Smart Sampling without Reinforcement

Stefani Nellen (snellen@andrew.cmu.edu)
Marsha C. Lovett (lovett@cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Ave
Pittsburgh, PA 15213 USA

Abstract

We conducted an experiment to compare the effect of natural frequencies and active, trial-by-trial sampling on participants' estimation of proportions in a population. The two information formats evoked similar degrees of accuracy, sensitivity to subtle statistical properties of the environment, and sensitivity to the estimates' accuracy. However, in some situations participants who didn't have to sample information showed a greater desire for this information, suggesting profound differences between the two information formats.

Background and Objective

Consider two explorers, each on their own island. They don't know of each other, but they have identical objectives: To determine how many of the birds on their island can both fly and swim under water for extended periods. Let's call such birds "swimming birds". These explorers don't want to destroy the rich wildlife around them, and so have to rely on their own observations of birds for making their final judgments.

Assume one of the explorers happens to be on an island that hosts only a small proportion of swimming birds, the true proportion being around 20%, or p=0.2. The other one is on an island where about half of the birds are also swimming birds, p=0.5. Both don't want to submit their reports until they can be sure they are reasonably accurate. Which explorer should collect more evidence? Which explorer *should* collect more evidence?

An intuitive answer to that question would be to say they both should collect the same amount of evidence, and as much as reasonably possible. After all, the bigger the sample-size, the more reliable and accurate the information from that sample will be. However, our explorers are interested in proportions, more specifically one "target proportion" (swimming birds). In the case of binary variables (such as swimming birds vs. other birds, a binomial 1/0 notation) that form the basis of estimates of proportions, the variance associated with the estimate of a proportion is a function of the estimate itself, and it increases monotonically as the proportion becomes closer to 0.5. Consider two binomial distributions A and B, each corresponding to a set of observations in which there is a certain proportion of "positive instances", e.g. successes, or swimming birds. Let this proportion be .5 for distribution A and .2 in distribution B. Even if there is an equal number of observations underlying the two distributions, distribution A will have a greater variance than distribution B. Because of this, the explorer on the island where half of the birds are swimming birds should be collecting more evidence than his colleague on the other island, at least if we assume that they both aspire for a similar level of accuracy. This is a relatively subtle statistical characteristic, and one objective of this research is to find out whether people are sensitive to this kind of feature in their environment.

In a broader context, the "explorers scenario" represents a familiar situation: a situation in which we have to estimate proportions on the basis of information we encounter sequentially. The format in which such sequentially encountered information can be described is called "Natural Frequencies" (Kleiter, 1994), and the argument has been made (Gigerenzer & Hoffrage, 1995; Kurzenhaeuser & Hoffrage, 2002) that information that is presented in a Natural Frequencies format is easier to process and therefore leads to more accurate quantitative judgments than information that is presented in a more "formal" statistical probabilistic format. For instance, judgments based on information of the form "2 out of 10 birds are swimming birds" are on average more accurate than judgment based information of the form "the probability of a bird being a swimming bird is 0.2", even though the two are equivalent.

The beneficial effect of a natural frequencies format on probabilistic inference, reasoning, and information search tasks is well documented by now, (Gigerenzer & Hoffrage, 1995; Hertwig & Gigerenzer, 1999; Oaksford & Chater, 2003). However, the comparison between the frequency format and the probabilistic notation was usually made by giving participants descriptions of the same information in the two different formats, as in the example given above. The effect of having to sample information trial by trial has not been directly compared with that of "just reading" information in a frequency format.

The effect of accumulating information trial by trial (let's call it "active sampling" for short) on judgments and decisions has, however, been investigated separately, and it has been found to reduce judgment biases such as the "illusion of control" or base-rate neglect (see Koehler, 1996, for an overview). Again, the "competitor", in terms of information format, has been the probabilistic format.

