The distorting effect of deciding to stop sampling

Anna Coenen and Todd M. Gureckis
Department of Psychology, NYU, 6 Washington Place, New York, NY 10003 USA
{anna.coenen, todd.gureckis}@nyu.edu

Abstract

People usually collect information to serve specific goals and often end up with samples that are unrepresentative of the underlying population. This can introduce biases on later judgments that generalize from these samples. Here we show that goals influence not only *what* information we collect, but also when we decide to terminate search. Using an optimal stopping analysis, we demonstrate that even when learners have no control over the *content* of a sample (i.e., natural sampling), the simple decision of when to stop sampling can yield sample distributions that are non-representative and could potentially bias future decision making. We test the prediction of these theoretical analyses with two behavioral experiments.

Keywords: information search; stopping rules; sampling; decision-making

Introduction

Information search typically improves judgment and decision making by reducing uncertainty about the the world. Imagine trying to research what qualities make a good business manager prior to making a hiring decision. To achieve this goal, it might make sense to look at the characteristics of CEOs of successful companies and find out what they have in common. In this case, your research seems to add information, improving your ability to decide. However, no matter how much data you collect, it would be a fallacy (and one that people often commit) to make the inverse inference that managers with those characteristics will necessarily lead companies to success (Denrell, 2003). In this case, the way in which information is collected can strongly bias judgment and decision making.

This scenario points to a common problem we encounter when sampling the world. In most cases, the information we obtain is not random, but rather tailored to our specific goals. Biases can arise when the information we collect with one goal or objective in mind is used to make later decisions pertaining to different goals. To give another example, imagine that we had sampled more broadly CEOs from both successful and unsuccessful companies to help answer the question what makes companies successful. Even though that broader sample is now more suitable to answer that question, it would still be unhelpful for answering other questions, like whether college dropouts do better than graduates in their subsequent careers. In this paper, we examine the relationship between search goals, information sampling decisions, and prediction.

Selective versus natural sampling

Examples of biased sampling such as those reviewed above typically involve *selective sampling*, when learners have control over *what* they sample (e.g., only successful managers). In addition to the example of conditional reasoning, selective sampling biases have also been shown when people have to trade-off reward and information (e.g. Denrell & March, 2001; Rich & Gureckis, 2014) or when feedback is asymmetric across different choice options (Le Mens & Denrell, 2011).

Unlike selective sampling, *natural sampling* refers to the process of drawing samples directly from the generating distribution without conditioning queries on any particular aspects of the sample. For example, a natural sample of businesses could be obtained by randomly selecting a number of companies that were founded in a specific year, irrespective of their subsequent success. As noted above, natural sampling is often considered a remedy for biases introduced by selective sampling because it enables learners to "conserve the properties of the universe" (Fiedler, 2008), such that the distribution of the sample will mirror the distribution of the source in an unbiased fashion.

However, even during natural sampling learners can often make the decision when to *terminate search* which may in turn be influenced by their goals (see e.g., Juni et al., in press; Lee & Paradowski, 2007; Vul et al., 2014). Here, we will argue that optimal goal-induced stopping decisions can lead to potentially non-representative samples even when they are generated by natural sampling. In particular, using an optimal stopping analysis we show that goals can have a powerful impact not only on the size of samples that learners collect, but also on the content and distributional characteristics of those samples. We then show how such samples may not reflect the statistical properties of the original distributions and how they could later on produce biased decisions in a naive learner. We finally examine these model predictions with two behavioral experiments.

An optimal stopping analysis for natural sampling of binary outcomes

To demonstrate the impact of stopping on sample composition, consider a simple information search task in which learners repeatedly observe binary outcomes from a distribution of interest. For example, learners could be picking either good or bad apples in order to learn something about the quality of the tree. The tree will have some probability of yielding good and bad apples, but learners have to rely on a finite sample of apples to estimate this probability.

Such a binary task can be modeled as a Bernoulli process (a coin flip essentially) with two possible outcomes (*h*eads or *t*ails) with the outcome probability θ (probability of heads). Assume that the learner incurs a small cost c for every draw of the distribution (every coin flip). Let's also assume that the learner will subsequently have to answer one of the following two questions (sampling goals).

