling dilemma revealed what it had to reveal on a priori grounds, namely, that there is little chance to evade this dilemma of the empirical world. Computer simulations and two experiments confirm the predictions derived from an ecological analysis of the biases and deficits of the information samples on which decision makers have to rely.

Computer simulations outlined the scope of the ultimate sampling dilemma. Although the simulations merely demonstrated what could be derived analytically, they provided a systematic picture of the generality and scope of the biases resulting from all kinds of sampling strategies. Two experiments provided an empirical answer to the question of how human decision makers deal with the dilemma. Experiment 1 corroborated the expectation that hardly anybody engages spontaneously in a pure natural-sampling strategy. Rather, people tended to sample information predominantly from those cells that were relevant to the decision task. When the task focus was on the relative positivity of a specific provider, participants would gather mostly positive information about the provider under focus. Although focusing on a specific provider alone might have resulted in unbiased forward inferences of that provider's positivity, a simultaneous (output-bound) focus on positive valence rendered even forward judgments distorted. Conversely, when the focus was on the explanation of negative outcomes, output-bound sampling of negative events was facilitated. As a consequence, forward judgments were biased toward negative information. However, backward (diagnostic) inferences of the origins of negative outcomes were also biased to the extent that sampling was at the same time input-bound, concentrating on a provider of main interest. Thus, mixed and changing sampling strategies, rather than producing a balanced picture of reality, actually contaminated judgments in either direction.

In Experiment 2, the same task instructions were used, but natural sampling was enforced. Although the resulting overall sample was indeed unbiased, information about infrequent event classes was scarce and often insufficient. As a consequence, decisions pertaining to such rare event classes were especially inaccurate and regressive. However, in addition to unsystematic error leading to inaccuracy, the highly unequal frequencies of observations induced a new kind of bias reflecting differential regression. Although positive outcomes were constantly frequent and negative outcomes constantly infrequent across all providers, this difference was more effectively recognized for the most frequent provider but often missed and underestimated for the rarest provider. Thus, the regressive tendency to underestimate real frequency differences, which characterizes all memory-based frequency estimates (cf. See, Fox, & Rottenstreich, 2006), was most apparent where samples were most impoverished, consistent with previous demonstrations of differential regression due to unequal sample sizes (Fiedler, 1996; Fiedler et al., 2002; Fiedler & Walther, 2003; Zuckerman, Knee, Hodgins, & Miyake, 1995). Therefore, natural sampling did not provide a useful remedy at all but led to strong unsystematic error as well as new biases coming in through the back door.

Can Decision Makers Evade the Sampling Dilemma?

In reality, decision makers generally face the same dilemma as the participants in the present experiments. On the one hand, to gather sufficient information about the crucial aspects of a decision problem within a reasonable time, they may try to adopt the strategy of a clever naïve research designer and actively sample those events that are most relevant to the decision problem in front of them. The resulting samples, which are inevitably selective, can yield unbiased answers to the problems for which they were designed. But as new problems arise and the same sample is used to answer them, the responses can be seriously misled. Specifically, to the extent that decision makers have engaged in input-bound sampling, forward inferences are likely to be accurate but backward inferences are biased toward those input levels that are overrepresented in the sample. Conversely, to the extent that decision makers engage in output-bound sampling, their backward inferences are likely to be accurate, but forward inferences will be biased toward the outcome valence that is overrepresented in the sample (Dawes, 1993).

Therefore, decision makers ought to use a sample only for the very purpose for which it was tailored. However, this rule is hard to handle, because world knowledge is usually not organized by specific sampling procedures. Neither human memory nor external archives of knowledge regularly keep separate samples for different search strategies tailored for distinct purposes. Metacognitive myopia often prevents decision makers from even noticing the constraints imposed on their information input.

