A Memory Model of Belief Formation[*]

Maxim Bakhtin
Stanford University

Markus Mobius
Microsoft Research

Muriel Niederle
Stanford University

May 30, 2024

Abstract

The agent in our model retrieves memories and combines them with the prior to form a belief. The agent is fully Bayesian and rational but faces a constraint on memory retrieval—she can only sample observations one at a time instead of retrieving all of them simultaneously. Retrieval is primarily random, but the agent can partially target retrieval using an index. The index splits the database of memories into two (or more) groups based on the values of one (or more) attribute. The agent chooses which indexed group to sample in each period to ensure that her beliefs are as accurate as possible. We show that the agent will generically oversample one group and characterize three forces that determine which group the agent samples more intensely. We then show that oversampling translates directly into belief distortion. We use this insight to explain well-known biases in beliefs across individuals, such as the "depression babies" effect, rational stereotypes, and the dependence of beliefs on the history of previously encountered problems.

JEL Classification: D81, D82

Keywords: memory, belief formation, optimal sampling, rational stereotypes, history dependence

[*] We are grateful to participants of the 34th Stony Brook International Conference on Game Theory and Learning and Bounded Rationality Workshop and 2023 briq Summer School in Behavioral Economics for helpful comments.

> . . . [A]s we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say, we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know . . . [I]t is the latter category that tends to be the difficult ones.
>
> (Donald Henry Rumsfeld, US Secretary of Defense (2002))

1 Introduction

Donald Rumsfeld's well-known quote captures the basic idea that we are commonly asked to voice opinions on issues we have not considered before. For example, a colleague organizing a department event might ask us if economists are more or less likely to enjoy board games than the rest of the population. The standard model of decision-making under uncertainty in economics essentially assumes that we have priors on the propensity of the average economist and non-economist to like board games, which we constantly update whenever we observe someone playing (or not playing) board games. We call this form of reasoning that treats beliefs as primitives *statistical reasoning.*

However, in reality, there are too many states of the world to keep track of, and we doubt that anyone but board game enthusiasts has a ready-formed belief about the relative enthusiasm of economists for board games. A natural way to construct an on-demand belief is to *sample* one's memory. We call this process *anecdotal reasoning.* Of course, if sampling from memory is perfect, then the agent can reconstruct the same beliefs she would have held under statistical reasoning.

However, this is no longer the case if memory is imperfect. There are two basic ways to introduce memory imperfections. First, *storing* memories could be costly, so the agent can only recall a limited number of memories. Wilson (2014) uses this approach and analyzes *which* anecdotes to store in one's memory. Second, there could be a cost for *retrieving* memories. Our paper builds on this friction, and our analysis focuses on *how* to sample from memory in a constrained optimal way.

The agent in our model retrieves memories and combines them with the prior to form a belief. The agent is Bayesian and rational but faces a constraint on memory retrieval—she can only sample observations one at a time instead of retrieving all of them simultaneously. Retrieval is primarily random, but the agent can partially target retrieval using an index. The index splits the database of memories into two (or more) groups based on the values of one (or more) attribute. The agent chooses which indexed group to sample in each period to ensure that her beliefs are as accurate as possible.

We show that the agent optimally samples one group more than the other, which we call *oversampling*, and identify three effects that contribute to it. First, the variability effect pushes the agent to oversample the group that has more uncertainty about it. Second, the availability effect pushes her to oversample the group with observations that are harder to access with her index. This oversampling only partially compensates for the difficulty in accessing observations. The agent ends up with more observations that are more available than observations that are less available. Third, the importance effect pushes the agent to oversample the group that is more important for the statistic the agent is estimating.

The agent's beliefs are correct in the limit, but in finite samples, they are distorted, and the magnitude of the distortion depends on the sampling strategy. The agent uses an unbiased estimator relative to her belief, but she is distorted relative to the realized value of the statistic. This distortion is what an outside observer who knows the true statistic would see. The agent's belief combines retrieved observations with her prior. As a result, in finite samples, the agent is distorted relative to the actual value of the statistic she is estimating. For linear problems, her distortion about each group is always toward her prior about that group. The magnitude of the distortion for each group decreases in the number of observations the agent samples from that group. Despite this monotonicity, the total distortion may be non-monotonic in the sample size if the agent learns about the different groups at different rates.

Figure 1 illustrates the basic mechanics of the model. The agent receives a problem about the difference in the share of Runners among Male and Female Scientists, $P(\mathit{Runner} \mid \mathit{Scientist}, Male) - P(R \mid \mathit{Scientist}, Female)$. The agent starts with a prior belief $P_0(\mathit{Runner} \mid \mathit{Scientist}, Male) = P_0(R \mid \mathit{Scientist}, Female) = 0.1$. Suppose the truth is that $P(\mathit{Runner} \mid \mathit{Scientist}, Male) = P(R \mid \mathit{Scientist}, Female) = 0.4$. The agent has a Gender index, so she chooses whether to sample a man or a woman in each period. With each sample from a given gender, her beliefs move away from the prior and closer to the truth about that gender. In the limit, the agent's beliefs converge to the truth. In finite time, however, her beliefs will be distorted. Suppose Scientists are more prevalent among men than women. Then, the availability effect causes the agent to oversample Male Scientists compared to Female Scientists[1]. As a result, the agent learns faster about Male Scientists than Female Scientists. So, her belief about the share of Runners among Male Scientists increases faster than her belief about Female Scientists. This belief trajectory is illustrated by the arrow from the 'Prior' point to the 'Truth' point in Figure 1. At a finite time $t > 0$, her beliefs are distorted up: $P_t(\mathit{Runner} \mid \mathit{Scientist}, Male) - P_t(R \mid \mathit{Scientist}, Female) > 0$. This distortion disappears in the limit as t goes to infinity.

We apply our model to explain various biases. The model provides a memory-based explanation for why some people are optimistic and others pessimistic. Optimists have easier access to memories of good days. Formally, optimists have an index that splits memories into good and non-good periods, which combine neutral and bad periods. As a result, observations of good periods are more available than observations of bad periods. Therefore, beliefs about the good periods are more accurate than beliefs about the bad periods. If the prior beliefs about an outcome of interest are medium for all periods, the agent's belief is distorted positively about bad periods and negatively about good periods. Since the agent learns faster about good periods, her total distortion is driven mainly by the distortion about bad periods, which is positive. As a result, an agent with an index for good periods is more optimistic about the uncertain outcome. In the financial domain, this explains why people who experienced the Great Depression at a young age ("depression babies") are more pessimistic about the stock market and less likely to invest in it (Malmendier and Nagel

[1] See Section 4.2 for more details.

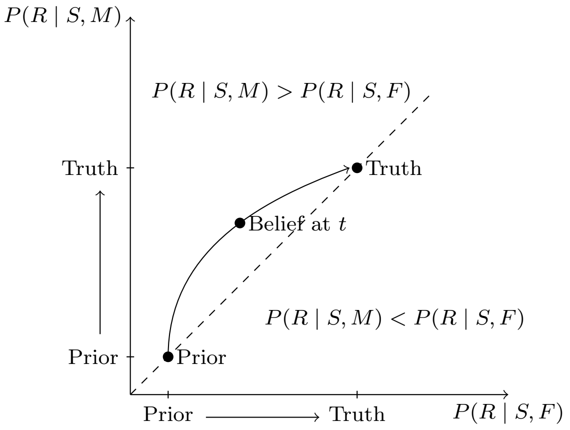


Figure 1: Evolution of agent's beliefs over time

(2011)). Someone who grew up during a period of depression is more likely to have easier access to memories about depression than someone who grew up during a period of economic boom. In our model, depression babies would have an index that separates memories of depression from non-depression. As a result, an agent with a depression index is more pessimistic about the stock market than an agent with a boom index.

The model also rationalizes stereotypes based on generalizations. Consider an agent with an index for nationalities combining all Europeans in one group. For this agent, it is harder to retrieve observations on specific European nationalities, like Germans, than for someone with an index for each nationality separately. As a result, the agent with the coarse index has less accurate beliefs about specific nationalities. Furthermore, she has a stronger distortion toward the common European prior. Therefore, the agent with the coarse index is more likely to give similar answers about all European nationalities. Although this may look like stereotyping, the agent gives her best answer under the constraints. If the agent is constrained to having a coarse index, she may want to index observations probabilistically, so that all nationalities are equally represented in the coarse index. Probabilistic indexing would allow the agent to optimize for the worst-case scenario—when she needs to form a belief about very rare nationalities.

Apart from generalization stereotypes, the model explains stereotypes based on anchoring. Consider an agent who has an index for Germans but needs to form a belief about Austrians. The agent knows that Germans and Austrians are similar with some probability. The agent could search for Austrians in the non-German group, but it would take a long time to find enough Austrians. Alternatively, the agent could form a belief about Germans and adjust it in the direction of her prior about Austrians. The benefit of this approach is that the agent has easy access to observations of Germans, so she can easily form an accurate belief about Germans. This approach is better if Austrians are similar to Germans with sufficiently high probability. Thus, the agent may form a stereotype about Austrians by anchoring to Germans. Moreover, another agent with an index, for example, for the French, may anchor to her beliefs about the French. The two agents will have different beliefs about Austrians: one is closer to Germans, and the other is closer to the French.

Our model also explains why people with the same experiences and priors may have different distortions depending on their previous problems. If the agent faces several problems, she may keep some retrieved observations in the short-term memory, available for instant use. Someone who faced questions about female scientists is likely to have several observations of female scientists readily available. This agent will likely have beliefs distorted up about the average mathematical skills of all women. In contrast, someone who recently formed beliefs about female non-scientists will likely have beliefs distorted down.

So far, we have treated an index as exogenously given, but the agent may prefer one index to another. We highlight three properties of an attribute that make it a good candidate for an index, which can be interpreted as salience. The first property is that the agent needs to form beliefs about a group defined by this attribute frequently. An index is most helpful for statistics that condition on the indexed group. Therefore, if such problems arise frequently, the attribute becomes a helpful basis for indexing. The second property is informativeness. An informative attribute splits the database into two groups that are substantially different. For example, gender and age are informative, while the first letter of the name is not. Informativeness helps for two reasons. First, an informative index makes it easier to target non-indexed attributes. For example, a gender index helps retrieve observations of football players by focusing on men. Second, more extreme parameters are increasingly easier to estimate with low variance. Therefore, an estimate of an average parameter based on two different groups is more accurate than its estimate based on two similar groups. The third property is unbalancedness. An attribute is unbalanced if it has a rare group, for example, minority status or rare skill. An index for an unbalanced attribute is especially useful for forming beliefs about a rare group. Without this index, the agent would spend long searching for relevant observations. The penalty for not having an index for an unbalanced attribute, in terms of required sample size, is larger than for a balanced attribute. These three properties suggest an interpretation of what it means for an attribute to be salient. A salient attribute is informative and unbalanced and defines a frequent target group for belief formation.

Lastly, we connect our model to models of cued recall based on similarity and representativeness (e.g., Kahana (2012); Bordalo et al. (2023b)). A fundamental assumption in those models is that the likelihood of recalling a given observation after receiving some cue increases in the similarity between the observation and the cue and decreases in the similarity between other observations and the cue. The optimal sampling strategy in our model has a similar structure, which arises endogenously as a result of optimization. This result allows us to pin down the measure of similarity that rationalizes the central assumption in the cued recall models. Relatedly, we show that representativeness also plays a role in our model. A characteristic is representative of a group if it is more prevalent in this group compared to another (Tversky and Kahneman (1983)). If some characteristic is representative of an indexed group, the index is helpful for retrieving observations with that characteristic. As a result, the agent may optimally sample more observations with the representative characteristic than their share in the population. Her beliefs may be driven more by the representative observations compared to someone who does not use an index.