Natural Frequencies and active sampling are, obviously, related. In fact, the beneficial effect of frequency formats has often been explained by the pointing out that, among other things, they match the way in which we encounter information as we move through life). It has also been pointed out (Kleiter, 1994) that Natural Frequencies communicate reliable statistical information about the environment that can be the basis of point estimates (such as

a proportion) a well as second order statistics (such as the variance around the estimate of a proportion), and this true both if they are encountered sequentially and if they are presented in the form of a description. However, there are also differences between the trial-by-trial information search situation and reasoning with natural frequencies. For instance, trial-by-trial information search is more effortful than reading summary information of a sheet of paper, and explicit processes, such as computing proportions, are less likely to occur.

In this paper, we therefore attempt a comparison between two good things ("good" in terms of being an information source, and relative to the probabilistic format): Presenting information about a sample in natural frequency format on the one hand, and having participants sample information about the same sample trial by trial.

The particular task in which we apply this manipulation is essentially the explorer scenario mentioned in the introduction: We asked participants to estimate proportions within a population "with reasonable accuracy". The accuracy of these estimates aside, we were particularly interested in seeing whether participants in the two situations are sensitive to the subtle statistical property inherent in this type of estimation: the monotonic increase of the variance around an estimate of a proportion as this proportion approximates .5. In the trial-by-trial sampling situation, this sensitivity would be indicated if the number of instances participants sample before making their estimate was an inverted U-shaped function of the "target" proportion. In the situation in which information about a sample is presented in a frequency format (to return to the explorer scenario, an example would be: "2 swimming birds and 3 non-swimming birds have been encountered so far") such a sensitivity would be indicated if, given the same sample size, participants were more likely to judge the sample size as insufficient for making a good estimate if the target proportion (of swimming birds) is close or equal to .5.

The two situations we are comparing correspond to two classes of situations that often occur outside of the lab: Starting an active search for more information from scratch (and, at some point, deciding to stop), and deciding on the basis of already existing information whether or not more evidence is needed. The first class of situations might be encountered by a biologist who wants to find out the proportion of different species there are on an island by exploring the island, or by sociology research assistant who wants to find out about the proportion of women in a field of study by walking down the hallway of the department, or by an ordinary person who has just had a shocking breakfast experience and now wants to find out the degree of infestation by cockroaches in his cereal by taking spoonfuls of content out of the box. The second class of situations might be encountered by someone who has to judge the sufficiency of the information collected by the first person: The publisher of the Journal of Rare Species, who reads the explorers research paper; the Campus Diversity Initiative Intern who reads the observers report; the person who answers the cereal company's complaints hotline.

In summary, these are the research questions we seek to address:

1.) Does the overall accuracy of estimates differ between participants that are receiving information in a natural frequency format and participants that have to sample this information trial by trial?
2.) Do people in both or either of these two situations show sensitivity to the "target proportion" they wish to find out? In other words, do they sample more information if that proportion is equal or close to 0.5?
3.) Do people in both or either of these two situations show sensitivity to the accuracy of their own estimates? In other words, do people who are actively sampling information stop when their estimates have reached a specific level of accuracy, regardless of the true proportion in the sample? And, in the second situation, does the probability that people feel a sample is sufficient for making a good estimate decrease systematically as the accuracy of their estimates increases?

Method

Participants

100 CMU undergraduate students participated in the study in exchange for credit counting towards the fulfillment of their research requirements. Participants were randomly assigned to either a "sampling group" (N=51) or a "numbers group" (N=49). We provide a detailed definition of these terms below.

Procedure and Stimuli

Participants in both the "sampling group" and the "numbers group" worked on the computer, where they were presented with, and read, the instruction text and then made responses by clicking virtual buttons with the mouse.

Sampling Group

In the "sampling group", participants were assigned the goal of estimating a proportion within a population with reasonable accuracy. They were allowed to sample as many instances from that population as they wished before making the estimate.