On the other hand, decision makers may try to circumvent the biases inherent in purpose-bound samples, refraining from all active information search and resorting instead to the completely passive strategy of natural sampling. Ideally, this produces an unbiased sample; however, it is impoverished with regard to rare events. If real decisions do pertain to such rare events, they inevitably suffer from the scarceness of data. Such unbiased but often skewed natural samples may induce, in addition to unsystematic error, systematic judgment bias of a different kind, due to differential regression. The same trend (e.g., the same high rate of, say, 80% positive feedback to two consumer products) will be more likely recognized and judged less regressively when based on a large rather than a small sample, because regression decreases with reliability and sample size. Thus, an 80% positivity rate is more readily inferred from 24 positive and 6 negative observations about one product than from, say, 8 positive and 2 negative observations about another product, even when the true proportion is 80% for both products. Note that from a Bayesian point of view, the inequality can be justified as rational, because the posterior probability for the positivity hypothesis, given 24 positives and 6

negatives, is indeed higher than the posterior probability given 8 positives and 2 negatives. Ironically, however, what is "rational" in the Reverend Thomas Bayes' name does not protect one from missing the true state of nature when sample sizes are unequal.

Does a Mixed Sampling Strategy Offer a Remedy?

Could a combination or compromise of both natural and tailored sampling help to overcome these problems? The answer is, very likely, no. Engaging in mixed strategies will not ameliorate the problem; in fact, it may even worsen the situation, causing biases in either direction. Given a complex mix of sampling strategies, even the possibility of Bayesian correction may be no longer viable. If each of, say, 100 observations is conditional on a specific combination of search constraints, such as time, media, information source, set of providers considered, focus on valence, product domains, and so forth, and if most observations do not even reveal their underlying constraints, then how could even the best Bayesian statistician reconstruct the underlying population parameters?

Does a Two-Stage Sampling Strategy Provide a Remedy?

Rather than sampling strategies that are mixed up and changed from trial to trial, a systematic two-stage sampling scheme may provide a more viable alternative. When entering a new task or environment, organisms may first engage in an extended period of natural sampling, in a purposeless exploration process. Once the basic parameters of the environment have been extracted, organisms may then in a later stage tackle specific problems, adopting the purposeful strategies of a clever research designer. Background knowledge about domain-specific base rates may sometimes even allow for calculations of inverse inferences from samples designed for different purposes.

Whether such a two-stage sampling process is feasible and realistic is an open empirical question. In the absence of direct evidence, introspection suggests that we do not administrate separate memory systems for general base rate knowledge and specific problem knowledge. If anything, some pertinent findings on reasoning about Simpson's paradox and other complex contingency problems (Fiedler & Freytag, 2004; Fiedler et al., 2007; Fiedler, Walther, Freytag, & Nickel, 2003; Spellman, Price, & Logan, 2001) suggest that people have tremendous difficulty understanding that different relationships may hold at different levels of analysis. In any case, a two-stage sampling process places extremely strong demands on the accessibility of information (i.e., when ecologies do not render all information equally observable), on source memory (when knowledge has to be split by sampling stages), and on memory administration (when changing environments call for continuous updating of base rates and contingencies). Nevertheless, an interesting goal of future decision research is to study adaptive changes of information sampling over time and as a function of different affordances.

Is Natural Sampling Possible at All?

At a more fundamental level, truly natural sampling may not be possible at all. At any point in time or space, information about different objects is not equally available. This restricts nature's ability to provide us with fully unrestricted samples. Consider the Internet for a nice thought experiment. Whoever has tried to search for specific targets on the Internet will agree how hard and actually impossible it is to solicit a natural, unrestricted sample that warrants "true" base rates. The resulting sample of Internet sites is always conditional on, and biased toward, the specific keywords used, the position of different sources in the hierarchy of search engines, the communicability of different contents, and the availability of Internet sites for particular topics. Literally, the Internet does not allow for natural sampling.

