SonderForschungsBereich 504

Rationalitätskonzepte,
Entscheidungsverhalten und
ökonomische Modellierung

No. 07-51

The Ultimate Sampling Dilemma in Experience-Based Decision Making

Klaus Fiedler*

July 2007

The present research was supported by various grants from the Deutsche Forschungsgemeinschaft. Thanks to Ralph Hertwig, Ulrich Hoffrage, Tobias Vogel, Peter Freytag, and Yaakov Kareev for their helpful and constructive comments on a draft of this article. Correspondence should be addressed to Klaus Fiedler, Psychologisches Institut, Universität Heidelberg, Hauptstrasse 47-51, 69117 Heidelberg, Germany. Fax: + 49 6221 547745. Email: kf@psychologie.uni-heidelberg.de

*Sonderforschungsbereich 504/ Universität Heidelberg, email: Klaus_Fiedler@psi-sv2.psi.uni-heidelberg.de

Universität Mannheim
L 13,15
68131 Mannheim

The Ultimate Sampling Dilemma in Experience-Based Decision Making

Klaus Fiedler,

(University of Heidelberg)

Running Head: Ultimate sampling dilemma

Author's Note: The present research was supported by various grants from the Deutsche Forschungsgemeinschaft. Thanks to Ralph Hertwig, Ulrich Hoffrage, Tobias Vogel, Peter Freytag, and Yaakov Kareev for their helpful and constructive comments on a draft of this article. Correspondence should be addressed to Klaus Fiedler, Psychologisches Institut, Universität Heidelberg, Hauptstrasse 47-51, 69117 Heidelberg, Germany.

Fax: + 49 6221 547745. Email: kf@psychologie.uni-heidelberg.de

Abstract

Computer simulations and two experiments are reported to delineate 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 three providers in two different product domains. When information sampling (from a computerized data base) was over, they had to make inferences about actual differences in the data base 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 two 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.

The Ultimate Sampling Dilemma in Experience-Based Decision Making

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 strong 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 actually 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, which 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 relevant information that happen to be available is not new. Sampling error and sampling bias are have long been

recognized as prominent topics in the methodology of the social sciences (Macrae, 1970), in epidemiology (Schlesselman, 1982), in econometrics (Manski, 1995), and in recent cognitive research in particular (Denrell, 2005; Fiedler, 2000; Fiedler & Juslin, 2006; Juslin, Winman & Hansson, in press). 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 inaccurate and 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 not only result from "nasty environments" (Hogarth, 2001) that obscure the real world or do not allow decision makers to access relevant data. Rather, the ultimate sampling dilemma emerges even in completely "benevolent environments" that render all information available and do not mislead or constrain decision makers in their information search process.

To illustrate this task setting, take the perspective of a manager, or entrepreneur, supposed to represent an "expert consumer", as it were, who is accuracy-motivated and got access to relevant data bases. 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, P_3, are available in a large database, from which the manager can draw a sample of any size. On each trial, 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 not least, he/she

may leave all three aspects unspecified and just ask for the next randomly drawn observation 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) – 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 + to – entries remains constant across all providers and product domains. Would a manager who can gather as many 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 information search consistently to a strategy that has been called natural sampling (Gigerenzer & Hoffrage, 1995). Natural sampling means to draw random events from the entire database, without ever restricting the baserates of providers, product domains, and 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 indeed conserve the properties of the universe (as in Figure 1). However, the price for such representative, unbiased, natural sampling is that information search cannot be tailored to the task at hand. When the baserate of observations about the provider of main interest, or the product domain of main interest, is very rare, focusing on that specific provider and product domain is not allowed, nor is it possible to concentrate selectively on positive (+) and negative (–) events if required in a certain problem context. 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. The information resulting from selective sampling can only be trusted for the restrictive purpose for which it was selected (e.g., a specific provider in a specific domain). Using such a purpose-constrained sample for other purposes will often lead to seriously distorted and inaccurate decisions.

Selective sampling. To anticipate an important result, hardly any decision maker engages systematically in 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, information search will focus on the rare negative outcomes. Such a selective sampling scheme will change the valence baserate, turning 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 selective search. That is, judgments of the origins of negative outcomes, like judging of p(provider P_1 / –) or p(domain C / –), will be unbiased, just as odds ratios such as p(P_1 / –) / p(P_2 / –) will be unbiased. However, 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. Depending on the proportion of trials on which search has been constrained by valence, the sample proportions of + and – outcomes reflect the decision maker's own selective search focus. As a general rule, to the extent that any variable constrains the sampling of information, the subsequent estimation of that variable is no longer unbiased. To that extent, the distribution of this variable reflects the decision maker's search strategy rather than the true environmental parameter.

Why then do people not refrain from conditional information search? Why do they not exploit the advantage of natural sampling? An apparent answer is because 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 natural design (e.g., in deficits (–) of provider P_3 in domain T; cf. Figure 1), natural sampling would require one to collect a huge number of observations until a sufficient information for the focal cell is attained. This pragmatic problem is further exacerbated when more complex designs include, for instance, the Cartesian product of valence (maybe more than 2 levels) x providers (maybe more than 3), x product domains (maybe more than 2) x recency of information (old vs. recent feedback) x prize level (high, medium, low) x validity of information source x further variables. The number of design cells can easily increase to an insurmountable number. Information about the rarest cells would be hardly accessible through natural sampling. If the decision problem (e.g., analyzing deficits of P_3 in product domain T) calls for data from particular cells, a selective sampling strategy may be necessary and adaptive, 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 impoverished evidence from rare cells, is but one problem of natural sampling. Another, equally severe problem lies in human learning, which is of course dependent on a sufficiently large sample of learning trials. Even when observation time were unrestricted and inexpensive, very rare events are likely to evade learning and memory, 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 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 trend (i.e., the preponderance of positive evaluation) will be more readily learned 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 after a long period of natural sampling from rare event classes the resulting sample is sufficiently large, learning and memory may still be impaired.