More specifically, participants were asked to look through a virtual stack of photographs on the computer. All photos showed animals. Participants had to estimate the proportion of photos in this stack that showed fish. Each participant repeated this for nine photo-stacks. The procedure was the same each time. Participants were able to look through the current stack by selecting a "next" button with the mouse. Whenever they did so, a new photo from the stack, showing either a fish or another animal, was displayed in the middle of the computer screen. (These photos were drawn from a large database, so a different photo was displayed each time). At each trial, participants either decided to look at

another photo or to give their estimate and thereby end the sampling process. If they wanted to look at another photo, the previous photo was removed from the display and a new photo was displayed. Participants couldn't see how many photos they had already looked at. They could theoretically know the number of photos they had inspected by counting them. We wanted to prevent this behavior in the sampling condition, because we were primarily interested in the effect of the sampling process itself, and less in the effect of mathematical or counting strategies. We instructed participants not to count, and post-experiment interviews indicated that participants complied with these instructions. Moreover, a counting strategy would have obliterated any sensitivity to the target proportion, and the fact that we discovered this sensitivity lends additional credibility to participants' self-reported compliance.

Each photo-stack was associated with an underlying probability of sampling a fish-photo. These probabilities varied between .1 and .9 with steps of .1 between them, and the order of probabilities was randomized for each participants. To relate the underlying, or "true" fish-probabilities to participants' actual experiences, our experimental software generated a random number between 0 and 1 at each trial. If the number was lower than the current "true" probability, a fish-photo was shown. Therefore, the fish-proportions in the actual sequences of photos encountered by participants could actually deviate slightly from the underlying proportions; a fact we have taken into account in our analyses.

Participants were giving their estimates by using a slider they could drag with the mouse on a scale with 5 categories: 0-20%, 20-40%, 40-60%, 60-80% and 80-100%. The entire procedure (i.e. Reading the instructions, sampling through all nine photo stacks and the debriefing) took approximately 10 minutes for all participants.

Numbers Group

In the "numbers group", participants also had to estimate the proportion of fish-photos in a series of virtual photo stacks. However, instead of sampling photos themselves, they were given summary-information about a sample from the current stack. More specifically, they were told how many photos with fish and other animals had been found in the stack so far. Then, they were asked to indicate whether they thought that this information was sufficient for making a reasonably accurate estimate. After having given this assessment (by selecting either a button saying "Yes, I know enough" or one saying "No, I need to see more"), they made an estimate using the same scale as the participants in the numbers group. Two factors varied between stack-summaries: Sample Size (i.e. the sum of photos that had "already been found") and the true proportion of fish-photos. The proportions again varied between .1 and .9 with intervals of .1. We wanted to generate a reasonable range of sample sizes, with a preference for smaller over larger samples, since we expected the shift between "insufficient" and "enough" to be located in a small to medium sample size. To this end, we took the sum of 3 plus (1,…,8) to the power of 1.88 and rounded the result downwards. This gave us the following set of sample sizes: 4, 6, 10, 16, 23, 32, 41, 52, and a total of 72 different stacks per participant (9 proportions * 8 sample sizes). The order of the stacks was randomized for each participant. Again, we generated the information about the sample sat each trial. A number of photos equal to the sample size was generated, where each individual photo had a probability equal to the true probability of being a fish-photo. Basically, this was the same procedure as in the sampling group, but used to generate an entire sample of a pre-specified size instead of one experience after the other.

Results

Before reporting the results, we will explain how we defined several variables of interest.

Two dependent measures that are important for exploring participants' sensitivity to the target proportion in both groups are:

"*Extent of Sampling*" (sampling group): average number of photos participants chose to look at before making their estimate.

"*Desire for information*" (Numbers Group): Overall proportion of trials at which participants decided that they would need more information to make a reasonable estimate.

"*Desire for more Information-Standardized"* (Numbers Group): proportion of trials corresponding to a certain target proportion at which the participant indicated that s/he wanted "more" information, computed for each participant individually.

As an independent measure in these analyses we used

"experienced proportion" (both groups): Participants' actual experience, determined by the generative processes described above. For clarity, we binned the measure into bins of .1 in the figures and in some analyses, but in reality it is continuous.