1. *Binary*: Find out if θ is greater than 0.5 ("Is the coin biased towards heads or tails?").

2. *Estimation*: Find out the value of θ ("What is the bias of the coin?").

Each goal is associated with a different reward that is a function of the true value of θ and the participant's estimate. The learner's task is to decide when to stop sampling and provide their estimate given this cost function, the current sample (heads and tails), and the sampling cost. Assuming that there exists a maximum number of samples that learners are allowed to draw this can be framed as an finite-horizon *optimal stopping problem*. At every possible state (defined by the size of the current sample, n, and number of heads in that sample, h_n), an optimal decision maker should compare the expected value of stopping and of continuing and choose whichever is higher. Thus the expected value of a state, given a horizon of a maximum of T flips is

$$V_n^{(T)}(n,h_n) = max\{V_{stop}, V_{cont}\} \quad (1)$$

where

$$V_{stop} = u_{stop}(n,h_n) - nc \quad (2)$$

and

$$V_{cont} = E[V_{n+1}^{(T)}(n+1,h_{n+1})] \quad (3)$$

and where $u_{stop}(.)$ is the expected utility of the post-sampling task, which depends on a learner's sampling goal (see below). The value of continuing to obtain another sample is a learner's expectation over possible future states given their current knowledge. The probabilities in the expectation are based on a beta distribution parameterized by the outcomes observed so far. The value of the final state (when $n = T$) is just the expected value of stopping, which, along with the Markov property, means that this problem is solvable by backwards induction (Ferguson, 2012).

To compute the utility of stopping under the binary goal, note that the probability of a coin being biased towards heads is $P(\theta > .5) = 1 - I_{0.5}(\alpha,\beta)$, where $I_x(\alpha,\beta)$ is the cumulative distribution function of the beta distribution with parameters $\alpha = h+1$ and $\beta = (n-h)+1$. Assuming that the learner chooses heads when $P(\theta > .5) > 0.5$ and tails otherwise, their expected utility for stopping is

$$u_{stop}^{bin} = max\{1 - I_{0.5}(\alpha,\beta), I_{0.5}(\alpha,\beta)\}\, r \quad (4)$$

where r is the reward for making a correct choice.

The expected utility from stopping under the estimation goal requires specifying a cost function over the distance between the participant's estimate and θ. For simplicity's sake, assume that the learner's answer counts as correct whenever their response lies within a .2 interval surrounding the true value, and incorrect otherwise. The expected utility from stopping under this goal is

$$u_{stop}^{est} = \max_{\theta}\left\{\int_0^1 Beta(x;\alpha,\beta) w_{0.2}(\theta - x)dx\right\} r \quad (5)$$

where $w_{0.2}$ is a boxcar function with a .2 wide interval. By convolving it with the posterior over θ it can be used to find the interval with the largest posterior density.

Predictions

To predict behavior using this model, we first need to choose values for sampling cost, c, potential reward of the secondary task, r, and the maximum number of trials (the length of the horizon), T. To generate more realistic predictions, we additionally assume some stochasticity in people's choice behavior by using a probabilistic choice rule instead of the $max()$ operator in Equation (1). For example, using a softmax choice rule yields the following probability of stopping at each state, $P_{stop} = \frac{exp(V_{stop}/\tau)}{exp((V_{stop}+V_{cont})/\tau)}$. It requires specifying an additional temperature parameter, τ which governs the degree of probabilistic responding (when $\tau = 0$ it chooses the maximum value, as $\tau \to \infty$ it chooses randomly). In this paper we will use the following parameter settings to derive model predictions: $T = 24, c = \$0.03, r = \$2, \tau = .02$. These values were chosen for illustrative purposes and many qualitative results hold across a broader range of values (as long as the trade-off between c and r leads to stopping after a number of samples but before the horizon is reached). The model was used to derive predictions for the expected sample size, stopping probabilities, and the composition of samples after stopping.