At a more fundamental level, even the very possibility of truly natural, unconditional sampling may be questioned. In reality, unlike the ideal world of statistics text books, 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, in press). A consumer will be hardly able to sample information about all products, providers, markets, and product attributes non-selectively. 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 information ecology than 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. For these and many other reasons, literally unconstrained sampling is pragmatically impossible. What makes the situation even worse is that the consumer normally has no knowledge whatsoever about the sampling constraints imposed on newspapers, TV-advertising, or the Internet. Thus, when encountering the positivity rates of providers or product domains in the media, or when assessing the relative proportion of deficient products associated with a particular provider, the consumer does not know if and to what extent the media have been sampling naturally, and by what factor the observed rates have been selectively over-sampled or under-sampled. Therefore, even when given a free choice, real decision makers can hardly 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, P_3, then one is on save 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 baserates in the population. To repeat, sample-based judgments are unbiased as long

as the dependent variable of the judgment task (proportion of + valence) has not been used for selective sampling. 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 P_1 vs. P_2, P_3, based on a sample that greatly over-represents the baserate of P_1. Having understood this simple rule, the consumer might avoid all biases and shortcomings just by tailoring the sampling process to the judgment problem at hand and by not misusing samples drawn for one purpose for another purpose.

Simple and straightforward as this solution might appear, it is hardly feasible for several reasons. First, consumers (like organisms in general) normally do not know the constraints imposed by nature on the observations to which they are exposed in reality. When hearing someone tell a bad story about a specific car, or when reading a comparative article about cars, or when drawing on his or her own memory for car experiences, one hardly ever knows to what extent the underlying sampling process was constrained. Second, even if it were known exceptionally, the constraints may change with each and every new observation. And third, the maxim to utilize for every judgment only those samples that were drawn with the independent variable in mind, and to fully ignore samples that were drawn with the dependent variable in mind, implies an untenable model of knowledge representation. 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 some logical reflection tells us that knowledge is not organized this way by sampling strategies and that it would be impossible to administrate such a multiply split, uneconomical memory. In general, 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 do not rely on prior knowledge but have unrestricted control over the sampling process and access to the entire universe.

Impact of Particular Sampling Schemes

Thus, the ultimate sampling dilemma is like sailing between skylla and charybdis. Natural sampling is potentially unbiased but very expensive, insensitive to rare events, and in reality often not feasible. Selective sampling is more feasible, especially when focusing on rare events or pursuing specific hypotheses, but the resulting sampling biases will very likely carry over to judgments of variables that somehow contributed to 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 nature of these sampling biases and their consequences for the decision process. So what will a manager do when facing the task depicted above? Granting that he or she will not refrain from active information search, what alternative search strategies might be used?

Output-bound sampling. One typical strategy is to make information search contingent on certain outcomes. An individual motivated by the goal to avoid regret and not to make mistakes could mainly look at 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 meta-cognitive insight into the sampling bias. However, as will soon be apparent, such meta-cognitive monitoring and control of sampling bias will be hardly 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., in press; 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 $p(+ / P_1 \& C) = 2/3 = .67$], estimates will follow the visible sample proportion (Fiedler, Brinkmann, Betsch & Wild, 2000; Juslin, Winman, & Hansson, in press; Kareev & Fiedler, 2006), with little attempt to correct for bias.

Thus, output-bound search by valence will lead to biases in judgments that use valence as a dependent variable, such as judgments of $p(+ / P_1 \& C)$. However, while backward judgments using valence as an independent variable, like estimates of $p(P_1 \& C / –)$ or $p(P_1 \& C / +)$, should be unbiased, they may be distorted for different reasons. 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 focus been on diagnosing origins of positive outcomes, in contrast, P_1 and C might have been most strongly associated with positive information. However, note that the latter judgment effect would reflect biases in associative memory rather than sampling biases proper.

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 x 2 cells of the design. Although an experimental design is commonly considered optimal, it is not bias-free at all (Brunswik, 1955; Dhami, Hertwig & Hoffrage, 2004; Hoffrage & Hertwig, 2006). On one hand, when information search leaves valence open and constrains providers and domains to be orthogonal, then the resulting estimates of valence (conditional or unconditional) are indeed unbiased. On the other hand, 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 have been blurred through the "experimental" scheme.

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. On one hand, pooling across product domains, it is true that the positivity rate is higher for P_3 than P_2 than P_1. On the other hand, when product domains are taken into account, it turns out that the positivity rate is markedly higher in domain C than T and that the seeming advantage of P_3 is merely due to P_3's mostly providing products from the superior domain, C. As the mediating impact of domains is partialled out, comparing providers separately within both domains, then P_3 is no longer dominant. Such a spurious correlation, 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, 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 should result in a flexible 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 environment from the sample. 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 sampling constraints.

The ultimate sampling dilemma can thus be formulated 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 to 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 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 would have to be corrected downward or upward dependent on the degree to which it has been over-sampled or under-sampled, respectively, using Bayesian statistics. However, the computational work required for such a Bayesian correction would exceed human capacity for most real problems, and the necessary statistics (baserates, likelihood ratios, conditional dependencies) are hardly ever known. For instance, to re-compute 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 x domains x valence) it was restricted. Separately for each combination of sampling constraints, the degree of over- or under-sampling would have to be calculated, and the correction algorithm would have to be a weighted average 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 be solved because in reality we seldom know the sampling constraints including all conditional dependencies for every single observation.

It may be for this reason that decision makers have evolved what might be called meta-cognitive myopia (Fiedler, 2000; Fiedler, Freytag & Unkelbach, in press; Fiedler & Wänke, 2004; Juslin, Winman & Olsson, 2000; Kareev et a., 2002; Winman & Juslin, 2006). Thus it appears as if even highly motivated judges do not care about the origin and the history of the sample data on which they base their judgments. They are often remarkably accurate relative to sample itself. But they are short-sighted, if not blind, regarding the way in which a stimulus

sample was generated. They take the sample statistics for granted, accurately but naively, as if were an unbiased snapshot of the underlying reality (cf. Fiedler et al., 2000; Juslin et al., in press; Kareev et al., 2002; Winman & Juslin, 2006).