We also consider two measures of accuracy, the former as a dependent measure and the latter as both a dependent and an independent measure:

"*%correct*" denotes the proportion of trials at which the "true" proportion is within the category chosen by the participant at that trial. For instance, if the true proportion is .3, selecting the bin 20-40% would count as correct, or "1" and all other would count as "0". In the case of proportions that are at the boundaries between the available categories, such as .2, both neighboring categories count as a correct estimate. In the case of .2, this would be the categories 20-40% and 0-20%.

"*p(correct)*" denotes the probability that the choice participants have made, given the instances they have seen so far, is actually correct. This measure is independent of the true probability and views p(correct) from the perspective of the participant. We can calculate this

probability using the beta function and the current evidence as follows:

$$p(\text{correct}) = \text{beta}(\square, \square, p_{upper}) - \text{beta}(\square, \square, p_{lower})$$

Here, beta is the cumulative beta function, $\square$ is the number of fish-photos + 1, $\square$ is the number of non-fish-photos + 1, p_{upper} is the upper bound of the bin the participant chooses, and p_{lower} is the lower bound of that category.
An alpha-level of .01 was used for all statistical tests.

Overall accuracy

How accurate are participants' estimates in both groups?
Figure 1 shows the overall accuracy in both groups. Since it is reasonable to assume that accuracy increases with sample size, we plotted accuracy as a function of sample size. Overall, participants' accuracy was rather high, around 70% in both groups. There seems to be no effect of sample size on accuracy in either group. This impression was confirmed by the result of a logistic regression of accuracy on sample size as a predictor, which was performed for both groups. In the sampling group, there was no significant increase of accuracy with sample size (p = .714*)*. The same was true for the numbers group (p = .481). Because the sample sizes were pre-defined in the numbers group and essentially self-defined in the sampling group, a direct statistical comparison between the two groups is not appropriate. However, eyeballing the data makes it seem unlikely that accuracy as a function of sampling size differs significantly between the groups. However, this issue awaits further investigation.
We already knew that active sampling and natural frequencies formats lead to a better understanding of quantitative information than probabilistic formats. Now we have evidence that their benefits might be comparable, at least using as coarse a measure of accuracy as "%correct". In the following sections, we will investigate how the two information formats compare on the level of more detailed processes.

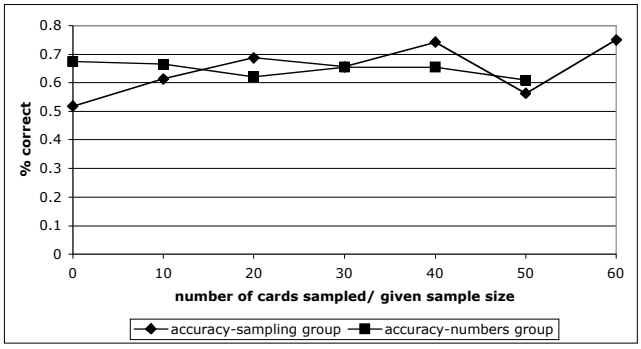

Figure 1: Extent of overall sampling as a function of current sample size for both sampling and numbers group.

Sensitivity to target proportion

An important objective of this research was to see whether participants under either or both of the two conditions investigated here would be sensitive to the variance associated with their current target proportion and adjust the extent of their sampling, or the desire for more information, accordingly. If this were the case, both measures would be an inverted U-shaped function of the experienced proportion.
Figure 2 shows the extent of sampling in the sampling group as a function of experienced proportion. Note that, in this analysis, we computed how much of their total sample each participant devoted to stacks with each of the 11 experienced proportions. In other words, we analyzed the extent of sampling evoked by each proportion relative to each individual participants total sampling size, controlling for inter-individual differences in the overall extent of sampling.
We see a clear effect of the target proportion: middle proportion lead to larger samples. There is one interesting exception to this pattern: A "dip" at the point where the experienced proportion is very close to .5. This seems to be at odds with normative prescriptions, because variance increases monotonically until .5. There are three possible reasons for the "dip": a) it might be the result of random noise (note that it is absent from the corresponding curve in the sampling group, Fig. 3), b) participants might use by the familiar sensation of a "fifty-fifty" situation as a sign that they "know enough", c) the fact that .5 is in the *middle* of the answer category "40-60%" facilitates the estimation process to some degree.
We conducted an ANOVA with experienced proportion as factor and found its effect on the standardized extent of sampling significant (F(10, 430) = 5.24, p <.001. Moreover, regression analysis revealed a significant quadratic trend (F(10, 430) = 18.4, p <.001. The best fitting linear model did not fit the data well, F(10, 430) = .26, p = .61.