Stopping Figure 1A shows the model's predictions for the learner's decisions to stop (white) or continue (green) given the current state (outcomes observed so far). Under the binary goal, learners' should continue sampling if the current sample is balanced (similar proportion of heads and tails) and be more likely to stop when a sample is more extreme. In contrast, the estimation goal leads to a greater probability to continue sampling in a much broader range of states. This is unsurprising because estimation requires a representative sample and overall more data than binary choice. A perhaps more surprising and subtle prediction is that the probability is not uniform for each sample size, but shows a slight pattern of earlier stopping for both very extreme and very balanced samples. Both patterns are caused by a learner's expected success at the estimation task. When samples are extreme, expected accuracy is high and learners can stop earlier with a reasonable chance at success. When samples are very mixed (close to 0.5 average), a high chance of good performance would require too many costly samples, at which point stopping early and making a best guess might have higher expected value. Basically mixed samples tend to take too many costs samples to resolve accurately.

From the stopping matrices in Figure 1A one can now derive expected probabilities for stopping points, that is learners' expected final state after stopping, for specific values of θ. Figure 1B shows the likelihood of different stopping points assuming $\theta = 0.5$. As expected, binary learners are predicted to end up in early extreme states (all heads or all tails) or later mixed states, whereas estimation learners show a wider dis-

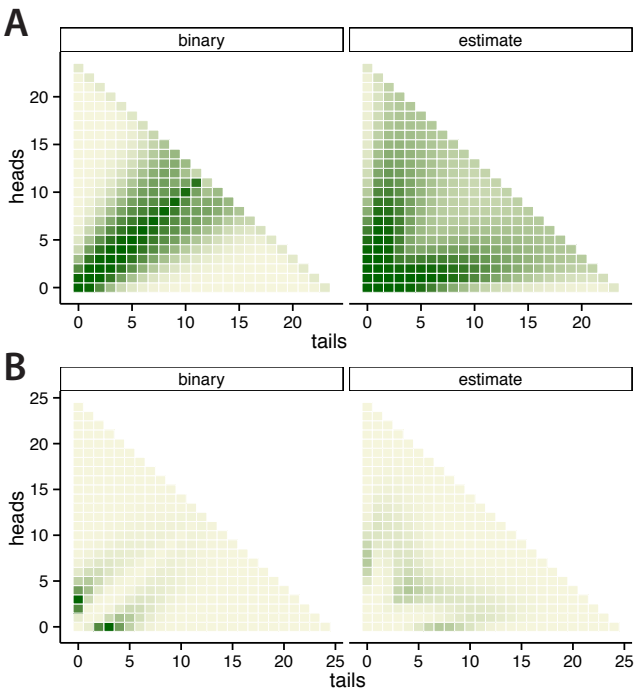

Figure 1: Model Predictions. A: Probability of continuing (green) or stopping (white) given observed data. B: Probability of having stopped in any given state. Modeled using uniform θ.

tribution of stopping states.

Sample Size Figure 2A shows the expected distribution of number of total samples taken, assuming learners are sampling from a process with underlying $\theta = 0.5$. Estimation learners are expected to collect more samples than binary learners, who face an easier secondary task and therefore terminate earlier on average.

Sample Composition The fact that stopping rules affect expected *size* of samples is not surprising given the different loss functions of the two goals. A more intriguing question is whether sampling goals also preserve the *properties of the distribution* that generates each sample. To investigate this, we will consider the proportion of heads and tails within each sample, after repeatedly sampling and stopping under the two goals. We expect that binary sampling would on average lead to more unequal samples (all heads, or all tails) since its stopping rule terminates early when outcomes are extreme (after two or three heads one can be pretty confident that heads is more common, for example) and continues when early outcomes are mixed (after one tails and two heads a decider might want to flip the coin at least one more time to be sure). On the other hand the estimation condition predicts a wider array of stopping points because the goal is to get an accurate picture of the average outcome probability.