A related issue is whether absolute base rates—base rates that hold across time, space, cultures, markets, and decision contexts—exist at all. To the extent that base rates change over time or between regions or cultures, any "true" population base rates must be arbitrary. Is the true base rate today's base rate, or tomorrow's? Is it the monthly or the annual base rate? And what level of domain specificity should be chosen for "the" base rates? Given all these open questions, whether base rates are really natural can be hardly determined for sure.

What Is the Constructive Lesson to Be Gained From the Ultimate Sampling Dilemma?

What insights and implications can be gained from this inquiry into the ultimate sampling bias? Is the inherent message really so pessimistic? I believe that the message is in fact not that pessimistic, and that quite a few optimistic aspects, both theoretically and practically, deserve to be pointed out.

On the one hand, it is important to note that the ultimate sampling dilemma is not a deficit of the human mind but a genuine property of empirical reality. Nonhuman, artificial-intelligence systems suffer from similar problems when fed with the same environmental input. In its most radical form, the dilemma reflects the fact that "true" or normatively correct answers to many probabilistic problems actually do not exist. They are indeterminate, as there is often no single reality behind the sample. There is no objectively correct sampling scheme to estimate the "true" probability that somebody will die from a car accident, or that the stock market will show a decline during the next 5 years. Any sampling scheme is conditional on a search strategy that focuses on specific sources, time frames, geographical frames, media, and categories that are arbitrarily used to define the respective reference set. Whether these sampling constraints are representative of the population cannot be determined, because the latent reality is invisible.

On the other hand, much can be learned from the more refined part of the message concerning the moderators of the sampling dilemma. Input-bound strategies in general, and experimental strategies in particular, will normally provide appropriate samples for forward (causal) inferences. Likewise, output-bound samples inform accurate backward (diagnostic) inferences. However, crucially, any given sample with a specific generation history does not warrant an unbiased, omnidirectional perspective on all possible inference directions. This restriction has to be kept in mind when responsible decisions are made in science, business, and politics. By exploiting scientific analyses and methods to recognize and admit the existence of the sampling dilemma in many practical contexts, many erroneous consequences can be avoided. In science and in expert systems at least, it must be possible to describe samples not only in terms of size and general representativeness, but also in terms of their inherent directionality and conditionality, and to explicate their inherent restricted uses.

The last and maybe most serious point, though, refers to the one aspect of the dilemma for which the human mind is to blame to a reasonable degree, namely, the metacognitive myopia that often prevents people from recognizing the pitfalls of sampling biases.

Decision makers—whether lay people or experts—take sample information for granted, uncritically and naively, even when it is obvious that samples are severely biased (Fiedler et al., 2000; Kareev et al., 2002). Maybe one of the most prominent goals of research on rationality and intellectual emancipation is to sensitize decision makers to major sampling biases in the environment (Denrell, 2005; Taylor, 1991)—due to media coverage (Combs & Slovic, 1979), restricted information access (Fiedler, 2007b; Fiedler & Walther, 2003), selective memory (Tesser, 1978), unequal communicability (Kashima, 2000), or restricted research designs (Wells & Windschitl, 1999)—to educate people about how to treat samples cautiously and recognize problems that call for a new sample tailored for the decision at hand. When it becomes clear that correcting a biased sample is not feasible, then ignoring a sample may be better, and more rational, than utilizing a biased and ill-designed sample accurately.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

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

Combs, B., & Slovic, P. (1979). Newspaper coverage of causes of death. *Journalism Quarterly, 56,* 837–843, 849.

Cuddeback, G., Wilson, E., Orme, J. G., & Combs-Orme, T. (2004). Detecting and correcting sample selection bias. *Journal of Social Service Research, 30,* 19–33.

Dawes, R. M. (1993). Prediction of the future versus an understanding of the past: A basic asymmetry. *American Journal of Psychology, 106,* 1–24.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112,* 951–978.

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

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review, 103,* 193–214.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107,* 659–676.