Plan of the Present Research

In the remainder of this article, I elaborate on the interplay between inevitable sampling effects and meta-cognitive myopia. I first present a simulation study that illustrates the strength and scope of the biases resulting from different sampling strategies. Then, another section will be devoted to experimental evidence about the way human decision makers deal with the ultimate sampling dilemma. Both simulations and experiments will keep within the same task setting that was used in the introduction – buying computers and telephone equipment offered by three providers based on stored positive and negative feedback – within the ecology depicted in Figure 1.

Biases Resulting from Different Sampling Strategies: A Simulation Study

Methods and Design

To provide a systematic analysis, a simulation study was conducted. Sampling biases were studied 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(+/P1,C), p(+/P2,C), …, p(+/P3,T), 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(P1/–), p(P2/–), …, p(C/+), … p(P3,T/–). Here, the resepctive sample statistics are used to diagnose the origin (in P1, P2, P3 x C,T) of + or – outcomes.

In accordance with Figure 1 (upper chart), a data base included 480 positive instances (+) about provider P_1 in domain C; 240 instances of +, P_2, C; 120 instances of +, P_3, C and so forth. Each simulated sample consisted of n = 100 instances, drawn randomly within the

constraints imposed by the different strategies. Two sets of indicators (for causal and diagnostic inferences) were calculated for each sample.

The manipulation of sampling strategies was based on the following assumptions. 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 in all three dimensions. Thus, the decision maker may just ask for the next piece of evidence, leaving open whether the valence is + or –, 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 population, with each item in the universe having the same probability of being drawn. Alternatively, she might want to see an observation about P_1, leaving open domain and valence, or ask for a + item from the C domain, or a – item about P_3 in the T domain, or solicit any other combination of {+, – , open} x { P_1, P_2, P_3,open} x {C, T, open}. The computer makes a random draw from the restricted subset of all items in the population (e.g., from all +, P_1 items when + and P_1 are asked for).

The simulation involves different combinations of restrictions in all three dimensions. Altogether, the study uses a 7 (restrictions on providers) x 7(restrictions on domains) x 7 (restrictions on valence) design. The seven restriction levels on providers 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 (level 2-4) to 0 (level 5-7), and within these three blocks, the

enforced proportions of items drawn for the three providers change. Likewise, for the dichotomous domains and valence factors, respectively, the seven constraint levels are:

(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

Altogether, then, we simulated all 7 x 7 x 7 = 343 strategies, or combinations of constraint levels, running 100 replications per strategy and calculating two sets of indicators for each 100-item sample, corresponding to both types of judgment task:

p(+ / provider, domain): forward inferences of the likelihood of positive valence given all combinations of 3 providers and 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 is skewed in all three dimensions (cf. Figure 1): more P_1 than P_2, than P_3 data, more C than T data, and more + than – data. However, all pairwise correlations are zero; the ratio of + to – is the same (2:1) for 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 Table 1). Similarly, the correct backward inferences to the three providers, both from positive and negative valence, in both domains (see Table 2), 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 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 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 + versus – 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 + 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).

For another search strategy, decision makers might leave valence unrestricted on 40 trials and restrict the valence of the outcome only on the remaining 60 trials (e.g., gathering 75% negative outcomes when the aim is to diagnose origins of deficits). As evident from the next block in 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% +) with 60% trials that impose only 25% + (and 75% –) yields an overall positivity rate of only about 42% (Table 1), well below the original population value of 67%.

Thus, output-bound sampling, even when only applied to a subset of trials, leads to systematic 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 from the other dimensions 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 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, over-sampling P_3 cases and under-sampling 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 baserates, rather than selective attempts to induce expected or desired contingencies. It goes without saying that a strategy that looks for many + outcomes in domain C but mostly looks at – 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, I exclude these blatant cases from consideration.

Summary. Thus, simulations confirm that, by definition, natural sampling is principally unbiased but may not be feasible for different reasons, due to the paucity of infrequent event classes and the impossibility to focus selectively on the most interesting event classes. As this major disadvantage of natural sampling is avoided through 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 / –). Mixed

strategies that result from input-bound sampling on some trials and output-bound sampling or multiply constrained sampling on other trials lead to biases in both forward and backward inferences (see bottom blocks in Tables 1 and 2). Such mixed samples render the reconstruction of the original population characteristics almost impossible.

Experimental Investigation of the Ultimate Sampling Dilemma

The simulation presented so far merely corroborates what can be seen from the algebraic notation alone, but it nevertheless helped to recognize the direction and strength of the biases and to understand some of its boundary conditions. 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, human decision makers will exhibit the same sampling problems. They might engage in purely natural sampling, never constraining their search on any dimension or focussing on specific problem aspects. Assuming such a strategy, the samples informing their decisions would be unbiased. However, such 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 focussed search is that the resulting samples can only be trusted for some judgments but not for 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 intransparent environments (Fiedler, Brinkmann & Betsch, 2000; Fiedler, Walther, Freytag & Plessner, 2002; Juslin et al., in press). However, prior studies did not tackle the impact of sampling biases in situations in which information search is completely transparent and fully under the judges' control. Particularly, no prior research has addressed the ultimate sampling dilemma, that is, the trade-off between natural sampling and

selective, focussed 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 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, 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 try 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, it was expected that consistent natural sampling should be very rare. Instead, participants should tune their information search to the task focus, taking into account that (some of their) judgments reflect severe sampling biases. Moreover, due to meta-cognitive myopia (cf. Fiedler et al., 2000; Fiedler & Wänke, 2004), judges 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) of the judgment problem. Consequently, output-bound sampling (i.e., obscuring the valence baserates) should result in biased forward judgments of p(+/providers, domains). Similarly, input-bound sampling (i.e., constraining 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 baserates 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) 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 are three providers, MediaCom, EMG, and Hi-Tech (in the following denoted P_1, P_2, 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 data base. Participants were free to gather as many observations from the data base 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 origins of negative outcomes, p(Providers / –), were called for. Another manipulation, provider focus, pertained to a specific provider that had to be compared with the others. The focal provider was either P_1 (most frequent provider, see Figure 1) or P_3 (rarest provider).

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 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 task focus 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 blame the most frequently encountered provider, P_1, unless input-bound sampling focuses on another provider. Whether this occurs should depend, finally, on the provider-focus manipulation. A focus on P_1 as the provider of main interest should increase the skew of the data base, strengthening the association between P_1 and the other prevalent aspects 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, so that the major role played by P_1 is obscured in the sample.