Figures 2B and C confirm this effect of stopping rule on sample composition. It shows the expected frequency distribution of sample averages under each sampling goal, when the true $\theta = 0.5$(A) or $\theta = 0.7$(B). In the estimation condition, sample means fall around the expected value of the distribution, whereas the binary condition leads to many samples with only one type of outcome (all heads or all tails). Note that these more extreme samples tend to be small (that's when learners terminate early), but they still mostly contain more than one observation, which can be seen from Figure 1A. Thus, the difference in sample mean distributions between the two goals is not just due to the fact that the binary condition leads to samples of size 1 which necessarily have an average of 0 or 1.

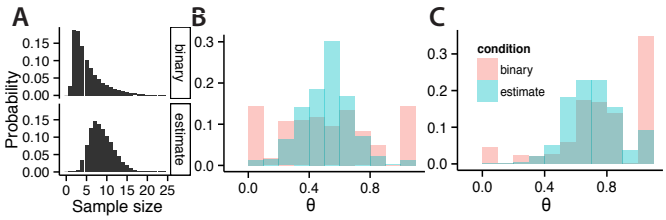

Figure 2: A: Distribution of sample size by condition, using uniform θ. B and C: Predicted distribution of sample averages, using $\theta = 0.5$, and $\theta = 0.7$.

Summary This analysis shows how sampling goals can influence on learners' expected stopping strategies even in a very simple binary sampling task. Furthermore, even though the model gives learners no influence on *what* to sample and it assumes they are choosing optimally, the resulting distribution of samples is *not* necessarily representative of the underlying generating probabilities. Instead, sampling with the binary goal leads to far more samples with extreme averages than the estimation goal, which generates more representative samples.

This illustrates that even in natural sampling tasks, in which samples generally "conserve the properties of the universe" Fiedler (2008) optimal stopping rules can distort the experienced data depending on the learner's goal during the sampling. In other words, even the minimal decision of when to stop can introduce systematic discrepancies between a sample and the population distribution.

Experiments

These theoretical results yield a number of interesting empirical questions, which will be addressed with two experiments. First, the model uses a relatively sophisticated forward-looking, optimal decision-making process to govern people's stopping decisions. The behavioral patterns in Figure 1 require not only that learners can assess the cost-accuracy trade-off and adjust the number of samples accordingly. They need to also be able to assess the expected performance *given the current sample content*. In the following two experiments we will therefore examine to what degree participants are actually engaging in such forward looking behavior when deciding to terminate or continue sampling.

Both experiments will also examine how goals affect sample composition, compared to the predictions of the model. A

key intention of the modeling effort above was to show that distributions of sample means differ depending on the goal, which suggests that even natural sampling does not preserve population characteristics in a straightforward manner.

Another question, albeit one that is more difficult to answer, is whether people's judgments about the distribution that generates the samples are affected by the different sampling strategies. For example, since binary sampling, according to the model, should produce more extreme samples, does this lead people to expect more extreme outcomes? Experiment 2 starts to address this question and yields some preliminary results.

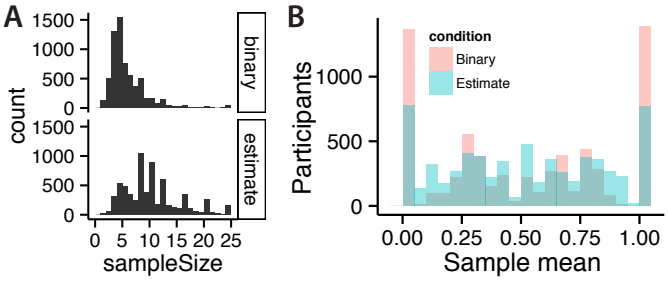

Figure 3: A: Histogram of sample sizes in Experiment 1. B:Distribution of sample averages in Experiment 1.

Experiment 1

Experiment 1 manipulated sampling goals in a simple repeated Bernoulli sampling task that shared all the characteristics of the task described in the modeling section. The sampling goal was manipulated between participants who either had to estimate the overall value of θ (*estimation condition*) or decide whether θ was greater or lower than 0.5 (*binary condition*).

We predicted that sampling goals would affect both the number of samples collected on average (higher for the estimation vs. binary goal), and the relationship between stopping points and current sample composition. As outlined in the previous section, we expected participants engaged in binary sampling to be more likely to stop early when strong evidence is encountered (average closer to 0 or 1) and more likely to continue when evidence is mixed (average closer to 0.5).