Section 2 discusses related literature. Section 3 describes the model setup. Section 4 identifies sampling distortions. Section 5 describes the resulting belief distortions. Sections 6-8 illustrate various applications of the model.

2 Related Literature

There is a large body of experimental and empirical evidence on memory imperfections and their effects on beliefs and decision-making. Kahana (2012) provides a broad review of findings about memory from the psychology literature. Bordalo et al. (2016, 2021, 2023b) experimentally show that memory retrieval is context- and cue-dependent and biased by representativeness.[2] They further show that these recall patterns lead to belief biases. Enke et al. (2023); Afrouzi et al. (2023) show that people overreact to recent news in their forecasts because memory is associative — current signals trigger recall of similar signals from the past. Graeber et al. (2022) compare stories and statistics. In their experiment, stories are more easily recalled than statistics and have a more persistent effect on beliefs.Malmendier and Wachter (2021) review empirical evidence of the effects of past experiences on financial choices. For example, Malmendier and Nagel (2011) show that people who grew up during the Great Depression are more pessimistic about the stock market and invest less. Charles (2022) shows that associations and recency of memories affect trading behavior. Kwon and Tang (2020) also use investor data to document systematic over- and underreaction to news consistent with beliefs based on the representativeness heuristic. Bordalo et al. (2016) review an extensive literature on stereotypes, many of which can potentially be attributed to memory. Our model explains some of these observations in a fully Bayesian model. We show that various stereotypes and biases can arise in a rational and Bayesian model due to sequential memory retrieval.

There are various approaches to modeling memory imperfections. One approach directly incorporates empirical regularities in the memory models. Dougherty et al. (1999); Nilsson et al. (2005), as well as models described in Kahana (2012) are some examples of this approach. Mullainathan (2002a) formalizes two empirical findings about memory — rehearsal and associativeness — and incorporates them into a consumption model. Fudenberg et al. (2022) assume that the agent is more likely to remember some experiences than others but does not account for it when updating beliefs. In a series of papers, Gennaioli and Shleifer (2010); Bordalo et al. (2016, 2021, 2023b) develop models of memory based on cued recall. The central assumption is that the likelihood of recalling an observation increases in the similarity between that observation and the cue and decreases in the similarity between other observations and the cue. These models capture the representativeness heuristic and generate many documented biases. Although our model does not assume specific memory retrieval patterns, it is related to the cued recall models. We show that the

[2] ‘An attribute is representative of a class if it is very diagnostic, that is, if the relative frequency of this attribute is much higher in that class than in a relevant reference class’ (Tversky and Kahneman (1983), p. 296). For example, having red hair is representative of the Irish population because red hair is more common among the Irish than among other populations.

optimal sampling strategy is consistent with the cue-guided recall, thus providing a microfoundation for this assumption. Furthermore, our model pins down a similarity measure that rationalizes cued recall.

Another approach to modeling memory limitations imposes "technological constraints" on human memory but otherwise assumes rationality and sophistication. Wilson (2014) assumes that the agent has a limited memory capacity and summarizes all information using several states. Limited memory size leads to various behavioral phenomena, such as stickiness and polarization of beliefs, and confirmation bias. Da Silveira et al. (2020) assume that memory has limited complexity. This assumption leads to biased beliefs and overreactions in forecasts. Afrouzi et al. (2023) impose a cost on retrieving past observations, leading to overreaction. In their model, the agent chooses how much to remember, while in our model, the agent chooses what to remember.Neligh (2022) assumes that memory decays over time, but the agent can exert costly effort to preserve a memory for longer. This model generates the recency effect. Our model is closely related to these papers in the spirit of using the "technological constraint" approach. The constraint we impose is that the agent cannot retrieve all observations at once but has to retrieve them one by one to update her beliefs. The agent is otherwise rational and Bayesian. Compared to other models, we focus on a new memory limitation and show that it can explain new biases and stereotypes.

Our model also relates to the literature on dynamic information acquisition but microfounds the signal structure, which is usually given exogenously (e.g., Che and Mierendorff (2019); Azevedo et al. (2020); Gossner et al. (2021); Mayskaya (2022); Liang et al. (2022)). These papers also characterize the optimal learning process. For example, Liang et al. (2022) derive the optimal strategy of allocating attention to different signals in closed form. The properties of this optimal strategy agree with our results. However, existing papers model the learning process as observing signals with exogenously specified informativeness. In contrast, we model learning as observing samples of data. As a result, the informativeness of each observation is not fixed exogenously but is determined endogenously by the population distribution and the agent's index. This approach allows us to explain why learning about some groups is more difficult than others and connect this to known stereotypes and belief biases.

Another related strand of literature is models of categorization. Mullainathan (2002b); Fryer and Jackson (2008); Mohlin (2014) model beliefs and decisions based on coarse categories. Categories combine different observations so that the agent's beliefs about a specific category element are biased toward the category average. We do not assume that the agent in our model forms explicit categories, but the index effectively splits memories into categories. As a result, our model explains stereotypes based on categories as well. However, in contrast to the categorization models, our model also predicts that the magnitude of the belief distortion depends on the problem the agent faces and that the distortion disappears in the limit.

We are also related to papers on applications of memory models to various domains. Kőszegi et al. (2022); Gottlieb (2014) apply memory imperfections to self-esteem, Wachter and Kahana (2019) to financial markets, Bordalo et al. (2020) to consumer choice, Bordalo et al. (2023a) to forecasting risks; Malmendier and Wachter (2021) review models with applications to financial choice. We also apply our model to financial choice and show that it can explain the pessimism of Depression Babies (Malmendier and Nagel (2011)).

3 A Model of Belief Formation through Sampling

3.1 Model Setup

Consider an environment with a database $\mathcal{D}$ containing individuals with K binary attributes. All individuals can be grouped into 2^K subgroups with identical attribute values. Let $\boldsymbol{x} = (x_1, ..., x_{2^K}) \in \Delta^{2^K-1}$ denote the shares of each subgroup.

There is one long-lived agent, who does not know the realization of $\boldsymbol{x}$ and has a Dirichlet prior $\boldsymbol{x} \sim Dir(\alpha_1, ..., \alpha_{2^K})$. One special case is the flat prior $Dir(1, ..., 1)$, under which all realizations of $\boldsymbol{x}$ are equally likely. When called to take an action, the agent must choose an action $a \in \mathbb{R}$ to match a given statistic $f(\boldsymbol{x})$, where f is a twice differentiable function known to the agent. The agent minimizes expected quadratic loss:

$$\min_{a \in \mathbb{R}} \mathbb{E}(a - f(\boldsymbol{x}))^2 \tag{1}$$

Before taking the action, the agent can refine her estimate of the statistic $f(\boldsymbol{x})$ by retrieving observations from the database. The crucial limitation is that the agent can only retrieve one observation at a time. In each period t, unless the agent is called to choose an action, she samples one unique observation from the database belonging to subgroup i with probability x_i. She observes the values of all its K attributes. The agent combines the prior with the retrieved observations using Bayes' rule to form a posterior belief G. Given the posterior G, the agent's optimal action a^* is

$$a^* = \mathbb{E}_G(f(\boldsymbol{x})) \equiv \widehat{f(\boldsymbol{x})} \tag{2}$$

An *expert* agent has a memory retrieval technology that allows her to target groups of observations instead of sampling random ones. We call this technology an *index*.

Definition 1. *An* ***index*** $Ind = \{A, \tilde{A}\}$ *is a partition of the subgroups in database* $\mathcal{D}$ *that allows the agent to target memory retrieval from groups* A *and* $\tilde{A}$*.*

An index $Ind = \{A, \tilde{A}\}$ partitions the database into two groups of observations, A and $\tilde{A}$. For example, if the agent has a gender index, all observations are divided into $A = Male$ and $\tilde{A} = Female$. An expert can choose in each period which group to sample from, A or $\tilde{A}$. In contrast, a *non-expert* does not have an index and can only retrieve observations randomly, i.e., the non-expert samples A and $\tilde{A}$ with probabilities equal to population shares. We assume that the total shares of the indexed groups, $P(A)$ and $P(\tilde{A})$, are known.[3]

[3] For individuals with an index, knowledge of $P(A)$ and $P(\tilde{A})$ is crucial in deciding which subgroup to sample

In each period t, the expert agent chooses whether to sample an observation from group A or group $\tilde{A}$, $s_t \in \{A, \tilde{A}\}$. Her objective is to minimize the expected next-period[4] loss from choosing the optimal action $a^* = \widehat{f(\boldsymbol{x})}$ conditional on her current belief G_t, where the expectation is taken over the distribution of the next sampled observation:

$$\min_{s_t \in \{A, \tilde{A}\}} \mathbb{E}_{G_t}[\mathbb{E}_{G_{t+1}}(\widehat{f(\boldsymbol{x})} - f(\boldsymbol{x}))^2] \tag{3}$$

The expected loss from taking the optimal action is the Mean Squared Error of $f(\boldsymbol{x})$ given the statistic estimator $\widehat{f(\boldsymbol{x})}$, so the agent's objective when choosing whom to sample is to minimize the expected next-period MSE:

$$\min_{s_t \in \{A, \tilde{A}\}} \mathbb{E}_{G_t}[MSE_{G_{t+1}}(f(\boldsymbol{x})|\widehat{f(\boldsymbol{x})})] \tag{4}$$

If the agent uses the unbiased estimator $\widehat{f(\boldsymbol{x})} = \mathbb{E}_G(f(\boldsymbol{x}))$, the agent's objective function reduces to posterior variance. Therefore, in each period t, the agent chooses to sample from the group that minimizes the expected next-period posterior variance:

$$\min_{s_t \in \{A, \tilde{A}\}} \mathbb{E}_{G_t}[Var_{G_{t+1}}(f(\boldsymbol{x}))] \tag{5}$$

3.2 Optimal Sampling

The solution to the agent's exact problem (5) is hard to characterize and interpret for a general statistic f. Instead, we analyze the approximate problem, which is more tractable and interpretable. Solutions to the approximate and exact problems converge in the limit and are similar in finite sample simulations.

We approximate the exact problem using two simplifications. First, in each period t, the agent treats the current estimate of subgroup shares $\hat{\boldsymbol{x}}_t$ as constant, ignoring the fact that she will update it based on an extra sampled observation. Second, instead of the exact variance, the agent minimizes its Taylor approximation.

Specifically, consider a general Dirichlet prior $\boldsymbol{x} \sim Dir(\alpha_1, ..., \alpha_{2^K})$, and denote $\alpha_A = \sum_{i \in A} \alpha_i$, $\alpha_{\tilde{A}} = \sum_{i \in \tilde{A}} \alpha_i$. Let $P(A)$ and $P(\tilde{A}) = 1 - P(A)$ be the population shares of groups A and $\tilde{A}$, known to the agent. Let $g(i)$ be the indexed group that subgroup i belongs to, i.e.,

$$g(i) = \begin{cases} A \text{ if } i \in A \\ \tilde{A} \text{ if } i \in \tilde{A} \end{cases} \tag{6}$$

Given a sample of observations, let n_i be the number of observations from subgroup i, and let

from. By endowing the non-index individuals with $P(A)$ and $P(\tilde{A})$ as well, we ensure that differences in beliefs are only due to the different memory retrieval technologies.