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:

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.

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.

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.

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 instructions in the Appendix show 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 completely free in their search strategy. On every trial they could either call for an item drawn at random from the data base (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 x 3 x 2 cube was presented graphically on the screen, with the rows labelled “Computers” and “Telecommunication”, the columns labelled “MediaCom, EMG, Hi-Tech” and the foreground and background slice labelled “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 key in the upper row to select domain C or T; the G, H, 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 + or – outcome). 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 had indicated his or her constraints, the computer randomly selected one out of all items in the database that met the constraints chosen. Altogether the

database was defined by the same population distribution as in the simulation above (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, like “If one needs maintenance, nobody is available.” This item was then presented for three seconds on the screen, inserted in the cube position that corresponded to the domain, the provider, and the valence. Participants knew that they could terminate information search at any time, by pressing the Escape key.

Dependent measures. The main dependent measures were percentage inferences from the samples observations to the data base. 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, among over all information stored about this provider?" In addition to these forward inferences, they were then asked to make 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 also solicited for the origins of positive outcomes, with similar results, not reported here).

Finally, at the end of the session, the three-dimensional cube appeared again on the screen, just as during stimulus presentation, and judges were asked to estimate (in cardinal frequencies) how many observations they had sampled from each cell of the 2 x 3 x 2 scheme. Although possibly 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.

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 occurred very rarely. The proportion of trials on which the average participant engaged in 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, another one on 98% of the trials. All other participants chose natural sampling on less than 50% of their trials.

Instead, virtually everybody constrained 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 to .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 on which one specific provider was fixed was higher for a 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 backward-diagnosing instructions.

Whereas the tendency to conditionalize 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 + or – outcomes is more surprising. On average, the proportion of trials on which participants restricted the outcome (to + or –) was .718 across all conditions.

Curiously, 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 baserates in the sample beyond the correctly assessed fact that provider was rarely left unspecified (i.e., only in .40 of the trials).

Sampling biases. Given that natural sampling occurred so rarely and that almost all samples were restricted in one or more dimensions, we can next examine whether the resulting samples were biased systematically, that is, whether the rates of specific domains, providers, and evaluations in the samples deviated from the original baserates in the population. Recall that the population distribution was skewed in all three dimensions (i.e., 2:1 baserate ratios for domains and valence, and the 4:2:1 ratio for providers). This original skew was clearly reduced in the samples acquired, reflecting regression toward more equal baserates (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 baserate 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. As evident from Table 4, estimates of the observed frequencies of the 12 event combinations were clearly 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 than on diagnosing deficits. All these differences proved significant in a domains x providers x task focus x provider focus ANOVA that is not reported here to save space.

More importantly, the sample-frequency estimates allow for a first check on the major phenomena encountered in the simulation study, pertaining to biases in forward evaluations due to output-bound sampling and biases in backward diagnoses due to input-bound sampling. Because judges varied in the degree to which they solicited positive data, the consequences of output-bound sampling is evident in a significant correlation ($r = .413$, $p < .01$) between the individual proportion p*(+ / .) 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 twelve 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 ($r = .404$, $p < .01$) with the pooled estimated frequency proportion of C items in the sample.

Judgment biases. Of most interest is the question whether biases in the samples actually led to errors and biases in the eventual 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). Accordingly, the average rated percentages of positive information in the data base (across all other conditions), was higher under the former ($M = 48.56$) than the latter task focus ($M = 38.10$; cf. Table 5).

For an appropriate statistical test, an ANOVA was conducted with task focus and provider focus as between-participants factors and a contrast between ratings P_1 and average

ratings of P_2 and P_3. The predicted judgment bias was apparent in a task focus main effect, $F(1,52) = 9.33$, $p < .01$, reflecting more positive inferences when the task focussed on positive evaluation rather than diagnosing deficits. Regardless of this self-generated bias in the stimulus input, the sample was taken as observed in making population inferences, turning the sampling bias into a judgment bias. The provider focus x task focus interaction was also significant, $F(1,51) = 15.11$, $p < .001$, as 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.

< ### Correlation across Ss with output-sampling bias >

As expected, the impact of input-bound sampling biases is also 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 x task type x provider focus interaction, $F(1,52) = 7.81$, $p < .01$, as well as a two-way task type x provider focus interaction, $F(1,52) = 12.09$, $p < .01$. Altogether, these findings corroborate the assumption of sampling biases carrying over to analogous judgment biases, due to meta-cognitive myopia, that is, judges' failure 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 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 major problem merely lies in the failure to apply natural sampling. However, the ultimate sampling dilemma entails good reasons to suspect that natural sampling may also lead to inaccuracy.

A second experiment was therefore conducted, which replicated Experiment 1 in all respects, except that natural sampling was enforced. 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 rate) should increase with decreasing sample size, judgment biases should come in through the backdoor, through differential regression. The same high degree of positivity should be judged to be lower for infrequent than for frequent providers.

Methods

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. However, 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 (including all instructions and dependent measures) was used as in Experiment 1, except for changes in the information search instructions and procedures. Rather than being allowed to actively search for information, participants observed a series of observations randomly drawn from the data base, without any restrictions. Stimuli appeared at a constant rate of 3 s per observation. Information search was terminated as the participant pressed the ESC key.

Results and Discussion

Basic sample data. The average participant sampled 35.21 observations (SD = 20.90). As expected, natural sampling led to impoverished data for the less frequent combinations of

product domains x providers x valence levels. Although the distribution of observations sampled across these 12 cells (pooling over participants) closely resembled the population distribution (cf. Table 6), the average individual sample included less than 2 observations in five out of the 12 cells. Even with doubled sample size, statistical inferences from these infrequent data would be extremely unreliable. This is also evident from the large number of judges (out of 39) basing their estimates on observed frequencies smaller than, or equal to, 0, 1, or 2 (cf. Table 6). Only when the lability of individual samples is eliminated by pooling over judges does natural sampling result in an accurate picture of the population.