Fiedler, K. (2007a). Construal level theory as an integrative framework for behavioral decision making research and consumer psychology. *Journal of Consumer Psychology, 17,* 101–106.

Fiedler, K. (2007b). Information ecology and the explanation of social cognition and behavior. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 176–200). New York: Guilford.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General, 129,* 399–418.

Fiedler, K., & Freytag, P. (2004). Pseudocontingencies. *Journal of Personality and Social Psychology, 87,* 453–467.

Fiedler, K., Freytag, P., & Unkelbach, C. (2007). Pseudocontingencies in a simulated classroom. *Journal of Personality and Social Psychology, 92,* 665–667.

Fiedler, K., & Juslin, P. (2006). *Information sampling and adaptive cognition.* New York: Cambridge University Press.

Fiedler, K., & Walther, E. (2003). *Stereotyping as inductive hypothesis testing.* New York: Psychology Press.

Fiedler, K., Walther, E., Freytag, P., & Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. *Personality and Social Psychology Bulletin, 29,* 14–27.

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.

Fiedler, K., & Wänke, M. (2004). On the vicissitudes of cultural and evolutionary approaches to social cognition: The case of meta-cognitive myopia. *Journal of Cultural and Evolutionary Psychology, 2,* 23–42.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8,* 172–179.

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

Hoffrage, U., & Hertwig, R. (2006)). Which world should be represented in representative design? In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 381–408). New York: Cambridge University Press.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114,* 678–703.

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.

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General, 131,* 287–297.

Kareev, Y., & Fiedler, K. (2006). Non-proportional sampling and the amplification of correlations. *Psychological Science, 17,* 715–720.

Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin, 25,* 594–604.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Lee, H., Herr, P. M., Kardes, F. R., & Kim, C. (1999). Motivated search: Effects of choice accountability, issue involvement, and prior knowledge on information acquisition and use. *Journal of Business Research, 45,* 75–88.

Macrae, A. W. (1971). On calculating unbiased information measures. *Psychological Bulletin, 75,* 270–277.

Manski, C. (1995). *Identification problems in the social sciences.* Cambridge, MA: Harvard University Press.

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

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin and Review, 10,* 289–318.

Oaksford, M., & Moussakowski, M. (2004). Negations and natural sampling in data selection: Ecological versus heuristic explanations of matching bias. *Memory & Cognition, 32,* 570–581.

Schlesselman, J. J. (1982). *Case-control studies. Design, conduct, analysis.* New York: Oxford University Press.

Schulz-Hardt, S., Frey, D., Lüthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology, 78,* 655–669.

See, K. E., Fox, C. R., & Rottenstreich, Y. S. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory & Cognition, 32,* 1385–1402.

Spellman, B. A., Price, C. M., & Logan, J. M. (2001). How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition, 29,* 193–208.

Stuart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1–26.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin, 110,* 67–85.

Tesser, A. (1978). Self-generated attitude change. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 289–338). New York: Academic Press.

Unkelbach, C., Fiedler, K., & Freytag, P. (2007). Information repetition in evaluative judgments: Easy to monitor, hard to control. *Organizational Behavior and Human Decision Processes, 103,* 37–52.

Wang, J., & Lee, A. Y. (2006). The role of regulatory focus in preference construction. *Journal of Marketing Research, 43,* 28–38.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115–1125.

Winman, A., & Juslin, P. (2006). "I'm m/n confident that I'm correct": Confidence in foresight and hindsight as a sampling probability. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 409–439). New York: Cambridge University Press.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18,* 327–350.

Yarnold, P. R., & Soltysik, R. C. (2005). *Optimal data analysis: A guidebook with software for Windows.* Washington, DC: American Psychological Association.

Zuckerman, M., Knee, C. R., Hodgins, H. S., & Miyake, K. (1995). Hypothesis confirmation: The joint effect of positive test strategy and acquiescence response set. *Journal of Personality and Social Psychology, 68,* 52–60.