[4] We conjecture that the sample would be the same in most cases if the agent was optimizing multiple periods ahead.

$n_{g(i)} = \sum_{i \in g(i)} n_i$. The estimates for individual subgroup shares x_i are:

$$\hat{x}_i = \mathbb{E}_G(x_i) = \frac{n_i + \alpha_i}{n_{g(i)} + \alpha_{g(i)}} P(g(i)) \tag{7}$$

At each t, the agent's problem is to choose whether to sample from group A or $\tilde{A}$ to reduce the Taylor approximation of variance. An extra observation from a group decreases the part of the total variance that comes from that group. Hence, the agent selects s_t such that:

$$\begin{aligned} \min_{s_t \in \{A, \tilde{A}\}} & \left[\sum_{i \in A} (f_i'(\hat{\boldsymbol{x}})^2 \hat{x}_i (P(A) - \hat{x}_i) - 2 \sum_{j \in A : j > i} f_i'(\hat{\boldsymbol{x}}) f_j'(\hat{\boldsymbol{x}}) \hat{x}_i \hat{x}_j) \right] \frac{1}{n_A + \alpha_A} \left(\frac{n_A + \alpha_A}{n_A + \alpha_A + 1} \right)^{\mathbb{1}_{s_t = A}} \\ & + \left[\sum_{i \in \tilde{A}} (f_i'(\hat{\boldsymbol{x}})^2 \hat{x}_i (P(\tilde{A}) - \hat{x}_i) - 2 \sum_{j \in \tilde{A} : j > i} f_i'(\hat{\boldsymbol{x}}) f_j'(\hat{\boldsymbol{x}}) \hat{x}_i \hat{x}_j) \right] \frac{1}{n_{\tilde{A}} + \alpha_{\tilde{A}}} \left(\frac{n_{\tilde{A}} + \alpha_{\tilde{A}}}{n_{\tilde{A}} + \alpha_{\tilde{A}} + 1} \right)^{\mathbb{1}_{s_t = \tilde{A}}} \end{aligned} \tag{8}$$

The two simplifications may affect the solution in a finite horizon, but they do not affect the limit. Let $S_t(A)$ be the share of group A in the observations sampled up to time t in the solution to the approximate problem (8). Similarly, let $S_t^*(A)$ be the share of group A in the observations sampled up to time t in the solution to the exact problem (5). These shares converge to the same limit $S(A)$, which minimizes asymptotic variance given realized population shares $\boldsymbol{x}^*$:

$$\begin{aligned} AVar(f(\boldsymbol{x}) \mid \boldsymbol{x}^*) = \min_{S(A)} \quad & \frac{1}{S(A)} \left[\sum_{i \in A} (f_i'(\boldsymbol{x}^*)^2 x_i^* (P(A) - x_i^*) - 2 \sum_{i,j \in A : j > i} f_i'(\boldsymbol{x}^*) f_j'(\boldsymbol{x}) x_i^* x_j^*) \right] \\ & + \frac{1}{1 - S(A)} \left[\sum_{i \in \tilde{A}} (f_i'(\boldsymbol{x}^*)^2 x_i^* (P(\tilde{A}) - x_i^*) - 2 \sum_{i,j \in \tilde{A} : j > i} f_i'(\boldsymbol{x}^*) f_j'(\boldsymbol{x}^*) x_i^* x_j^*) \right] \end{aligned} \tag{9}$$

Theorem 1. *The sample compositions under the approximate and exact strategies converge in probability to the same limit, which minimizes asymptotic variance,* $S_t(A) \xrightarrow[N \to \infty]{p} S(A)$*,* $S_t^*(A) \xrightarrow[N \to \infty]{p} S(A)$*. The rate of convergence is* $\sqrt{N}$*.*

The proof is in Appendix A.

In the limit, our two simplifications do not affect the strategy. First, the estimate $\hat{\boldsymbol{x}}$ converges to the constant true value as the sample size grows, so treating the estimate as constant does not change the strategy. Second, the Taylor approximation keeps only those variance terms that decrease at the rate of $\frac{1}{N}$ and ignores smaller-order terms, which disappear in the limit.

For some statistics f, we can calculate the variance directly for any sample and solve the exact problem (5). We use these examples to test the performance of the approximation in finite samples with simulations. Specifically, we compare the approximate strategy to a strategy that accounts for the updating in $\hat{\boldsymbol{x}}$ and minimizes the exact variance for three examples of statistics, for which we can calculate the exact variance. We report simulations in Appendix B that demonstrate that the two strategies are almost identical even in a finite sample.

3.3 Discussion of Modeling Assumptions

Indexing as Associative Recall. Indexing can be interpreted as a form of associative recall (Kahana (2012)). For example, assume that the agent is an expert on gender and needs to estimate the differences in the propensity of men and women to pursue a career in science. Models of cued recall assume that the mention of "gender" in the problem instance serves as the cue that induces agents to selectively recall memories based on some similarity measure. The expert in our model will sample memories by gender in a way consistent with the common functional descriptions of cued recall in cognitive psychology and some recent papers in economics Bordalo et al. (2023a). Our model pins down an exact functional form of the similarity measure. We formally explore this connection to models of cued recall in Section 8.1.

Introspective versus Explicit Sampling. We describe our model in terms of *introspective* sampling from memory. An interesting alternative interpretation is to view it as a model of a market research agency that estimates a given statistic $f(\boldsymbol{x})$ for a client firm. The agency can selectively invite potential consumers for an interview according to a set of predefined criteria (such as gender). Each invitation has a fixed cost, and the agency wants to provide the most accurate prediction to the client firm at the lowest sampling cost. In this interpretation, the agency *explicitly* samples from a large set of consumers.

Memory Database. We treat sampling from the memory database $\mathcal{D}$ as drawing iid observations with each subgroup i having probability x_i. While the actual memory databases are finite, we deliberately abstract away from limited memory capacity. This model focuses on the constraints of memory retrieval rather than storage. Therefore, we think of the database as being sufficiently large so that issues of finite memory can be ignored.

Dirichlet Priors. Throughout most of the paper, we assume that the agent has a flat prior over group share realizations. We model this assumption using the flat symmetric Dirichlet distribution, $\boldsymbol{x} \sim Dir(1, ..., 1)$. The Dirichlet distribution ensures that $x_i > 0$ for $i = 1, ..., 2^K$, and $\sum_{i=1}^{2^K} x_i = 1$.

The prior easily generalizes to a non-flat Dirichlet distribution, $\boldsymbol{x} \sim Dir(\alpha_1, ..\alpha_{2^K})$. Given the Dirichlet parameter vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{2^K}) > 0$, the expectation and variance of the share of each group are:

$$E(x_i) = \frac{\alpha_i}{\sum_j \alpha_j},\ Var(x_i) = \frac{\alpha_i(\sum_j \alpha_j - \alpha_i)}{(\sum_j \alpha_j)^2(\sum_j \alpha_j + 1)} \tag{10}$$

Thus, the non-flat Dirichlet prior captures two additional features. First, subgroups may have different shares in expectation: the higher the α_i (holding α_j for all $j \neq i$ constant), the higher the subgroup share. Second, the agent may have different degrees of certainty about the subgroup shares: scaling all α_i by a factor greater than 1 decreases variance and thus increases certainty about the subgroup shares.

The non-flat Dirichlet prior has two interpretations. First, the agent may have some prior knowledge about the distribution of subgroup shares. This knowledge can be directly captured in parameters $\alpha_1, ..., \alpha_{2^K}$. Second, the agent may have immediate access to some observations from the database before sampling extra observations. Suppose the agent has a flat prior $Dir(1, ..., 1)$ but observes a random sample consisting of n_i observations from subgroups $i = 1, ..., 2^K$. Then the agent's belief is $\boldsymbol{x} \sim Dir(1 + n_1, ..., 1 + n_{2^K})$.

Single versus Multi-Period Optimization. We assume that the expert samples from her index to minimize the expected loss in the next period. At this point, we only conjecture that her sampling would be the same in most cases if she were optimizing multiple periods ahead.

4 Sampling Distortions

We now characterize sampling strategies and answer the question of *who* comes to mind. Since non-experts do not have an index, they randomly select anecdotes from memory. For example, these agents will sample men and women according to the share of both groups in the population.

The sampling strategies of experts, on the other hand, depend on the problem they need to solve. We will show, for example, that an expert with a gender index will generically over- or undersample men and women. We start by defining a class of problems that we call *simple problems* that allow us to characterize this sampling distortion efficiently.

4.1 Simple Problems

Consider an index $Ind = \left\{A, \tilde{A}\right\}$. Let $I_A \subseteq A$ and $I_{\tilde{A}} \subseteq \tilde{A}$ be subsets of the indexed groups. Let $I'_A \subseteq I_A$ and $I'_{\tilde{A}} \subseteq I_{\tilde{A}}$ be further subsets of those subsets. Let $P(I'_A \mid I_A)$ and $P(I'_A \mid I_A)$ be the shares of these smaller subsets conditional on the bigger subsets:

$$P(I'_g \mid I_g) = \frac{\sum_{i \in I'_g} x_i}{\sum_{i \in I_g} x_i} \text{ for } g = A, \tilde{A} \tag{11}$$

Simple problems depend only on these aggregate shares and not on the shares of individual subgroups x_i.

Definition 2. *Given an index $Ind = \left\{A, \tilde{A}\right\}$, a statistic $f(\boldsymbol{x})$ is a simple problem if it can be written as $\tilde{f}(P(I'_A \mid I_A), P(I'_{\tilde{A}} \mid I_{\tilde{A}}))$ where $I'_g \subseteq I_g \subseteq g$, and $P(I'_g \mid I_g) = \frac{\sum_{i \in I'_g} x_i}{\sum_{i \in I_g} x_i}$ for $g = A, \tilde{A}$.*

To illustrate the model, consider an example of a memory database that consists of people with three binary attributes (see Table 1): Gender (Male or Female)[5], Occupation (Scientist or Non-Scientist), and Hobby (Runner or Non-Runner). Therefore, there are $2^3 = 8$ distinct subgroups of observations, with shares $\boldsymbol{x} = x_1, ..., x_8$, $x_i \geq 0$, and $\sum_{i=1}^{8} x_i = 1$. The agent does not know the realization of $\boldsymbol{x}$ and has a prior $\boldsymbol{x} \sim Dir(\alpha_1, ..., \alpha_8)$.

Table 1: Database composition with three binary attributes (gender, occupation, hobby)

	F		M	
	R	NR	R	NR
S	x_1	x_2	x_5	x_6
NS	x_3	x_4	x_7	x_8

$\boldsymbol{x} \sim Dir(\alpha_1, ..., \alpha_8)$

There are three binary attributes $\{\text{Female}, \text{Male}\} \times \{\text{Scientist}, \text{Non-Scientist}\} \times \{\text{Runner}, \text{Non-Runner}\}$ and therefore $2^3 = 8$ subgroups.