Sample estimates. The major asset of natural sampling is apparent, by and large, in subjective estimates of the frequencies of the 12 domains x providers x valence combinations (Table 6). Thus, at an ordinal level, the average judge correctly reported that more positive (M = 14.83) than negative observations (M = 11.16) had been sampled, that there were more data for domain C (M = 14.03) than for domain T (M = 11.95), and that the frequency of sampled data decreased from P_1 to P_2 to P_3 (M = 14.46, 13.68, 10.85, respectively). However, in spite of the average judges' conserving these ordinal differences, and the absence of a crude bias, frequency estimates were highly regressive, yielding ratios much smaller than the actual ratios of 2:1 or even 4:1; cf. Table 6). When frequency estimates were transformed into proportions to render them comparable to population proportions, large frequencies were clearly underestimated whereas small frequencies were overestimated.

Note, however, that unsystematic regression error can turn into systematic bias when the strength and direction of regression effects varies across events. This is apparent from an analysis of inaccuracy scores, that is, from the signed differences between subjective estimates and objective proportions (cf. Table 6; all transformed into proportions). These inaccuracy scores, which tend to be negative for the more frequent levels on the domains, providers, and valence factors but positive for infrequent levels (reflecting regression), were subjected to a repeated-measures ANOVA (including all 39 participants). Biases resulting

from differential regression were apparent in strong main effects on all three factors, reflecting selective underestimation (i.e., negative scores) for domain C, $F(1,38) = 39.44$, $p < .001$, for the most frequent provider P_1, $F(2, 76) = 53.76$, $p < .001$, and for positive observations, $F(1,38) = 18.35$, $p < .001$. Similarly, the same basic regression tendency produced three two-way interactions. The tendency to underestimate positive and to overestimate negative observations increases from domain T to C, $F(1,38) = 9.73$, $p < .01$, from P_3 to P_2 to P_1, $F(2, 76) = 8.67$, $p < .001$, and the inaccuracy difference between providers is 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 severely 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 dependent measure, namely, the inferences about the population, as distinguished from the estimates of sample frequencies. It was expected that biased judgments should arise from extremely unequal cell frequencies, as the same trend (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, this implies that 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 focus manipulations. Higher P_1 than P_3 judgments – both in terms of more positive forwards 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 focus was on P_1, but not when the provider focus was on P_3, M = –2.82, $t(16)$ = –.365). Thus, when the focus was consistent with the prevalence of providers, inferred positivity decreased from P_1 to P_2 to 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, though, + outcomes were also attributed to P_1).

When backward inferences were analyzed as a function of the two between-participants factors, task focus and provider focus (see Table 7), the bias score tended to be strongest when the 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 towards P_1 was also 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, the results of Experiment 2 reveal that although natural sampling produces (by definition) unbiased samples, the judgments run into new problems. Information about infrequent 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

On summary, the present inquiry into the ultimate sampling dilemma revealed what it had to reveal on a-priori grounds, namely, that there is hardly a real chance to evade this dilemma of the empirical world. To be sure, decision makers – or more generally, organisms

– may refrain from the possibility to actively search for information tailored to test specific hypotheses and passively resort to natural sampling. Ideally, this produces an unbiased sample, but one that is impoverished with regard to rare events. If decisions refer to such rare events (e.g., accidents, crimes etc.), then the information of interest may be missing or overshadowed by large amounts of unwanted or irrelevant information. Moreover, in reality, it is unclear whether natural sampling is possible at all. At any point in time or space, information about the different objects is not equally available. This restricts nature's ability to provide us with truly natural samples. Consider the Internet for a nice thought experiment. Whoever tried to search for specific objects in the Internet will agree how hard and actually impossible it is to solicit a natural, unrestricted sample that warrants “true” baserates. 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 question is whether absolute baserates exist at all – baserates that hold across time, space, cultures, markets, and decision contexts. To the extent that baserates change over time or between regions or cultures, the "true" population baserates are hard to determine. As a consequence, whether a supposedly natural sample really conserves the natural baserates cannot be controlled.

Alternatively, rather than striving for natural sampling, organisms may follow the role model of a clever 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 over-represented 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 over-represented in the sample. Finally, engaging in mixed sampling strategies will not ameliorate these problems but may even worsen the situation, causing biases in either direction.

Given a sample that entails a complex mix of search strategies, Bayesian correction is not really viable. If each of, say, 100 observations about the same provider 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 not even reveal their underlying constraints – how should the Bayesian correction of a biased sample be accomplished? For instance, how should even the best Bayesian statistician correct for a biased sample of resulting from an Internet search, if each entry in a list of entries has different constraints and, crucially, those constraints are not transparent at all?

Computer simulations served to illustrate the ultimate sampling dilemma. Although the simulations did not demonstrate something that cannot be derived through analytical reflection alone, they served to illustrate the nature and the degree of the judgment biases arising from different search strategies. Two experiments provided an empirical test of the open empirical question of how human decision makers deal with the ultimate sampling dilemma. Experiment 1 corroborated our intuition that hardly anybody engages in a pure natural-sampling strategy when information search is fully unconstrained. Rather, people tended to sample information predominantly from those cells that were relevant to testing specific hypotheses. 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 focussing on a specific provider 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. The judges' uncritical reliance on the sample given reflected meta-cognitive myopia.

Experiment 2, then, provided complementary findings for the same decision tasks when natural sampling was enforced. Although the resulting overall samples was indeed unbiased, information about infrequent event classes was insufficient. Sample-based judgments were nevertheless biased through differential regression. For instance, although positive outcomes were constantly frequent and negative outcomes constantly infrequent across all providers, this difference was more clearly 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, was most apparent where samples were most impoverished, consistent with many previous demonstrations of differential regression in confirmation-bias studies (Fiedler, 1996; Fiedler et al., 2002; Fiedler & Walther, 2003; Zuckerman, Knee, Hodgins & Miyake, 1995). In any case, natural sampling did not provide a useful remedy at all, because unreliability and regression biases came in through the back door.