Participants 276 participants were recruited via Amazon Mechanical Turk. They were paid \$2 for participating with an option to win a bonus of up to another \$2 (explained below).

Stimuli Participants were told that they were repeatedly drawing cards from 50 different card decks consisting of 200 cards each. Cards could be either red or blue, and the distribution of red and blue cards in each deck was determined semi-uniformly (using an evenly-spaced distribution of Bernoulli probabilities θ that were then used to randomly draw the cards for each deck). Participants were explicitly told to assume a uniform distribution of numbers of red and blue cards in each deck, as well as being told that the order of cards in each deck was completely random ("well shuffled").

Procedure For each card deck, participants could repeatedly (up to 24 times) turn over cards using a button press to reveal their color. A counter on the screen would tell how much of potential bonus remained after each sample (which cost \$0.05). For each card deck, the potential bonus started at \$2. At any point participants could also decide to move on to the secondary task (binary choice or estimation) via a different button. In the binary task, participants would then make a two alternative forced choice decision of whether they thought there were more red or more blue cards in the deck. In the estimation task they gave an exact estimate using a bar slider whose handle had two lines that indicated the .2 interval within which their response would be counted as correct. After giving their estimate, all 200 cards of the card deck were revealed and participants received feedback on their choice or estimate. If their estimate was correct, their potential bonus was recorded as the bonus that remained after the sampling phase, otherwise the it was recorded as \$0. At the end of the task, one card deck was chosen at random and participants were actually paid the bonus earned on that deck.

Results & Discussion To recap, our two main predictions motivated by the modeling were that participants in the estimation conditions would on average collect larger samples, and that participants in the binary conditions would be strongly influenced by the composition of the current sample when deciding whether to stop or to continue sampling. Figure 3A shows the distribution of sample sizes by condition. As expected (compared to model predictions in Figure 2A), participants in the estimation group took larger samples than participants in the binary group ($t(211) = -11.18, p < 0.01$), suggesting that people were aware of the trade-off between sampling cost and accuracy and were willing to incur higher cost (more samples) in the more difficult task. There were also visible spikes for certain sample sizes (8, 10, and 12) in the estimation condition, suggesting that perhaps these were preferred, salient stopping points for a range of participants (cards on the screen were aligned in columns of four, so spikes at 8 and 12 may be the result of some aesthetically inspired fixed-sample-size stopping rule).

Figure 4 shows the proportion of times participants ended up in each state after terminating search. There was a clear difference between the stopping patterns of the two groups. Participants in the binary group were more likely to stop in early extreme states, but continued sampling if early evidence was mixed. Participants in the estimation group, on the other hand, showed no discernable stopping pattern based on current sample composition. For a given sample size, there appeared to be fairly equal stopping probability across different proportions of red/blue cards. Also, the previously mentioned average preference for certain sample sizes is reflected in diagonal "ridges" in the estimation goal data. What we did not observe were the earlier stopping patterns for very mixed or very extreme samples, which were predicted by the model.

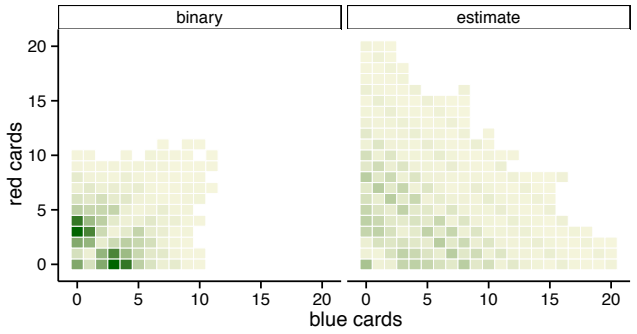

Figure 4: Stopping probabilities (dark means high) by sampling goal in Experiment 1

Possibly participants were unaware of the differences in uncertainty about the true average that emerge as a consequence of the sample composition.