Table 2: Examples of simple problems

	Importance	Variability	Availability
P(Scientist | Female) − P(Scientist | Male) *Conditional Full-Index*	No	Yes	No
P(Runner | Female, Scientist) − P(Runner | Male, Scientist) *Conditional Partial-Index*	No	Yes	Yes
P(Scientist) *Unconditional*	Yes	Yes	No

Suppose the agent is a gender expert, so she has a gender index $Ind = \{F, M\}$. Table 2 lists three examples of simple problems. The first example is a *conditional full-index problem*, which is a function of conditional estimates about certain traits among men and women, such as the difference in the propensity of women and men to become scientists. The second example is a *conditional partial-index problem* which conditions on subsets of women and men such as female and male scientists, respectively. The third example is an *unconditional* problem, which we can decompose as follows:

$$P(\text{Scientist}) = P(\text{Scientist} \mid \text{Female})P(\text{Female}) + P(\text{Scientist} \mid \text{Male})P(\text{Male}) \tag{12}$$

However, some problems are not simple. Most notably, problems conditioning on a non-indexed attribute are not simple. For example, P(Runner | Scientist) is not simple because it conditions on being a Scientist, which combines both genders.

The general solution to the agent's problem and other results are independent of the problem being simple. The main advantage of simple problems is that they allow us to separate and interpret different forces that affect sampling distortions, which we formalize in Theorem 2. These effects are also present in the optimal strategies for non-simple problems. For non-simple problems, however,

[5] We acknowledge that while for illustration purposes, we treat gender as binary, in reality, it is a diverse aspect of human identity.

the interactions between these forces do not allow us to isolate them.

4.2 Characterizing Sampling Distortions

To keep the exposition simple, we will focus on the gender index when stating results. The following theorem describes the optimal shares of male and female groups in the sample in the limit as the number of retrieved observations goes to infinity. It compares the sampling strategy to a natural and useful benchmark — equal-share sampling. We say that the agent *oversamples* one group if she samples it more than the other.

Theorem 2. *Consider a simple problem. The ratio of sampled women to men $S(F)/S(M)$ is increasing in the following three effects:*

1. *relative importance $\left|\frac{\partial\tilde{f}}{\partial P(I'_F|I_F)}/\frac{\partial\tilde{f}}{\partial P(I'_M|I_M)}\right|$ — the subgroup that affects the problem more intensely is sampled more*
2. *relative variability $\frac{P(I'_F|I_F)(1-P(I'_F|I_F))}{P(I'_M|I_M)(1-P(I'_M|I_M))}$ — the subgroup whose share is closer to $\frac{1}{2}$ is sampled more*
3. *relative availability $P(I_M \mid M)/P(I_F \mid F)$ — the subgroup that is rarer is sampled more*

The ratio of sample shares in the limit is:

$$\frac{S(F)}{S(M)} = \left|\frac{\partial\tilde{f}}{\partial P(I'_F \mid I_F)}/\frac{\partial\tilde{f}}{\partial P(I'_M \mid I_M)}\right|\sqrt{\frac{P(I'_F \mid I_F)(1-P(I'_F \mid I_F))}{P(I'_M \mid I_M)(1-P(I'_M \mid I_M))}\frac{P(I_M \mid M)}{P(I_F \mid F)}} \tag{13}$$

The proof is in Appendix C.

Theorem 2 characterizes the sample in the limit. In a finite horizon, the agent's sample is approximated by the same equation (13) but with her current estimates of probabilities instead of their true values. The agent's belief distortions may amplify the three effects in a finite sample.

We can now apply theorem 2 to the three simple problems listed in Table 2. For the conditional full-index example, we obtain the following sampling share:

$$\frac{S(F)}{S(M)} = \underbrace{\sqrt{\frac{P(\text{Scientist} \mid \text{Female})(1-P(\text{Scientist} \mid \text{Female}))}{P(\text{Scientist} \mid \text{Male})(1-P(\text{Scientist} \mid \text{Male}))}}}_{\text{variability effect}} \tag{14}$$

We can see that sampling is only driven by the variability effect – the importance effect is equal to 1 because the two subsets enter the question with equal weights. The availability effect is also equal to 1 because $I_F = F$ and $I_M = M$. It is, therefore, optimal to sample more from the gender whose share of scientists is closer to $\frac{1}{2}$.

We next consider the partial index problem:

$$\frac{S(F)}{S(M)} = \underbrace{\sqrt{\frac{P(\text{Runner} \mid \text{Female, Scientist})(1 - P(\text{Runner} \mid \text{Female, Scientist}))}{P(\text{Runner} \mid \text{Male, Scientist})(1 - P(\text{Runner} \mid \text{Male, Scientist}))}}}_{\text{variability effect}} \times \underbrace{\sqrt{\frac{P(\text{Male, Scientist} \mid \text{Male})}{P(\text{Female, Scientist} \mid \text{Female})}}}_{\text{availability effect}} \tag{15}$$

Sampling is now also affected by availability: if women are less likely to be scientists such that $P(\text{Female, Scientist} \mid \text{Female}) < P(\text{Male, Scientist} \mid \text{Male})$ then the agent will attempt to compensate for this by devoting more attention to retrieving women from her database. However, the agent compensates for the relative availability of female scientists only partially — she still retrieves fewer observations of them than of male scientists.

Corollary 1. *The availability effect leads to the agent oversampling the group with the rarer subgroup, but the rarer subgroup is still undersampled. Assuming $P(I_F \mid F) < P(I_M \mid M)$ and the importance and variability effects are absent, the ratio of subgroup shares in the sample is:*

$$\frac{S(I_F)}{S(I_M)} = \frac{S(F)}{S(M)} \frac{P(I_F \mid F)}{P(I_M \mid M)} = \sqrt{\frac{P(I_F \mid F)}{P(I_M \mid M)}} < 1 \tag{16}$$

This corollary highlights the trade-off when searching for rare observations. On the one hand, the rare observations are useful because the agent has little data on the rare subgroup. On the other hand, rare observations are less available, and searching for them is costly because the agent foregoes observations from the other groups. As a result, the agent oversamples the group that contains the rare subgroup, but not to the full extent. The rare subgroup is ultimately undersampled.

The availability effect is closely related to the availability heuristic (Tversky and Kahneman (1973)). It is optimal for the agent to oversample observations that are relatively easier to access. As we show in the next section, this leads to the agent's beliefs being driven more by the information on the subgroup that is more available. Even though the agent does not overestimate the share of the more available observations, they contribute more to her beliefs than the less available observations.

Finally, we consider sampling for the unconditional problem:

$$\frac{S(F)}{S(M)} = \underbrace{\frac{P(\text{Female})}{P(\text{Male})}}_{\text{importance effect}} \times \underbrace{\sqrt{\frac{P(\text{Scientist} \mid \text{Female})(1 - P(\text{Scientist} \mid \text{Female}))}{P(\text{Scientist} \mid \text{Male})(1 - P(\text{Scientist} \mid \text{Male}))}}}_{\text{variability effect}} \tag{17}$$

For this problem, the importance effect is not equal to 1, but it is equal to the ratio of gender population shares $P(\text{Female})/P(\text{Male})$. The larger the gender's share, the more important it is for answering the question about the whole population and the more it is being sampled.

Note that a truly gender-blind person, or a person with no gender index, samples women and men by their population shares. Population-share sampling is an alternative benchmark in addition to the equal-share benchmark. For the agent with an index, it is generically optimal to sample in different proportions, not by population or equal shares. The degree and direction of oversampling depends on the problem. To an outside observer, oversampling may look like a bias in favor of one group. However, our model shows that oversampling can be rational and result from an optimization problem.

5 Belief Distortions

In the previous section, we have shown that an expert will generically oversample one of the indexed groups. In the long term, oversampling does not matter because a fully Bayesian agent (an expert and a non-expert) will have correct beliefs as the number of retrieved observations goes to infinity. However, in the medium run, oversampling will give rise to what might look to an outside observer as a bias: the agent will have more accurate beliefs about the group that she oversamples, and her beliefs about the undersampled group will be more heavily affected by her priors.

5.1 Distortion

Although the agent uses an unbiased estimator given her beliefs, the expectation of the estimate is different from the actual realization of subgroup shares. We will refer to this difference as *distortion.*

Definition 3. *The* ***distortion*** *of an estimator $\widehat{f(\boldsymbol{x})}$ given realized $\boldsymbol{x}^*$ and sample sizes from two index groups n_A, $n_{\tilde{A}}$ is*

$$Dist(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}) = \mathbb{E}(\widehat{f(\boldsymbol{x})} - f(\boldsymbol{x}^*) \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}) \tag{18}$$

where the expectation is taken over the distribution of sampled observations conditional on $\boldsymbol{x}^*$, n_A, and $n_{\tilde{A}}$.

The distortion arises because the agent combines data from observations with the prior. While the observations are informative about the realized value of $f(\boldsymbol{x}^*)$, the prior generally distorts the estimate away from the realized $f(\boldsymbol{x}^*)$. For example, consider the distortion for one subgroup x_i:

$$\begin{aligned}
Dist(\hat{x}_i \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}) &= \mathbb{E}(\hat{x}_i - x_i^* \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}) &(19)\\
&= \mathbb{E}\left(P(g(i))\frac{n_i + \alpha_i}{n_{g(i)} + \alpha_{g(i)}} - x_i^* \middle| \boldsymbol{x}^*, n_A, n_{\tilde{A}}\right) &(20)\\
&= \frac{\alpha_{g(i)}}{n_{g(i)} + \alpha_{g(i)}}\left(P(g(i))\frac{\alpha_i}{\alpha_{g(i)}} - x_i^*\right) &(21)\\
&= \frac{\alpha_{g(i)}}{n_{g(i)} + \alpha_{g(i)}}\left(x_i^{prior} - x_i^*\right) &(22)
\end{aligned}$$

The distortion arises from the difference between the realized value x_i^* and its expected value based on the prior, x_i^{prior}. The more observations the agent retrieves, the more weight she puts on the data-driven component of the estimate and the less weight she puts on the prior. As the sample grows, the distortion decreases and disappears in the limit.

While we can calculate the distortion for estimators of individual group shares x_i, there is no closed-form expression for estimators of arbitrary problems $f(\boldsymbol{x})$. As in the previous section, we calculate the Taylor approximation of the distortion instead. To keep statements precise, we formulate results about the *asymptotic distortion*, $ADist(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*) = plim_{N \to \infty} N \cdot Dist(\hat{x}_i \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}})$, because the Taylor approximation of the distortion scaled up by N converges to its true value in the limit.

The distortion for a general problem can be decomposed into two components corresponding to the two index groups. For example, with the gender index, the two components are the distortions about women and men. Each component decreases in the number of observations sampled from the corresponding gender.

Theorem 3. *The agent's beliefs are distorted after sampling a finite number of observations. For linear problems*[6]*, the female and male components are distorted toward their prior, and the absolute values of distortion components decrease in sample size. For general problems, the ratio of the asymptotic distortion components for women and men, in absolute value, is proportional to the ratio of sampled men to women $S(M)/S(F)$.*

Proof is in Appendix D.

The agent's ability to optimize sampling leads to *rational* and *persistent stereotypes*. As we show in Theorem 2, different forces make it optimal for the agent to sample one gender more than the other. If the agent oversamples men, she has precise and observation-driven beliefs about men and imprecise and prior-driven beliefs about women. Although the agent looks like she has stereotypes about women and is unwilling to correct them, this strategy and the corresponding outcome are optimal for the agent. Furthermore, since the agent undersamples women, it takes longer for her to unlearn any stereotypes about women, making stereotypes more persistent.