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 in fact the message is not that pessimistic, and that quite a few optimistic aspects, both theoretically and practically, deserve to be pointed out. On 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. Non-human, artifical-intelligence systems would suffer from similar problems when fed with the same complexly biased input. In its most radical form, the ultimate sampling dilemma reflects the fact that “true” or normatively correct answers to many probabilistic problems actually do not exist. They are indeterminate, because there is 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 an academic career will be successful, or that the stock market will show a decline during the next five years. Whatever sampling scheme is applied will be conditionalized on a search strategy that focusses on specific sources, time frames, geographical frames, media, and on the 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, omni-directional perspective applicable to all decision problems. Inferences will be biased to the extent that the sampling process was contingent on the dependent variable to be inferred. This restriction has to be kept in mind when decisions are made in science, economy, and politics. Beyond the common practice to describe samples in terms of size and general quality or representativeness, it is essential to indicate the inherent directionality and conditionality. For instance, even a very large carefully selected clinical sample of, say, depressive patients (matched with the same number of appropriate controls), however representative it is, must not be used uncritically for ethiological (causal) inferences about the influence of genetic factors, learning, stress, or attribution on depression.

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 meta-cognitive myopia that often prevents people of all intelligence levels from recognizing the pitfalls of sampling biases. The present experiments corroborate the basic finding from many other sampling approaches that decision makers – whether lay people of 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, in press; Fiedler & Walther, 2003), selective memory (Tesser, 1978), unequal communicability (Kashima, 2000), or restricted designs (Wells & Windschitl, 1999) – to educate people in what samples are good for and to engage in corrections of biased samples wherever this is possible. In those remaining cases where corrections of biased samples are not available, the most prominent goal is to understand that ignoring a sample may be better, and more rational, than accurately utilizing a sample tailored for the wrong purpose.

Appendix

General instructions provided at the beginning of the experimental session:

"Dear participant:

Thanks for your willingness to participate in this study.

In this study, all information is transparent. That is, the goal and purpose of the study is not kept secret. You will not be distracted from the actual purpose and no deception will be involved. And we do not tray 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. Analogously, you will get access to a data base containing all stored experience with the offered products. This data base 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 correspond to the "reality" of the data base.

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 trees before the forest.

Rather than assessing and utilizing everything – which in the era of the Internet is impossible anyway – we almost always base our decisions on a s a m p l e 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 data base, enabling to you make a correct decision. Promised: the actual differences in the data base are strong enough so that a commonly chosen sample size allows you to detect these differences – provided the sample is drawn in a way appropriate to the problem. This is exactly the key to success: collecting data that are useful for the problem at hand in that they reveal the actual relations that hold in the data base.

… 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 computer 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 data base 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 data-base 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 poviders?"

Task focus = positive evaluation – Provider focus = P_3 (called XXxxx)

"MediaCom" replaced by "XXxxx", otherwise identical.

Task focus = diagnosing deficits – Provider focus = P_1 (called MediaCom)

"You are to find out whether in the total data-base 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 XXxxx)

"MediaCom" replaced by "XXxxx", 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 data base. The judgments you are supposed to make below are always meant as judgments about the entire data base 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 …?"

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*, 837-43; 849.

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

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.

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. (in press). Information ecology and the explanation of social cognition and behavior. In E.T. Higgins & A. Kruglanski (Eds.), *Handbook of basic principles.* New York: Guilford.

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

Fiedler, K., Freytag, P., & Unkelbach, C. (in press). Pseudocontingencies in a simulated classroom. *Journal of Personality and Social Psychology*.

Fiedler, K., & Juslin, P. (2006). Taking the interface between mind and environment seriously. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 3-29). New York: Cambridge University Press.

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

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

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.

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.

Hogarth, R.M. (2001). *Educating intuition*. Chicago: University Press.

Juslin, P., Winman, A., & Hansson, P. (in press). The Naïve Intuitive Statistician: A Naïve Sampling Model of Intuitive Confidence Intervals. *Psychological Review*.

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.

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.

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

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.

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.

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.

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 Constraints								
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

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 Constraints								
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

Table 3: Characteristics of spontaneously gathered samples

Task Focus: Provider Focus:	Forward P_1 Focus	Forward P_3 Focus	Backward P_1 Focus	Backward $P3_1$ Focus	Overall
Mean Sample Size (SD)	52.00 (33.76)	62.71 (41.65)	33.86 (27.24)	56.57 (38.75)	51.29 (36.43)
Natural Sampling Proportion of Trials	.06	.10	.19	.08	.11
p(Domain unspecified) p(Domain specified) p(C called for) p(T called for) p(C in sample)	.18 .82 .45 .37 .576	.20 .80 .43 .37 .559	.49 .51 .25 .26 .566	.49 .51 .28 .23 .607	.34 .66 .35 .31 .577
p(Provider unspecified) p(Provider specified)	.26 .74	.27 .73	.51 .49	.55 .45	.40 .60
p(Valence unspecified) p(Valence specified) p(+ called for) p(– called for) p(+ in sample)	.17 .83 .51 .32 .64	.44 .56 .33 .23 .63	.28 .72 .24 .48 .41	.24 .76 .15 .61 .30	.28 .72 .31 .41 .495

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	
	+	–	+	–	+	–	+	–	+	–	+	–
Population	.254	.127	.127	.063	.063	.032	.127	.063	.063	.032	.032	.016
Forward P_1 Focus	.127	.097	.077	.070	.087	.072	.097	.062	.094	.067	.091	.058
Forward P_3 Focus	.119	.060	.094	.074	.100	.069	.098	.062	.108	.057	.105	.055
Backward P_1 Focus	.059	.173	.068	.119	.061	.074	.060	.112	.063	.098	.056	.058
Backward 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	.06

Table 5. Mean Population Inferences (in Percent) By Conditions in Experiment 1

	Forward Inferences p(+ / Providers)			Backward Inferences p(Providers / –)		
	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 P_1 Focus	54.14	37.71	40.86	40.43	30.86	35.14
Forward P_3 Focus	53.71	51.57	53.36	40.93	36.29	39.86
Backward P_1 Focus	35.71	40.64	41.50	44.71	32.00	26.71
Backward P_3 Focus	39.64	36.07	35.00	24.81	33.88	40.24