Finally, we also examined the sample composition resulting from these stopping rule differences. Figure 3B shows the distribution of sample averages by condition. The binary sampling group showed a much larger proportion of extreme samples (all red/blue) compared to the estimation condition. In summary, people's stopping decisions lead to differences in the resulting sample content in a manner that is qualitatively consistent with the model predictions.

Experiment 2

Experiment 1 established that people adapt their stopping rule to different sampling goals in a way that is broadly consistent with the predictions of our optimal stopping analysis. However collecting a sample is usually not the end goal of learning. Instead learners use a sample to enable predictions about more general properties of the world. For example, people might listen to a few songs by a band then attempt to generalize from those song to the entire catalog. Experiment 2 was designed to explore how goal-induced stopping rule differences might affect people's subsequent judgments of more general population parameters.

A critical challenge was ensuring that participants were motivated to learn more general properties from the samples they collected. To that end, we altered the Experiment 1 design by grouping samples under two different generative processes (specifically two different schoolteachers who had the ability to influence the test scores of students in their class). The intention was to encourage participants to form summary representations of these parameters based on the samples collected for each. By varying the average outcome probabilities of these two groups/parameters (one high, one low), we would then be able to compare the potential impact of sampling goals on the estimates of each group's average. We expected binary samplers to be exposed to a greater difference between outcome probabilities due to a stopping rule that leads to more extreme outcomes and wanted to know if this would bias their population estimates to also be more more extreme (even higher, and even lower) as a consequence.

Participants 58 participants were recruited via Amazon Mechanical Turk and paid $2 for participation (plus bonus).

Stimuli To encourage participants to form summary representations of two separate distributions a cover story asked them to repeatedly sample binary results (correct or incorrect test scores) from a set of high school students, who were each taught by one of two teachers (teacher A or teacher B). That is, rather than randomly and uniformly drawing values for θ for each student (as in Experiment 1), outcome probabilities were now drawn from one of two hierarchical distributions. The mean accuracy of students from each teacher differed, such that one teacher yielded a higher average ($\mu_{high} = 0.7$) of correct test scores than the other ($\mu_{low} = 0.3$). Individual students' outcome probabilities were distributed as $\theta \sim Beta(3\mu_t, 3(1-\mu_t))$, depending on a student's teacher type, $t \in \{high, low\}$. Participants were each given different random draws from this distribution.

Procedure The main part of the experiment was similar to that of the model and previous experiment. Participants repeatedly sampled 100 students' test answers (correct or incorrect), while losing $0.05 per question from a potential $2 bonus per student. Each student was presented along with their respective teacher, which alternated between teacher A and B. Participants were again assigned to a binary choice or estimation condition, but unlike in the previous experiment they did not receive feedback on the true distribution for every student. After sampling 50 students per teacher, participants also rated the average quality of each teacher's students (using a slider between 0% and 100% correct). Again, they were paid a bonus based on one randomly chosen student from the sampling task.

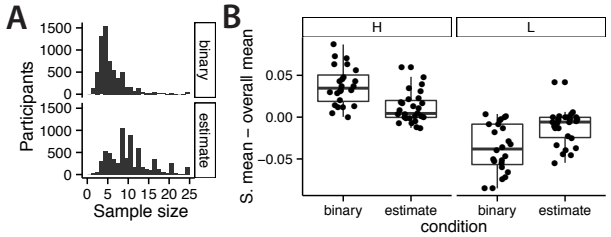

Figure 5: A: Number of samples collected in Experiment 2. B: Difference between mean of the sampling means, and mean of all samples by condition and teacher type (*H*igh or *L*ow).

Results & Discussion Figure 5A shows the distribution of sample size by condition and Figure 6 depicts participants' stopping probabilities by state. Despite substantial changes compared to Experiment 1 (addition of teachers, two different outcome distributions, omission of feedback) the results are qualitatively similar. Again, binary learners took smaller samples on average and were more likely to stop given highly

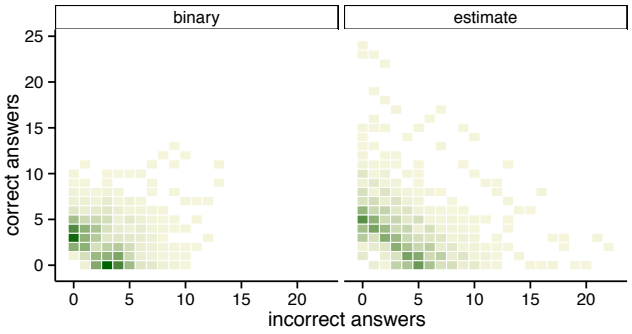

Figure 6: Stopping probabilities (dark means high) by sampling goal in Experiment 2.

unbalanced samples (all correct or all incorrect).