The agent's distortions are also problem-dependent. The statistical problem f affects the optimal sampling strategy $S(F)/S(M)$, which in turn affects the resulting distortions. As a result, the agent's belief may be relatively more distorted about women than men for some problems and more distorted about men than women for other problems.

5.2 Non-monotonic Distortion

Theorem 3 states that the distortions of beliefs about men and women are decreasing in sample size in absolute value. However, the aggregate distortion may be non-monotonic if the agent learns about the two genders at different rates.

[6] A problem $f(\boldsymbol{x})$ is linear if f is a linear function. For the three simple problems listed in Table 2, the full-index and unconditional problems are linear, and the partial-index problem is not linear.

Table 3: Parameters in simulations

(a) Shares conditional on gender

	F		M	
	R	NR	R	NR
S	0.2	0.1	0.5	0.3
NS	0.4	0.3	0.1	0.1

(b) Unconditional shares

	F		M	
	R	NR	R	NR
S	0.09	0.045	0.275	0.165
NS	0.18	0.135	0.055	0.055

Corollary 2. *The agent's distortion may be non-monotonic in sample size.*

The aggregate distortion consists of the distortions about men and women. The two components of the aggregate distortion may balance each other out, at least partially. As the sample size grows, the agent reduces her distortion about both components. If one of the distortions decreases faster than the other, they may stop balancing each other out, and the aggregate distortion may increase.

Consider an agent with a general prior $Dir(\alpha_1, ..., \alpha_{2K})$ and a linear problem $f(\boldsymbol{x})$. Appendix D shows that the distortion can be written as a weighted sum of the prior distortions about the two genders:

$$Dist(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, n_F, n_M) = \frac{\alpha_F}{n_F + \alpha_F} Dist_F(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, 0, 0) + \frac{\alpha_M}{n_M + \alpha_M} Dist_M(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, 0, 0) \tag{23}$$

Suppose the prior distortions are equal but have opposite signs: $Dist_F(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, 0, 0) = -Dist_M(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, 0, 0)$. Then, the aggregate distortion is proportional to:

$$Dist(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, n_F, n_M) \propto \left(\frac{\alpha_F}{n_F + \alpha_F} - \frac{\alpha_M}{n_M + \alpha_M} \right) \tag{24}$$

The agent may learn about the two genders at different rates for two reasons. First, the agent may be (optimally) sampling one gender more often than the other. For example, suppose $n_F = \frac{2}{3}N$, $n_M = \frac{1}{3}N$, but $\alpha_F = \alpha_M$. Second, the agent may be more certain in her prior about one gender than the other. For example, suppose $\alpha_M = 2\alpha_F$ but $n_F = n_M = \frac{N}{2}$. Both reasons generate non-monotonic aggregate distortion. The initial aggregate distortion is zero. As sample size N increases, the aggregate distortion increases, reaches its peak, and decreases toward zero.

5.3 Simulations

We run simulations to illustrate the results' magnitudes for the three problems introduced in Table 2 of the previous section. First, we fix the population shares of different groups at levels specified in Table 3. We assume the agent has a flat prior and the gender index and hence knows the population gender shares: $P(F) = 0.45$, $P(M) = 0.55$[7].

[7] We introduce slight imbalance in the gender shares to illustrate the importance effect in the unconditional problem.

For each of the three questions, we run 10,000 simulation rounds, each consisting of 100 periods (i.e., retrieved observations). In each period, we choose whether to sample a man or a woman to minimize the expected next-period variance (calculated using Taylor approximation) based on the current sample. Given the choice of whom to sample, we draw a random observation of the corresponding gender according to the population shares. The variation between simulation rounds comes from the randomness of the drawn observations.

Figure 2 shows how the simulated share of men in the sample changes over time for the three problems. In all three cases, the share converges to a constant. For the full-index problem in panel a), only the variability effect determines sampling. The probability of being a Runner is 0.6 for both genders, so the variability effect equals 1. As a result, the agent samples men and women in equal shares in the limit. For the unconditional problem in panel c), sampling is also influenced by the importance effect, i.e., by population shares of men and women. Since men constitute 55% of the population, the agent optimally accounts for their importance and samples men proportionally. For the partial-index problem in panel b), sampling is determined by the availability and variability effects. The problem is about scientists, who are rarer among women than men. This effect pushes the share of men in the sample down.

Figure 3 shows the simulated evolution of distortion over time by gender. Specifically, each line plots the distance between the agent's belief and the true value of the female and male components of each problem in percentages of the true value. For all three problems, distortion decreases as the sample grows. After $N = 100$ observations, the relative distortion is between 8% and 12%, depending on the problem and gender. Consistent with intuition, the evolution of distortion is the same for both genders in the full-index problem. For the unconditional problem, distortion about men decreases faster because the agent samples more men. For the partial-index problem, although the agent samples more women, she retrieves fewer observations of women who are scientists than men who are scientists. Because there are more men in the 'effective' sample, distortion about men decreases faster.

Figure 4 shows the simulated variance evolution of the agent's beliefs about the male and female components of the three problems. After $N = 100$ observations, the variance approaches zero. For the full-index problem, there is no difference between genders. For the unconditional and partial-index problems, the difference in variance between genders reflects the difference in parameters and the agent's sampling strategy.

6 Exogenous Indexing

Different people may have different indices. This can be caused by exogenous factors such as growing up in a particular place and time or a history of previous beliefs the agent had to form. It can also result from the agent endogenously choosing an index from the outset. This section focuses on exogenous indices and illustrates how they can explain various biases and stereotypes.

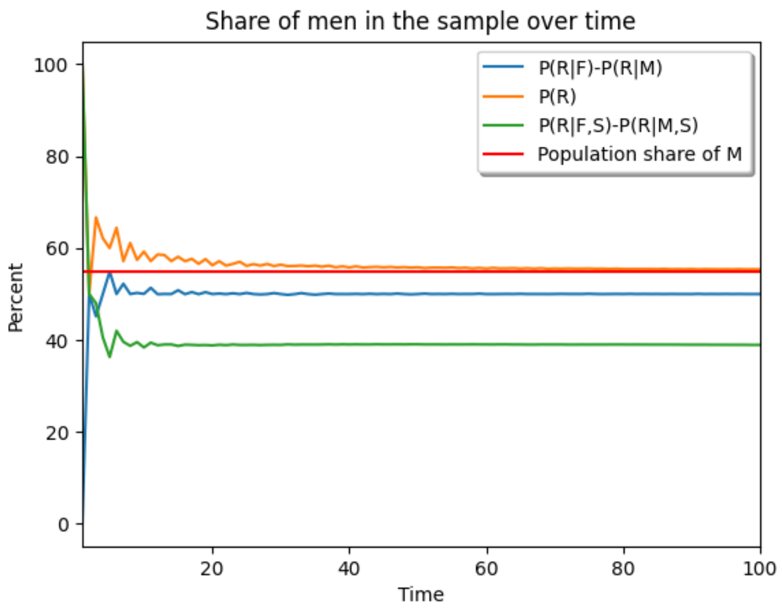
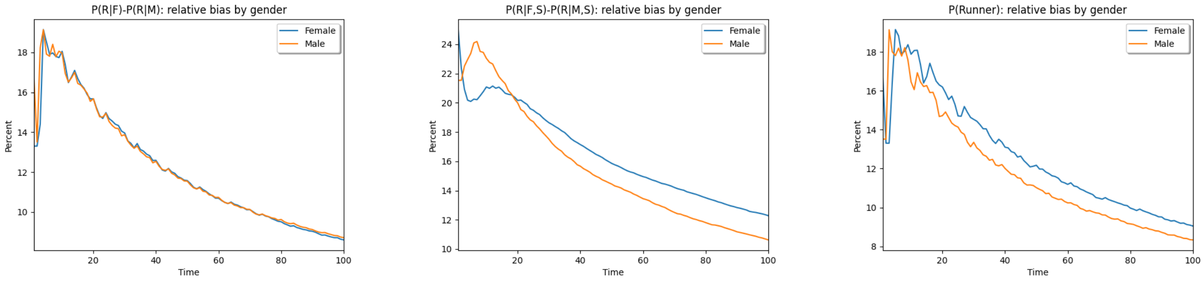


Figure 2: Simulated share of men in the sample

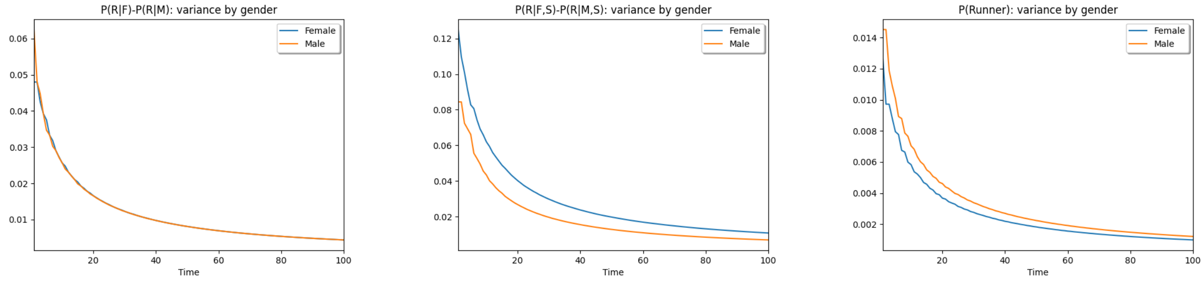


(a) Conditional Full-Index

(b) Conditional Partial-Index

(c) Unconditional

Figure 3: Simulated relative distortion over time by gender for the three problems



(a) Conditional Full-Index

(b) Conditional Partial-Index

(c) Unconditional

Figure 4: Simulated variance over time by gender for the three problems

6.1 Optimism vs Pessimism

Why are some people optimistic and others pessimistic? Our model explains this difference in views through the lens of distorted beliefs due to memory imperfection.

Suppose the world can be in one of three equally likely states: Good, Neutral, and Bad. The outcome of some event of interest can be High or Low. The agent's memory consists of observations of the state and the outcome[8]. There are two types of agents. Optimists have an index for the Good state: $Ind_O = \{\text{Good}, \text{Not Good}\}$, where Not Good includes Bad and Neutral. Pessimists, in contrast, have an index for the Bad state: $Ind_P = \{\text{Bad}, \text{Not Bad}\}$, where Not Bad includes Good and Neutral. The difference in indices could reflect, for example, different life experiences or different world views.

The agent is deciding whether to take a risky action. The expected payoff from the action is positive if the High outcome is sufficiently likely. We assume that the agent is indifferent about the action in the Neutral state, so she needs to estimate the chance of the High outcome in the non-Neutral states:

$$f(\boldsymbol{x}) = P(\text{High} \mid \text{non-Neutral}) = \frac{1}{2}P(\text{High} \mid \text{Good}) + \frac{1}{2}P(\text{High} \mid \text{Bad}) \tag{25}$$

Assume, for symmetry, that $P(\text{High} \mid \text{Good}) = 1 - P(\text{High} \mid \text{Bad}) > 0.5$. The agent starts with an uninformative prior.

Proposition 1. *Optimists overestimate the chance of High returns. Pessimists underestimate it.*

Proof is in Appendix E.