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

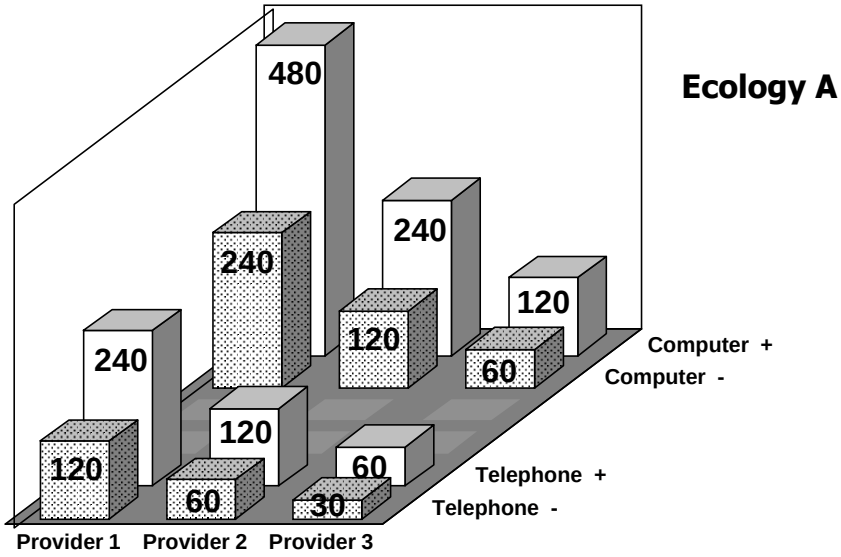
	Computers						Telecommunication					
	P_1		P_2		P_3		P_1		P_2		P_3	
	+	–	+	–	+	–	+	–	+	–	+	–
Population proportion	.254	.127	.127	.063	.063	.032	.127	.063	.063	.032	.032	.016
Sample proportion	.280	.107	.121	.065	.066	.035	.132	.057	.062	.028	.034	.014
Estimated proportion	.157	.093	.119	.079	.080	.050	.103	.072	.093	.062	.055	.038
Estimation prop. – Popolation prop.	-.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
No Ss with n=0	0	2	0	6	5	11	1	7	7	14	15	24
No Ss with n≤1	0	9	2	17	17	29	3	18	20	28	26	35
No Ss with n≤2	1	19	11	24	24	35	9	31	24	36	33	39

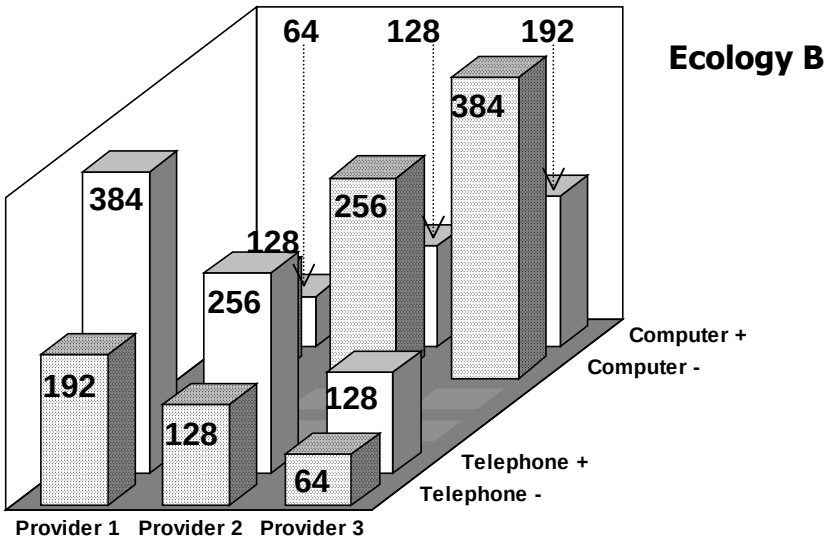
Table 7. Mean Population Inferences (in Percent) By Conditions in Experiment 2

	Forward Inferences $(+1)p(+ / P_1) + 0p(+ / P_2) + (-1)p(+P_3)$	Backward Inferences $(+1)p(P_1/–) + 0p(P_2/–) + (-1)pP_3/–)$
Population	0.00	42.86
Rating overall	6.69	3.46
Forward P_1 Focus	18.25	20.08
Forward P_3 Focus	-1.67	-1.67
Backward P_1 Focus	9.00	6.20
Backward P_3 Focus	-4.13	-19.13

Figure Captions

Figure 1: Two distinct environments to demonstrate the ultimate sampling dilemma: The skewed ecology (A) and the spurious ecology (B).





Nr.	Author	Title
07-54	Klaus Fiedler	Pseudocontingencies - A key paradigm for understanding adaptive cognition
07-53	Florian Kutzner Peter Freytag Tobias Vogel Klaus Fiedler	Base-rate neglect based on base-rates in experience-based contingency learning
07-52	Klaus Fiedler Yaakov Kareev	Implications and Ramifications of a Sample-Size Approach to Intuition
07-51	Klaus Fiedler	The Ultimate Sampling Dilemma in Experience-Based Decision Making
07-50	Jürgen Eichberger David Kelsey	Ambiguity
07-49	Tri Vi Dang	Information Acquisition in Double Auctions
07-48	Clemens Kroneberg	Wertrationalität und das Modell der Frame-Selektion
07-47	Dirk Simons Nicole Zein	Audit market segmentation and audit quality
07-46	Sina Borgsen Martin Weber	False Consensus and the Role of Ambiguity in Predictions of Othersí Risky Preferences
07-45	Martin Weber Frank Welfens	An Individual Level Analysis of the Disposition Effect: Empirical and Experimental Evidence
07-44	Martin Weber Frank Welfens	The Repurchase Behavior of Individual Investors: An Experimental Investigation
07-43	Manel Baucells Martin Weber Frank Welfens	Reference Point Formation Over Time: A Weighting Function Approach
07-42	Martin Weber Frank Welfens	How do Markets React to Fundamental Shocks? An Experimental Analysis on Underreaction and Momentum