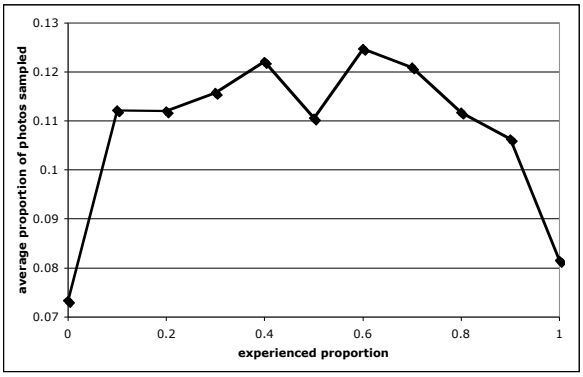

Figure 2: Extent of sampling (standardized) as a function of experienced proportion (sampling group)

Figure 3 shows the "Desire for more information" as a function of the experienced proportion in the numbers group. The two most extreme data-points (corresponding to experienced proportions of 0 and 1, respectively, are strikingly different from the corresponding ones in the sampling group: Whereas participants in the sampling group sample very little information when they experience these extreme proportions, participants who are presented with

them all at once apparently feel that they are lacking some crucial piece of information. This does not necessarily mean that they are insensitive to the statistical information in the sample: from the behavior of the participants in the sampling group we can conclude that a few experiences would have been enough to satisfy these participants. Indeed, statistical analyses confirm the intuition that these participants are just as sensitive to the changes in variance as the ones in the sampling group, with the notable exception of the two most extreme proportions:

An ANOVA in which *all* levels of the experienced proportions are included shows that there are significant differences in the standardized desire of information measure, depending on the current proportion ($F(10, 506) = 3.805$, $p < .001$. However, there is no significant quadratic trend in the data (best fitting curve: $F(10, 506) = .85$, $p = .456$).

However, if we exclude the two most extreme proportions, the overall effect of experienced proportion remains significant ($F(8, 414) = 3.575$, $p = .001$), and we can identify a significant quadratic trend ($F(10, 506) = 10.61$, $p = .000$). A linear model did not show a good fit to these data.

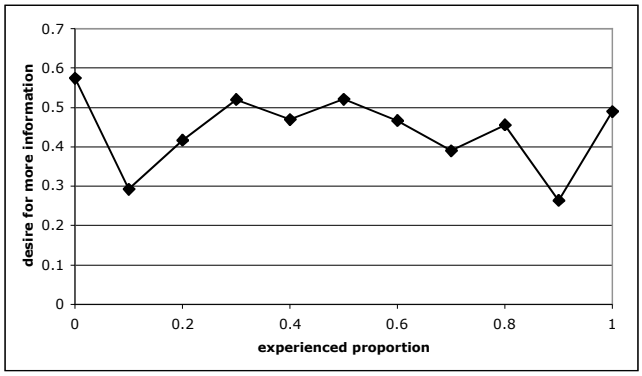

Figure 3: Desire for more information as a function of experienced proportion (numbers group)

Sensitivity to Accuracy

We have seen that participants in both groups do, on average, sample/desire more information when the target proportion is around 0.5. Can it be argued that they do this in order to achieve a comparable subjective accuracy for all estimates?