Appendix

General Instruction Letter Provided at the Beginning of Experimental Sessions

Dear participant:

Thanks for your willingness to participate. In this study, all information is transparent. That is, the goal and purpose of the study are not kept secret. You will not be distracted from the actual purpose and no deception will be involved. And we do not try for a moment to manipulate or direct your behavior.

Your task entails a role play – you are supposed to take the role of an entrepreneur who has to make purchasing decisions – but this is quite a natural task familiar to everybody. Before you buy something, you compare different providers with regard to advantages and disadvantages and you thereby rely on experiences that others have made with the same products and providers. Accordingly, you will get access to a database containing all stored experience with the offered products. This database constitutes the reality; it is physically available on the computer and provides the graduator for your achievement. Making accurate judgments means to make judgments that fit the "reality" of the database.

If you could assess and memorize all available data, then your decisions would have to be correct. That is, there would actually be one best decision.

Like in real life, however, it is not possible to take the entirety of all information into account. Sufficient time is often lacking, and it would be much too expensive and fully unusual to base each and every decision on the entirety of all relevant information. Besides, we would run into a storage problem. Our memory would suffer from overload just like the hard-disk of our computer, let alone the problem of how to derive a decision from the insurmountable quantity of information. We would soon miss the forest for the trees.

Rather than assessing and utilizing everything – which in the era of the Internet is impossible anyway – we almost always base our decisions on a sample of information. If the sample is not too small and not distorted selectively, this is usually no problem. Then the sample affords a rather reliable picture of the reality.

Your task also consists in drawing a sample from the database, enabling you to make a correct decision. Promised: the actual differences in the database are strong enough so that a commonly chosen sample size allows you to detect these differences – provided the way the sample is drawn is appropriate to the problem. This is exactly the key to success: collecting data that are useful for the problem at hand, revealing the actual relations that hold in the database.

. . . You have to equip your company with electronic devices . . . Two electronic domains have to be distinguished, computers and telecommunication. That is, you have to purchase both computers for work stations as well as telephones, picture telephones, and cell phones for conferences. You have to compare three providers, who all offer products in both domains. Thus, the database for this problem discriminates between:

3 providers (EMG, MediaCom, Hi-Tech)

2 product domains (computers and telecommunications)

2 possible outcomes (positive or negative)

After an extended explanation of the multiple ways of constraining information search, and how to handle the keyboard, instructions were manipulated between focus conditions: Task focus = positive evaluation – provider focus = P_1 (called MediaCom).

"You are to find out whether in the total database the provider MediaCom received better evaluations than the other two providers, pooling across product domains. That is, is the relative proportion of positive observations among all observations for MediaCom higher than for the other 2 providers?" Task focus = positive evaluation – provider focus = P_3 (called Hi-Tech) "MediaCom" replaced by "Hi-Tech," otherwise identical. Task focus = diagnosing deficits – Provider focus = P_1 (called MediaCom).

"You are to find out whether in the total database negative observations most frequently originate in the provider MediaCom. Only think of the set of negative outcomes. Does the provider MediaCom appear more frequently in this reference set than the other two providers, regardless of the positive information?" Task focus = diagnosing deficits – Provider focus = P_3 (called Hi-Tech) "MediaCom" replaced by "Hi-Tech," otherwise identical.

The general instructions preceding the dependent measures read as follows:

"Now, as indicated at the outset, you will be asked to draw inferences from what you have seen to the total database. The judgments you are supposed to make below are always meant as judgments about the entire database from which you have gathered observations."

Finally, the sample estimates of the frequencies of all 12 event combinations were solicited:

"And finally, now, a few more questions about the actually observed information. Now your task is not to make inferences concerning the entire database, but to estimate the a b s o l u t e frequencies of positive and negative observations you have seen for the different providers in both product domains. How many examples (not %) did you see for the following combinations . . .?"

Received October 19, 2006
Revision received August 16, 2007
Accepted October 2, 2007 ■