This result is ultimately driven by the availability effect, which causes the agent to oversample some of the observations depending on her index. Consider an Optimist agent with an index for the Good state. The agent can directly retrieve observations of the Good state but cannot directly target observations of the Bad state. Instead, she can only target non-Good observations, which yield observations of the Bad state with a probability of less than 1. From Corollary 1, the availability effect pushes the agent to oversample the non-Good state but not enough to fully compensate for the availability of the Good state. As a result, the agent ends up with fewer observations of the Bad state than of the Good state.

Because the agent has more observations of the Good state than the Bad state, her beliefs are driven by data more for the Good state and are driven by the prior more for the Bad state. The agent starts with a flat prior, so her prior for the High outcome in both states is 0.5. Therefore, her belief is distorted downward about the chance of the High outcome in the Good state and upward about the chance of the High outcome in the Bad state. The Optimists have more observations of the Good state than the Bad state, so they correct the negative distortion about the Good state faster than the positive distortion about the Bad state. As a result, the total distortion is positive, and the Optimists overestimate the aggregate chance of the High outcome.

[8] This setup slightly deviates from the baseline because the state is not binary but has three values. This extension is straightforward if the non-binary attribute is indexed and the agent knows the shares of all its values.

One example of this is the "depression babies" effect. Malmendier and Nagel (2011) show that people who experienced low stock market returns during their young age ("depression babies") are less likely to participate in the stock market and are more pessimistic about future stock returns. To put this story in our context, depression babies have an index for the Bad state of the economy, which is driven by their experiences in their formative years. Having the index for the Bad state makes them more pessimistic about the stock market performance and deters them from investing.

6.2 Generalization Stereotypes

Stereotyping may be a belief that two different groups are similar. Such stereotyping arises naturally in the current model and depends on the coarseness of the agent's index.

Suppose the indexed attribute has more than two values. Consider two agents with indices of different coarseness. Agent C has a coarse index $Ind_C = \{A, \tilde{A}\}$. Agent F has a finer index $Ind_F = \{A_1, ..., A_L, \tilde{A}_1, ..., \tilde{A}_{\tilde{L}}\}$, where $A_1 \cup ... \cup A_L = A$ and $\tilde{A}_1 \cup ... \cup \tilde{A}_{\tilde{L}} = \tilde{A}$[9]. We assume that both agents know the population shares $P(g)$ for $g = A_1, ...A_L, \tilde{A}_1, ..., \tilde{A}_{\tilde{L}}$, so the only difference between the agents is their ability to target memory retrieval. Agent F can target retrieval more precisely, which leads to a lower posterior variance of her belief.

Proposition 2. *The agent with a finer index has a weekly lower asymptotic variance for any question. The tight bounds on the ratio of asymptotic variances are* $\frac{AVar_F(f(\boldsymbol{x}))}{AVar_C(f(\boldsymbol{x}))} \in [\min\{P(A_1 \mid A), ..., P(A_L \mid A), P(\tilde{A}_1 \mid \tilde{A}), ..., P(\tilde{A}_{\tilde{L}} \mid \tilde{A})\}, 1]$.

Proof is in Appendix F.

Agent F can target her sampling strategy more finely than agent C. In the extreme case, this allows agent F to retrieve only useful observations, while agent C retrieves useful observations mixed with irrelevant ones. However, the share of useful observations agent C retrieves is bounded from below. For example, if agent C needs to retrieve observations from group A_1, she can target its superset — group A, for which she has an index. Then, she would retrieve observations from A_1 with probability $P(A_1 \mid A)$. This strategy imposes a bound on the ratio of useful observations agents F and C retrieve. The bound on the ratio of useful samples implies a bound on the ratio of the accuracy of agents' beliefs, measured by asymptotic variance.

The following example illustrates this result, demonstrates that the bound is tight, and shows how a coarse index exacerbates stereotyping. Suppose each person has two attributes: nationality and profession. Suppose there are three nationalities: American, German, and French, and two professions: Economists and Non-Economists.

To illustrate the effects of a coarse index, compare two agents with different coarseness of the index. Suppose an American agent has a coarse nationality index $Ind_A = \{\text{American}, \text{Non-American}\}$. A European agent, in contrast, has a finer nationality index $Ind_E = \{\text{American}, \text{German}, \text{French}\}$. This difference in coarseness could result from different life experiences. If an American rarely

[9] This setup also deviates from the baseline because the indexed attribute has multiple values, and the index is non-binary. This generalization is straightforward and does not affect previous results.

interacts with Europeans or solves problems that require thinking about Europe, she is likely to have a more coarse index than a European. Let $P(G \mid NA)$ be the share of Germans among Non-Americans. Suppose both agents receive a question about Germans: what is the share of Economists among Germans, $f(\boldsymbol{x}) = P(E \mid G)$.

Corollary 3. *A European agent has a more accurate belief about Germans than an American agent:* $\frac{AVar_E(P(E|G))}{AVar_A(P(E|G))} = P(G \mid NA)$*. Furthermore, the American agent's beliefs about Germans are more distorted towards the prior than the European agent's:* $\left|\frac{Dist_E(\widehat{P(E|G)}|\boldsymbol{x}^*)}{Dist_A(\widehat{P(E|G)}|\boldsymbol{x}^*)}\right| = P(G \mid NA)$*. Under the flat prior, the American agent's beliefs about Germans and French are closer than the European agent's.*

To form more precise beliefs about Germans, the agents need to retrieve observations of Germans. While the European agent can target Germans directly, the American agent can only target the Non-American group, which yields a German with probability $P(G \mid NA)$. Thus, the American agent retrieves only a $P(G \mid NA)$ share of observations of Germans compared to the European agent. The difference in the number of sampled Germans leads to the difference in the accuracy of estimates. Asymptotic variance and absolute distortion decrease inversely in the share of observations. Therefore, asymptotic variance and absolute distortion are proportionally smaller for the European agent than for the American agent. If Germans are rarer than the French, this example shows that the lower bound is tight.

The American agent treats Germans and the French more similarly than the European agent due to the distortion toward the flat prior. Suppose both agents receive a second question about the French. After sampling a finite number of observations, American and European agents' beliefs are distorted toward the common prior. However, the European agent learns about Germans and the French faster than the Americans. As a result, the American agent's beliefs about Germans and French will be more similar than the European agent's.

6.2.1 Probabilistic Indexing

One implication of Proposition 2 is that the agent may be able to strictly improve upon indexing all observations. If she indexes some of them probabilistically, she can make it easier to retrieve rare observations. Let $P_{Ind}(A_i \mid A)$ be the share of observations from subgroup A_i in the indexed group A *in the agent's index*, which could differ from the population share $P(A_i \mid A)$.

Corollary 4. *The worst-case relative penalty on the coarse index,* $\frac{AVar_F(f(\boldsymbol{x}))}{AVar_C(f(\boldsymbol{x}))}$*, is minimized by probabilistic indexing, such that the shares of subgroups in the index are equal,* $P_{Ind}(A_1 \mid A) = ... = P_{Ind}(A_L \mid A) = \frac{1}{L}$ *and* $P_{Ind}(\tilde{A}_1 \mid \tilde{A}) = ... = P_{Ind}(\tilde{A}_{\tilde{L}} \mid \tilde{A}) = \frac{1}{\tilde{L}}$.

Probabilistic indexing allows an agent to create an index with more balanced subgroups. As a result, the agent has easier access to observations with rare attribute values. Problems about rare observations drive the worst cases. Thus, an index with more balanced subgroups has a better worst-case performance.

6.3 Anchoring Stereotypes

Stereotyping can also take the form of extrapolating beliefs about one group to another group. This phenomenon also arises in our model as a rational response to the difficulty of finding rare observations. Instead of searching for rare observations, the agent can form a belief based on observations from a different but correlated group that are easier to retrieve and adjust the belief in the necessary direction. In this case, the agent's belief may not converge to the truth even in the limit, and the beliefs of two agents using two different groups may not converge to the same limit.

Suppose the indexed attribute has several values, one of which is rare, and the problem $f(\boldsymbol{x})$ is about this rare group. If the rare group does not have its own index but is correlated with an indexed group, it may be optimal for the agent to use a 'proxy index.' Instead of trying to find the rare observations, the agent can estimate an analogous statistic for the indexed group. The agent would then form a *proxy estimate* by adjusting her estimate for the indexed group toward the rare group. This approach could be a good approximation if the correlation is high.

Consider an agent with index $Ind = \{A, \tilde{A}\}$, where $A = A_1$, $\tilde{A} = \tilde{A}_1 \cup ... \cup \tilde{A}_{\tilde{L}}$. Let k_1 be a value of a non-indexed attribute K_1. Suppose the agent's problem is to estimate $f(\boldsymbol{x}) = P(k_1 \mid \tilde{A}_1)$. Suppose further that there is a chance that A_1 is identical to $\tilde{A}_1$ in dimension K_1: with probability p, $P(k_1 \mid \tilde{A}_1) = P(k_1 \mid A_1)$, and with probability $1-p$, $P(k_1 \mid \tilde{A}_1)$ and $P(k_1 \mid A_1)$ are independent. The higher the p, the more likely it is that learning about A_1 is informative about $\tilde{A}_1$.

One strategy for the agent is to sample from group $\tilde{A}$, keep observations in $\tilde{A}_1$, and thus estimate $P(k_1 \mid \tilde{A}_1)$ directly. Alternatively, the agent can sample from group A, estimate $P(k_1 \mid A_1)$, and partially adjust the estimate in the direction of her prior about $P(k_1 \mid \tilde{A}_1)$. The result would be a proxy estimate for $P(k_1 \mid \tilde{A}_1)$. Denote G_0 the prior distribution and G_N the posterior distribution after N observations.

Proposition 3. *The proxy estimate is* $p(P(k_1 \mid A_1)) + (1-p)P(k_1 \mid \tilde{A}_1)$*. Assuming* $\tilde{A}_1$ *is sufficiently rare, the proxy estimate has lower variance than the direct estimate after* N *observations if*

$$p(1-p)(\mathbb{E}_{G_0} P(k_1 \mid \tilde{A}_1) - \mathbb{E}_{G_N} P(k_1 \mid A_1))^2 < p(Var_{G_0}(P(k_1 \mid \tilde{A}_1)) - Var_{G_N}(P(k_1 \mid A_1))) \quad (26)$$

In the limit, the estimate based on the proxy index has a higher variance than the direct estimate.

Proof is in Appendix G.

When choosing between the direct and proxy estimates of $P(k_1 \mid \tilde{A}_1)$, the agent faces a trade-off between estimating $P(k_1 \mid \tilde{A}_1)$ on little data and estimating $P(k_1 \mid A_1)$, which is potentially informative, on more data. On the one hand, a proxy index allows to get an accurate estimate for $P(k_1 \mid A_1)$ with less data than what would be needed for $P(k_1 \mid \tilde{A}_1)$. The right-hand side of equation (26) captures this force. The proxy index reduces the variance by accurately estimating the case when the two subgroups are identical. On the other hand, the estimate for $P(k_1 \mid A_1)$ is not informative about $P(k_1 \mid \tilde{A}_1)$ if they are independent. The left-hand side of equation (26) captures this force after partial adjustment. Conditional on the two subgroups being independent, the proxy estimate introduces a bias because the expectations for the two subgroups are different. The balance between reducing variance and introducing a bias when the groups are different determines whether the proxy index is better than direct estimation.