Next, we considered the impact of stopping on the evidence collected for each of the teachers. Because each participant saw a different sequence of outcomes and the resulting distribution of samples varied as a consequence, we developed the following measure of a participants' *potential* for being biased given the data they saw: For each participant, we computed the difference between the average sample mean for each teacher and the overall teacher average irrespective of the sample boundaries (i.e. treating all students as one sample). This difference is 0 when a participant always takes samples of the same size, but may diverge when sample sizes are uneven, particularly when sample size and sample mean are non-independent, as we expect to happen under binary sampling. Figure 5B shows this difference by teacher and condition. As expected, in the binary condition, the average sample mean was more extreme (higher for the good teacher, lower for the bad teacher) than the overall average, compared to the estimation condition, (significant condition and teacher type interaction, $t(98) = 5.62, p < 0.01$), indicating again that stopping rules affected the sample distribution.

Next, we tried to investigate the impact of this difference on people's subsequent judgments of the teachers. To do so, we regressed participants' posterior estimates of the teacher mean (one estimates per participant per teacher) on their overall teacher mean and the difference between overall mean and average sample mean (from Figure 5B). If the latter had any additional positive effect of peoples estimates beyond the overall mean, this would indicate that participants were not correcting for the unequal sample sizes that introduced this difference. However, we failed to find a significant positive effect of the difference, $t(99) = 1.627, p = 0.11$. Due to considerable variation in the actual outcomes observed by each participant (all received different sequences) and the high variability of theses posterior estimates, it may be that we lacked sufficient statistical power to detect this effect. We are currently working towards an improved design of this study that reduces variance in people's posterior estimates of teacher quality and their sample composition.

Discussion

In this paper, we presented a theoretical analysis showing how sampling goals can have a profound impact on people's stopping strategies even under natural sampling. Crucially, our results go beyond showing a difference in the *amount* of evidence collected, but demonstrate that goals also affect also the composition and statistical properties of samples. This demonstrates that natural sampling does not necessarily produce samples that mirror the statistical properties of the environment. Instead, just the simple decision of when to stop sampling can lead people to collect samples with vastly different distributional characteristics depending on their goal. While Experiment 2 did not show a robust effect of this non-representative sampling on more general predictions, it is an interesting question if people can take into account the process by which samples were gathered and perform the necessary correction to remove possible biases. Prior work on sampling and estimating binary outcomes suggests that this might be difficult for people (Fiedler, 2008; Griffin & Tversky, 1992). If this turns out to be more generally true, it suggest that simply deciding when to stop sampling information from a natural process can strongly bias judgement.

Acknowledgments This work was supported by grant number BCS-1255538 from the National Science Foundation and a John S. McDonnell Foundation Scholar Award to TMG.

References

Denrell, J. (2003). Vicarious learning, undersampling of failure, and the myths of management. *Organization Science*, *14*(3), 227–243.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538.

Ferguson, T. S. (2012). *Optimal stopping and applications.* Electronic Text. Retrieved from `https://www.math.ucla.edu/ tom/Stopping/Contents.html`

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 186.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, *24*(3), 411–435.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (in press). Information sampling behavior with explicit sampling costs. *Decision*.

Lee, M. D., & Paradowski, M. J. (2007). Group decision-making on an optimal stopping problem. *The Journal of Problem Solving*, *1*(2), 06.

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the naivety assumption in sampling explanations of judgment biases. *Psychological review*, *118*(2), 379.

Rich, A. S., & Gureckis, T. M. (2014). The value of approaching bad things. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36rd annual conference of the cognitive science society.* Austin, TX.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.