Nr.	Author	Title
07-41	Ernst Maug Ingolf Dittmann	Lower Salaries and No Options: The Optimal Structure of Executive Pay
07-40	Ernst Maug Ingolf Dittmann Christoph Schneider	Bankers and the Performance of German Firms
07-39	Michael Ebert Nicole Zein	Wertorientierte Vergütung des Aufsichtsrats - Auswirkungen auf den Unternehmenswert
07-38	Ingolf Dittmann Ernst Maug Christoph Schneider	How Preussag became TUI: Kissing too Many Toads Can Make You a Toad
07-37	Ingolf Dittmann Ernst Maug	Valuation Biases, Error Measures, and the Conglomerate Discount
07-36	Ingolf Dittmann Ernst Maug Oliver Spalt	Executive Stock Options when Managers are Loss-Averse
07-35	Ernst Maug Kristian Rydqvist	Do Shareholders Vote Strategically? Voting Behavior, Proposal Screening, and Majority Rules
07-34	Ernst Maug Abraham Ackerman	Insider Trading Legislation and Acquisition Announcements: Do Laws Matter?
07-33	Dirk Simons	Independence, low balling and learning effects
07-32	Rainer Greifeneder Herbert Bless	Relying on accessible content versus accessibility experiences: The case of processing capacity
07-31	Rainer Greifeneder Herbert Bless	Depression and reliance on ease-of-retrieval experiences
07-30	Florian Heiss Axel Börsch-Supan Michael Hurd David Wise	Pathways to Disability: Predicting Health Trajectories
07-29	Axel Börsch-Supan Alexander Ludwig Mathias Sommer	Aging and Asset Prices

Nr.	Author	Title
07-28	Axel Börsch-Supan	GLOBAL AGING - Issues, Answers, More Questions
07-27	Axel Börsch-Supan	MIND THE GAP: THE EFFECTIVENESS OF INCENTIVES TO BOOST RETIREMENT SAVING IN EUROPE
07-26	Axel Börsch-Supan	Labor market effects of population aging
07-25	Axel Börsch-Supan	Rational Pension Reform
07-24	Axel Börsch-Supan	European welfare state regimes and their generosity towards the elderly
07-23	Axel Börsch-Supan	Work Disability, Health, and Incentive Effects
07-22	Tobias Greitemeyer Rainer Greifeneder	Why the Euro looked like a price booster: Differential perception of increasing versus decreasing prices
07-21	Patrick A. Müller Rainer Greifeneder Dagmar Stahlberg Herbert Bless	Relying on accessibility experiences in procedural fairness judgments
07-20	Volker Stocké	The Motive for Status Maintenance and Inequality in Educational Decisions. Which of the Parents Defines the Reference Point?
07-19	Jürgen Eichberger David Kelsey Burkhard Schipper	Ambiguity and Social Interaction
07-18	Jürgen Eichberger Willy Spanjers	Liquidity and Ambiguity: Banks or Asset Markets?
07-17	Patrick A. Müller Jana Janßen Dominique Jarzina	Applicantsí reactions to selection procedures ñ Prediction uncertainty as a moderator of the relationship between procedural fairness and organizational attractiveness

Nr.	Author	Title
07-16	Patrick A. Müller Dagmar Stahlberg	The Role of Surprise in Hindsight Bias ñ A Metacognitive Model of Reduced and Reversed Hindsight Bias
07-15	Axel Börsch-Supan Anette Reil-Held Daniel Schunk	Das Sparverhalten deutscher Haushalte: Erste Erfahrungen mit der Riester-Rente
07-14	Axel Börsch-Supan Dirk Krüger Alexander Ludwig	Demographic Change, Relative Factor Prices, International Capital Flows, and their Differential Effects on the Welfare of Generations
07-13	Melanie Lührmann	Consumer Expenditures and Home Production at Retirement: New Evidence from Germany
07-12	Axel Börsch-Supan Anette Reil-Held Christina Wilke	Zur Sozialversicherungsfreiheit der Entgeltumwandlung
07-11	Alexander Ludwig Dirk Krüger	On the Consequences of Demographic Change for Rates of Returns to Capital, and the Distribution of Wealth and Welfare
07-10	Daniel Schunk	What Determines the Saving Behavior of German Households? An Examination of Saving Motives and Saving Decisions
07-09	Axel Börsch-Supan Anette Reil-Held Christina Wilke	How an Unfunded Pension System looks like Defined Benefits but works like Defined Contributions: The German Pension Reform
07-08	Daniel Schunk	The German SAVE survey: documentation and methodology
07-07	Hans-Martin von Gaudecker Carsten Weber	Mandatory unisex policies and annuity pricing: quasi-experimental evidence from Germany
07-06	Daniel Schunk	A Markov Chain Monte Carlo Multiple Imputation Procedure for Dealing with Item Nonresponse in the German SAVE Survey
07-05	Hans-Martin von Gaudecker Rembrandt Scholz	Lifetime Earnings and Life Expectancy

Nr.	Author	Title
07-04	Christopher Koch Daniel Schunk	The Case for Limited Auditor Liability - The Effects of Liability Size on Risk Aversion and Ambiguity Aversion
07-03	Siegfried K. Berninghaus Werner Gueth M. Vittoria Levati Jianying Qiu	Satisficing in sales competition: experimental evidence
07-02	Jannis Bischof Michael Ebert	Inconsistent measurement and disclosure of non-contingent financial derivatives under IFRS: A behavioral perspective
07-01	Jörg Oechssler Carsten Schmidt Wendelin Schnedler	Asset Bubbles without Dividends - An Experiment
06-16	Siegfried K. Berninghaus Hans Haller	Pairwise Interaction on Random Graphs
06-15	Markus Glaser Philipp Schmitz	Privatanleger am Optionsscheinmarkt
06-14	Daniel Houser Daniel Schunk Joachim Winter	Trust Games Measure Trust
06-13	Markus Glaser Sebastian Müller	Der Diversification Discount in Deutschland: Existiert ein Bewertungsabschlag für diversifizierte Unternehmen?
06-12	Philipp Schmitz Markus Glaser Martin Weber	Individual Investor Sentiment and Stock Returns - What Do We Learn from Warrant Traders?
06-11	Siegfried K. Berninghaus Sven Fischer Werner Gueth	Do Social Networks Inspire Employment? - An Experimental Analysis -
06-10	Christopher Koch Carsten Schmidt	Disclosing Conflict of Interest - Does Experience and Reputation Matter?