However, using a proxy index is worse than direct estimation in the limit. As the sample size goes to infinity, the variance of the direct estimate goes to zero. For the proxy estimate, however, the variance does not go to zero in the limit. When the two subgroups are independent, the proxy index is uninformative about the subgroup of interest, so the variance is always strictly positive. Thus, with an infinite sample, direct estimation is better.

The use of proxy indices rationalizes the stereotyping of a small group. If it is too difficult to retrieve observations of the small group, it may be optimal to form a belief based on the observations of a similar group that is easier to retrieve. For example, suppose the agent does not have an index for Austrians but has an index for Germans. She may find it optimal to base her beliefs about Austrians on the observations of Germans and adjust them partially towards her prior about Austrians. Her beliefs about Austrians based on Germans may be accurate for those domains where the two nationalities are similar. In other domains, her beliefs may be very wrong.

One consequence of using a proxy is that beliefs do not converge to the truth even in the limit. Furthermore, if two agents use two different proxies to estimate the same statistic, their beliefs may not converge in the limit.

Corollary 5. *If an agent uses a proxy for estimation, her beliefs do not converge to the truth even in the limit. Beliefs of agents who use different proxies may not converge even in the limit.*

Continuing the previous example, while one agent may have a German index, another may have a French index. Their beliefs converge to a mixture of the truth about their indexed nationality and the prior about Austrians. Thus, their beliefs do not converge to the truth about Austrians. Furthermore, if Germans and the French are not identical, the beliefs of the two agents will diverge—one will be anchored to Germans, the other to the French.

6.4 History Dependence

If the agent answers several questions sequentially, she may store a set of recently retrieved observations and have them all in short-term memory for instant use. These observations may affect the agent's beliefs, and depending on their composition, the agent's beliefs will be distorted in different directions.

Suppose the memory database consists of women who are Scientists (S)/Non-Scientists (NS) and Good at Math (G)/Not Good at Math (NG). Suppose that $P(S) = 0.5$, $P(G) = 0.5$, but $P(G \mid NS) < 0.5 < P(G \mid S)$ — scientists are more likely to be good at math than non-scientists.

Consider two agents, S and NS, with a scientist index and a flat prior but different sets of observations in their short-term memory for historical reasons. Agent S has N_0 observations of scientists. Agent NS has N_0 observations of non-scientists. Suppose both agents receive a question about women: What is the share of women who are good at math, $P(G)$. The agents' beliefs are distorted in the direction of the observations in their short-term memory.

Proposition 4. *Agent S, who has scientists in short-term memory, has beliefs that are distorted up about $P(G)$. Agent NS, who has non-scientists in short-term memory, has beliefs that are distorted down about $P(G)$.*

Proof is in Appendix H.

The agents decompose the problem into two components — female scientists and female non-scientists — and use the available set of observations to estimate one of the components before sampling from the memory database. From the start, Agent S has accurate beliefs about female scientists, but her beliefs about female non-scientists are distorted up, i.e., toward the prior. This leads to a positive aggregate distortion of agent S. In contrast, Agent NS has accurate beliefs about female non-scientists, but her beliefs about female scientists are distorted down, i.e., toward the prior. This leads to a negative aggregate distortion of agent NS. Two agents who start with the same index and prior but focus on different observations for historical reasons can have beliefs distorted in opposite directions about the same question.

7 Optimal Indexing

So far, we assumed that the index is given exogenously. However, some indices perform better than others. If the agent had a choice, she may prefer to index one attribute over another. This section shows that good indices address frequent questions and represent informative and unbalanced attributes.

First, we introduce a general setup and then impose restrictive assumptions to highlight each factor. The memory database consists of people with four attributes: Scientist (S)/Non-Scientist (NS), Runner (R)/Non-Runner (NR), Female (F)/Male (M), Nationality ($A/\tilde{A}$). The agent receives one problem f from a set of possible problems $\mathcal{F}$.

For tractability, we assume that the agent can only have a gender index $Ind_G = \{F, M\}$ or a Nationality index $Ind_N = \{A, \tilde{A}\}$. We also restrict the set of possible problems to unconditional and conditional problems about being a Runner: $P(R)$ and $P(R \mid k)$, where $k \in \{S, NS, F, M, A, \tilde{A}\}$. In the following subsections, we introduce additional restrictive assumptions to turn off all but one factor at a time, including assumptions on the parameter realizations. The agent does not know these realizations from the beginning, but she learns them over time and adjusts her strategy accordingly. We compare the performance of the two indices given the assumptions on the parameters.

7.1 Distribution of Problems

One reason for an attribute to be a good candidate for an index is that it helps answer problems that arise frequently. For example, if most problems are conditional on gender, then having a gender index is useful because it facilitates efficient retrieval of relevant observations.

Suppose both indexable attributes are ex-post uninformative about the other attributes: conditioning on a value of the indexable attribute does not affect the distribution of values of other attributes. Suppose also they are ex-post balanced: both values of the indexable attributes are equally likely.

Definition 4. *An attribute K_i is **ex-post uninformative** if $P(k_j \mid k_i) = P(k_j)$ and $P(k_j \cap k_l \mid k_i) = P(k_j \cap k_l)$ for any $k_i \in K_i$, $k_j \in K_j$, and $k_l \in K_l$. An attribute K_i is **ex-post balanced** if $P(k_i) = P(\tilde{k}_i) = \frac{1}{2}$ for $k_i, \tilde{k}_i \in K_i$.*

The distribution of problems is the only factor determining which index is better. In particular, the only difference is in the questions that condition on one of the indexable attributes.

Proposition 5. *Assume that both indices are ex-post uninformative and balanced. The gender index has a lower expected asymptotic variance than the nationality index if and only if the probability of problems conditioning on gender is higher than the probability of problems conditioning on nationality.*

Proof is in Appendix I.

Intuitively, an index is good if it is useful to answer high-probability problems. Because of the uninformativeness and balancedness assumptions, neither index is useful for any problem except for problems that condition on the indexed attribute. Therefore, the only difference is coming from problems $P(R \mid k)$, where $k \in \{F, M, A, \tilde{A}\}$. If the agent has an index for k, she can sample a k observation each period. If the agent has a different index, this index is not useful because of the uninformativeness assumption. As a result, she can only sample a k observation with probability $P(k)$. Smaller effective sample size inflates asymptotic variance by $\frac{1}{P(k)}$ compared to the agent with the index on k:

$$\frac{AVar_G(P(R \mid k))}{AVar_N(P(R \mid k))} = \begin{cases} P(k) \text{ if } k \in \{F, M\} \\ \dfrac{1}{P(k)} \text{ if } k \in \{A, \tilde{A}\} \end{cases}$$

For clarity of comparison, we assume that gender and nationality attributes are ex-post symmetric. Therefore, the gender index is better if and only if problems conditioning on gender are more likely.

7.2 Informativeness

A good attribute for an index is informative about other attributes. An informative attribute helps answer problems not only conditioning on the attribute itself but also on other attributes and the unconditional problem. Higher informativeness rationalizes having an index based, for example, on gender or age rather than on the first letter of the name or eye color.

Suppose nationality is ex-post uninformative, while gender is fully informative about being a scientist: $P(S \mid M) = 1$, $P(S \mid F) = 0$. The nationality index is useful only for answering problems that condition on nationality. The gender index is useful for problems that condition on gender and occupation and for the unconditional problem.

Proposition 6. *Assume gender is fully informative about occupation, while nationality is ex-post uninformative. The gender index leads to a lower asymptotic variance than the nationality index for the unconditional problem and the problems conditional on the non-indexable attribute.*

Proof is in Appendix J.

First, consider a problem $P(R \mid S)$ (equivalently, $P(R \mid NS)$) — conditional on a non-indexable attribute. The nationality index is not useful because it is fully uninformative. On the other hand, the gender index is completely informative and turns the problem into estimating $P(R \mid M)$. Thus, the gender index allows the agent to sample a scientist each period, while the nationality index — only with probability $P(S)$. As a result, the gender index performs better:

$$\frac{AVar_G(P(R \mid S))}{AVar_N(P(R \mid S))} = P(S) \tag{27}$$

Second, consider the unconditional problem $P(R)$. The uninformative nationality index splits the population into two groups with identical shares of runners $P(R \mid A) = P(R \mid \tilde{A}) = P(R)$. The informative gender index splits the population into two groups: one has a higher share of runners, and the other has a lower share of runners. More extreme parameters are easier to estimate precisely: formally, asymptotic variance is concave in the share of runners. As a result, splitting the data into two different groups by gender reduces the total variance of the estimate compared to splitting into two identical groups by nationality.

7.3 Unbalancedness

Another feature of a good candidate for an index is unbalancedness. An unbalanced attribute has rare values, which are hard to retrieve without the appropriate index and, therefore, hard to learn about. This penalty on low availability makes it important to have an index for unbalanced attributes, such as minority status, rare skill, or rare nationality.

Suppose gender is more balanced than nationality: $\mid P(F) - P(M) \mid < \mid P(A) - P(\tilde{A}) \mid$. Suppose also that both attributes are ex-post uninformative. Thus, an index is useful only for problems that condition on the indexed attribute. For these problems, it is important to have an index for rare groups.

Proposition 7. *Assume both indices are ex-post uninformative, but gender is more balanced than nationality. Assume the problems conditioning on the indexable attributes are equally likely. Then, the nationality index has a lower expected asymptotic variance than the gender index.*

Proof is in Appendix K.

Because the indexed attributes are uninformative, the indices are useful only for problems that condition on the indexed attribute, i.e., $P(R \mid k)$, where $k \in \{F, M, A, \tilde{A}\}$. If the agent has the index for k, she can sample a relevant observation each period. If she has the other index, she can only sample a relevant observation with probability $P(k)$. As a result, the asymptotic variance gets inflated by $\frac{1}{P(k)}$ if the agent has the wrong index:

$$\frac{AVar_G(P(R \mid k))}{AVar_N(P(R \mid k))} = \begin{cases} P(k) \text{ if } k \in \{F, M\} \\ \dfrac{1}{P(k)} \text{ if } k \in \{A, \tilde{A}\} \end{cases} \tag{28}$$

Note that the relative penalty for the wrong index, $\frac{1}{P(k)}$, is convex. For the problems conditional on a fully balanced attribute, the average penalty is only 2. As the attribute becomes less balanced, the average penalty increases because of convexity. Intuitively, if one group is very rare, then not having an index for it is very costly. Therefore, if all problems conditional on the indexable attributes are equally likely, the index for the unbalanced attribute performs better.

8 Other Applications

8.1 Microfoundation for Cued Recall

A popular class of recall models is based on cues and the similarity between cues and memories (e.g., Kahana (2012); Bordalo et al. (2023b)). In our model, the agent's recalled sample looks like it is based on cued recall. Therefore, our model provides a microfoundation for the cued recall models and specifies the functional form of the similarity measure.

Models of cued recall assume that the probability of recalling a given observation increases in the similarity of this observation with a given cue and decreases in the similarity of other observations with the cue (e.g., Kahana (2012); Bordalo et al. (2023b)). Let d be an observation in a memory dataset D. Let c be a cue. Let $Sim(d, c)$ be the similarity between observation d and cue c. The standard assumption is that the probability of recalling observation d given cue c, $r(d, c)$, is

$$r(d, c) = \frac{Sim(d, c)}{\sum_{d' \in D} Sim(d', c)} \tag{29}$$

Our model does not impose any specific assumptions on the recall probabilities. Nevertheless, the agent's optimal recall strategy yields the same structure of the recalled sample as the models of cued recall.