One indication of such a sensitivity to accuracy would be if p(correct) remained constant for all proportions in both groups, especially for those estimates that are indeed correct. Since p(correct) is a measure of accuracy from the participants view, the finding that it remains constant across proportions would suggest that a single criterion is underlying participants' subjective decisions to stop sampling.

Figure 4 shows p(correct) as a function of experienced proportions in both groups. We split the data into incorrect and correct estimates. Trivially, p(correct) is higher for the correct estimates. It also appears that, for the correct estimates, p(correct) is quite constant across proportions, in the sampling as well as in the numbers group. Exceptions are the two most extreme proportions, .1 and .9. where p(correct) is highest. Note that this is a demonstration of the lower variance associated with extreme proportions: Even though participants in the sampling group sample little information when the experienced proportions are extreme, their probability of being correct is still quite high. Note also that, given the overall rather high accuracy, "being correct" is actually the more frequent situation in this study.

For both groups, we analyzed p(correct) as a function of experienced proportion and accuracy in a 2-way ANOVA. There was a significant interaction between the two factors in both groups, confirming the difference in the patterns (sampling: $F(10, 419) = 4.248$, $p = .000$; numbers: $F(10, 3434) = 118$, $p = .000$). To investigate whether p(correct0 is indeed "flat" across proportions, we excluded incorrect estimates and cases in which the experienced proportion was smaller than .1 and greater than .9 and conducted an ANOVA with proportion as factor and p(correct0 as dependent measure. In the sampling group, the differences in p(correct) between proportions were no longer significant ($F(6, 306) = .682$, $p = .664$). In the numbers group, the effect of proportion on p(correct) remained significant ($F(6, 2546) = 24.45$, $p =<.001$). Whether this is a systematic effect or an artifact of the extreme sample size is an issue that can't be decided upon on the basis of the current evidence.

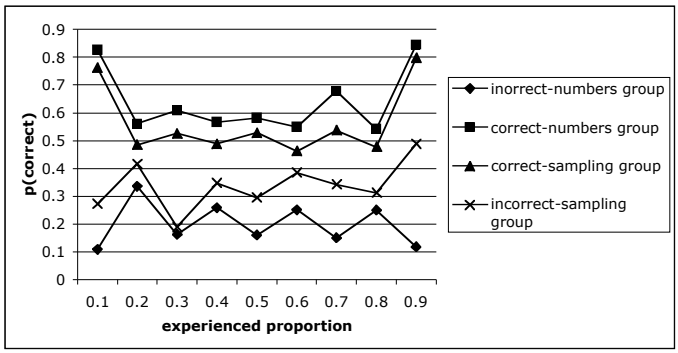

Figure 4: P(correct) as a unction of experienced proportion (both groups)

Participants in the numbers group do not produce samples but judgments in a binary format (do / do not want more information). This data format allows us to conduct another interesting analysis that is related to participants' sensitivity to accuracy: We can plot the probability that participants want more information as a function of p(correct) to see whether they are sensitive to differences in the reliability of their estimates. If their tendency to request information decreases systematically as p(correct) increases, this is an indicator for that sensitivity – and again, this should be mostly true for correct estimates. In Figure 5, we show the results of this analysis, plotted separately for correct and incorrect estimates. We see that there is a negative linear relationship between p(correct) and the probability of wanting more information, such that p(wanting more) is roughly 1-p(correct) + 0.1. In the case of the incorrect

estimates, there seems to be no relationship between p(correct) and the propensity to desire more information. Again, this is a systematic pattern, which suggests that, on average, participants are sensitive to the reliability of their estimates when they are indicating whether or not they need more information.

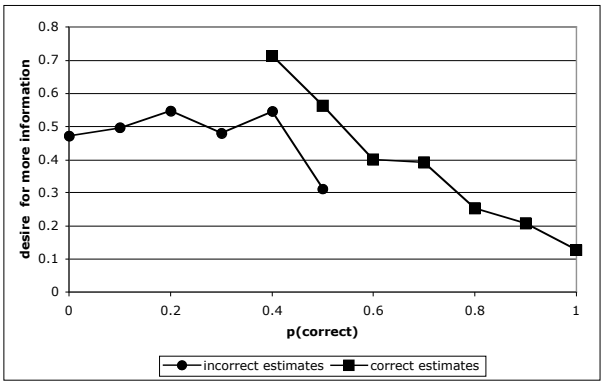

Figure 5: Desire for more information as a function of p(correct) for correct and incorrect estimates in the numbers group.

Discussion

So how do the active sampling paradigm and the Natural Frequency (or "numbers") Paradigm compare? They appear to evoke a similar level of accuracy in participants' estimates, which might be related to the fact that this talk wasn't insurmountably hard. However, we find it mildly surprising that the accuracy in the sampling group was relatively similar to that in the numbers group, given that participants in this condition possessed no external records of their experiences and were furthermore "forbidden" to count instances. The fact that even small samples led to relatively high levels of accuracy demonstrates the value and reliability of sampled information.

A striking difference between the group was the response to extreme proportions, such as 0 or 1. Participants in the numbers group seem to have felt compelled to ask for information in these cases (note that this was true across all sample sizes, not just for small samples), while participants in the sampling groups collected only very small samples when they experienced these extreme proportions. This shows two things: Participants in the sampling group were probably sensitive to the smaller variance of these extreme proportions, and sampling information is not always comparable to seeing that same information on a page. Participants in the numbers group probably felt as if they were missing information if one of the two categories (fish or other animals) was empty. Had they had the chance to sample for themselves, they would probably have stopped quite quickly. This supports the notion that active sampling is reducing biases, or, to put it more neutrally, certain reasoning patterns that might or might not be reasonable in a given situation. One possible mechanism behind this could be that the actual physical effort of collecting information provides people with a better clue about the "size" of their information and its reliability. It is also effortful to collect information yourself: information that is likely to be useful is more desirable under these circumstances. Participants in the numbers group only had to indicate whether or not they wanted more information, but they didn't have to collect it themselves. As a result, they might have had a lower threshold for demanding new information, a phenomenon that might well be true in all kinds of situations in which we are given information we haven't collected ourselves.

There are two issues that need to be addressed in future research. The first of these is the influence of statistical knowledge. In the present study, we have assumed that our participants were unaware of the fact that variance around proportion estimates increases as these approach .5, and post-experiment interviews revealed no reason to doubt this assumption. However, in future studies, this knowledge should be controlled for explicitly, for instance by administering a post-test. Another interesting issue is the effect of the response format on the sampling process. In this experiment we provided participants with 5 pre-defined bins. It would be interesting to see whether these findings generalize to situations in which there are more or fewer bins. We predict that the sensitivity would remain, but the overall sample size would increase with the number of available bins. In conditions identical to or similar to our sampling group, we expect to find the same inverted U-shaped relationship between sample size and experienced proportion, but with bigger sample sizes overall. In the numbers group, however, we expect the curve to become flatter, since more bins require more precision, which should be mirrored by a greater probability to ask for more information across all experienced proportions.

References

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

Hertwig, R. , & Gigerenzer, G. (1999). The "conjunction fallcy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12,* 272 – 305.

Kleiter, D.G. (1994). *Natural Sampling: rationality without base rates.* In G.H. Fischer & D. Laming (Eds.), Contributions to mathematical psychology, psychometrics, and methodology. New York: Springer.

Koehler, J.J. (1996). The base-rate fallacy reconsidered: descriptive, normative and methodological challenges. *Behavioral and Brain Sciences, 19,* 1 – 53.

Kurzenhaeuser, S. & Hoffrage, U. (2002). Teaching Bayesian Reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher, 24,* 531 – 536.

Oaksford, M. & Chater, N. (2003). Optimal data selection: Revision, Review, and Re-evaluation. *Pychonomic Bulletin and Review, 10,* 289 – 318.