It is helpful to think of recall in our model as happening in two stages. First, the agent chooses which indexed group to target — this stage is the one we are most interested in. Then, she retrieves an observation from that group uniformly at random. Suppose the agent has an index $Ind = \{A, \tilde{A}\}$ and receives a problem $f(\boldsymbol{x})$. We treat $f(\boldsymbol{x})$ as a cue for the first retrieval stage. Let observation d belong to the indexed group g. To adapt equation (29) to our setup, take the sample size to infinity. In the first stage, the agent chooses which group to target, so instead of looking at the probability of recalling a single observation d (which goes to zero), consider the share of recalled observations that belong to group g:

$$r(g,c) = \lim_{N\to\infty} \frac{\sum_{d\in g} Sim(d,c)}{\sum_{d'\in D} Sim(d',c)} \tag{30}$$

$$= \frac{P(g)Sim(g,c)}{\sum_{g'\in\{A,\tilde{A}\}} P(g')Sim(g',c)} \tag{31}$$

where $Sim(g',c) = Sim(d',c) = Sim(d'',c)$ for all observations d', d'' in indexed group g'.

The sample share of group g under cued recall coincides with the sample share of group g in the optimal sampling strategy. The following corollary is a direct consequence of the solution to problem (9), given in Appendix C.

Corollary 6. *The share of indexed group $g \in \{A, \tilde{A}\}$ in the recalled sample in the limit as $N \to \infty$ is*

$$S(g) = \frac{P(g)Sim(g, f(\boldsymbol{x}))}{\sum_{g'\in\{A,\tilde{A}\}} P(g')Sim(g', f(\boldsymbol{x}))} \tag{32}$$

where

$$Sim(g, f(\boldsymbol{x})) = \frac{1}{P(g)}\sqrt{\sum_{i\in g}(f_i'(\boldsymbol{x})^2 x_i(P(g)-x_i) - 2\sum_{i,j\in g:j>i} f_i'(\boldsymbol{x})f_j'(\boldsymbol{x})x_i x_j)} \tag{33}$$

This corollary highlights two results. First, we provide a microfoundation for the cued recall models. We show that the recalled sample under optimal memory retrieval has the same structure as in the cued recall models. Second, we derive the specific functional form of the similarity metric. Intuitively, the similarity between a group and a problem is given by the amount of variance attributed to this group. This metric pins down the definition of similarity between cues and observations, connecting it to the fundamental parameters. This similarity metric rationalizes the recalled sample based on a cued recall model.

In the second stage, the agent draws an observation from the target group uniformly at random. In the terminology of cues, this is equivalent to drawing an observation guided by the target group being the cue. The similarity measure is constant and positive for observations inside that group and zero outside. This assumption on the retrieval process within a group is the simplest and cleanest because it does not drive any of the results about distortions. Possible extensions of the model could impose other assumptions on within-group retrieval, such as associativeness and rehearsal effects. These extra assumptions would affect the resulting sample of recalled memories and lead to different model predictions.

8.2 Representativeness Stereotypes

One implication of the cued recall models is that people may form stereotypes based on the representativeness heuristic (Tversky and Kahneman (1983); Bordalo et al. (2016, 2021)). Following

Bordalo et al. (2016), we say that an attribute k is *representative* of group A if $\frac{P(k|A)}{P(k|\bar{A})}$, a measure of representativeness, is greater than 1.

The following simplified setup illustrates how representativeness affects the agent's distortions in our model. Consider our main example with three attributes and the gender index. Suppose the agent's problem is $P(\text{Runner} \mid \text{Scientist})$. To get cleaner results, analogously to the previous section, assume that being a runner is ex-post uninformative.

Proposition 8. *Suppose the problem is $P(Runner \mid Scientist)$, and being a runner is ex-post uninformative. If being a scientist is representative of men, the share of male scientists in the sample is greater than their share in the population, and it is increasing in representativeness:*

$$S(Scientist, Male) = P(Scientist, Male)\frac{\sqrt{\frac{P(Scientist|Male)}{P(Scientist|Female)}}}{P(Male)\sqrt{\frac{P(Scientist|Male)}{P(Scientist|Female)}} + P(Female)} \tag{34}$$

Proof is in Appendix L.

Even though the agent uses the optimal sampling strategy, an outside observer may think she uses the representativeness heuristic. If being a scientist is representative of men, the agent samples more male scientists (and fewer female scientists) than their population share. This strategy is optimal because it is easier to find scientists—the group of interest for the problem—among men than women, so representativeness works through the availability effect. As a consequence of representativeness, the agent's beliefs about scientists are driven more by the data on male scientists and by the prior about female scientists, compared to an agent who does not use an index for sampling.

9 Conclusion

In this paper, we proposed a model of belief formation based on memory. The agent in our model retrieves memories and combines them with the prior to form a belief. The agent is Bayesian and rational but faces a constraint on memory retrieval — she can only sample observations one at a time instead of retrieving all of them simultaneously. Retrieval is primarily random, but the agent can partially target retrieval using an index. The index splits the database of memories into two (or more) groups based on the values of one (or more) attribute. The agent chooses which indexed group to sample in each period to ensure that her beliefs are as accurate as possible.

We show that the expert will generically oversample one group and characterize the three forces determining the oversampling for simple problems. We then demonstrate that oversampling translates directly into belief distortion. We use this insight to explain well-known biases in beliefs across individuals, such as the "depression babies" effect, rational stereotypes, and the dependence of beliefs on the history of previously encountered problems.

Testing the model. A natural next step would be empirically testing our model's assumptions and implications. Our model implies that beliefs converge to the truth as the value of having correct beliefs increases. It assumes that the agent keeps retrieving observations indefinitely until an exogenous decision period. In reality, we expect people to stop retrieving observations when the marginal cost of recall is greater than the marginal benefit of improving beliefs. Thus, one can change the incentives to recall more observations by experimentally varying the stakes. If there are enough observations to be retrieved, our model predicts that beliefs should converge to the truth if the stakes are sufficiently high.

Another implication of the model is that the sampling strategy depends on the problem. The effects identified in Theorem 2 are not fixed for a given index but change from problem to problem. For example, while women may be undersampled when answering a problem about scientists, they may be oversampled when answering a problem about teachers. One can test this result by experimentally varying the beliefs participants must form, such as asking about scientists or teachers. Furthermore, Theorem 2 predicts not only that the sampling strategy should change in response to a new problem but also the specific direction of this change, which can also be tested.

One could also test the assumptions about the mechanics of the model. For example, one could induce different indices among participants in a lab experiment, elicit beliefs on full-index, partial-index, and unconditional problems, and test for the predicted sampling strategies and belief distortions.

These tests are challenging because recall is an internal process, which we can not observe. However, we can measure it indirectly in two ways. First, we can ask participants to list every observation they recall. Although imperfect, this approach makes the recall process more explicit. Second, we can indirectly test the model by measuring final beliefs, which depend on the recall strategy. A complementary approach could build on the alternative interpretation of our model from Section 3.3 where a "market research agency" explicitly samples consumers at a cost. By having experimental participants play the role of the agency and measuring their explicit sampling of different types of consumers, we could directly test the oversampling predictions of our model.

References

Afrouzi, Hassan, Spencer Y Kwon, Augustin Landier, Yueran Ma, and David Thesmar, "Overreaction in expectations: Evidence and theory," *The Quarterly Journal of Economics*, 2023, *138* (3), 1713–1764.

Azevedo, Eduardo M, Alex Deng, José Luis Montiel Olea, Justin Rao, and E Glen Weyl, "A/b testing with fat tails," *Journal of Political Economy*, 2020, *128* (12), 4614–000.

Bordalo, Pedro, Giovanni Burro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Imagining the Future: Memory, Simulation and Beliefs," Working Paper 2023.

_ , John J Conlon, Nicola Gennaioli, Spencer Y Kwon, and Andrei Shleifer, "Memory and probability," *The Quarterly Journal of Economics*, 2023, *138* (1), 265–311.

_ , Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *The Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.

_ , _ , _ , Frederik Schwerter, and Andrei Shleifer, "Memory and representativeness.," *Psychological Review*, 2021, *128* (1), 71–85.

_ , Nicola Gennaioli, and Andrei Shleifer, "Memory, attention, and choice," *The Quarterly Journal of Economics*, 2020, *135* (3), 1399–1442.

Charles, Constantin, "Memory and trading," *Available at SSRN 3759444*, 2022.

Che, Yeon-Koo and Konrad Mierendorff, "Optimal dynamic allocation of attention," *American Economic Review*, 2019, *109* (8), 2993–3029.

Dougherty, Michael RP, Charles F Gettys, and Eve E Ogden, "MINERVA-DM: A memory processes model for judgments of likelihood.," *Psychological Review*, 1999, *106* (1), 180.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann, "Associative Memory, Beliefs and Market Interactions," Working Paper 2023.

Fryer, Roland and Matthew O Jackson, "A categorical model of cognition and biased decision making," *The BE Journal of Theoretical Economics*, 2008, *8* (1).

Fudenberg, Drew, Giacomo Lanzani, and Philipp Strack, "Selective memory equilibrium," *Available at SSRN 4015313*, 2022.

Gennaioli, Nicola and Andrei Shleifer, "What comes to mind," *The Quarterly Journal of Economics*, 2010, *125* (4), 1399–1433.

Gossner, Olivier, Jakub Steiner, and Colin Stewart, "Attention please!," *Econometrica*, 2021, *89* (4), 1717–1751.

Gottlieb, Daniel, "Imperfect memory and choice under risk," *Games and Economic Behavior*, 2014, *85*, 127–158.

Graeber, Thomas, Florian Zimmermann, and Christopher Roth, "Stories, statistics, and memory," 2022.

Kahana, Michael Jacob, *Foundations of Human Memory*, Oxford University Press, 2012.

Kőszegi, Botond, George Loewenstein, and Takeshi Murooka, "Fragile self-esteem," *The Review of Economic Studies*, 2022, *89* (4), 2026–2060.

Kwon, Spencer Yongwook and Johnny Tang, "Extreme events and overreaction to news," *Available at SSRN 3724420*, 2020.

Liang, Annie, Xiaosheng Mu, and Vasilis Syrgkanis, "Dynamically Aggregating Diverse Information," *Econometrica*, 2022, *90* (1).

Malmendier, Ulrike and Jessica A Wachter, "Memory of past experiences and economic decisions," *Available at SSRN 4013583*, 2021.

_ **and Stefan Nagel**, "Depression babies: Do macroeconomic experiences affect risk taking?," *The Quarterly Journal of Economics*, 2011, *126* (1), 373–416.

Mayskaya, Tatiana, "Dynamic choice of information sources," *California Institute of Technology Social Science Working Paper, ICEF Working Paper WP9/2019/05*, 2022.

Mohlin, Erik, "Optimal categorization," *Journal of Economic Theory*, 2014, *152*, 356–381.

Mullainathan, Sendhil, "A memory-based model of bounded rationality," *The Quarterly Journal of Economics*, 2002, *117* (3), 735–774.

_ , "Thinking through categories," Working Paper 2002.

Neligh, Nathaniel Leigh, "Rational memory with decay," *Available at SSRN 4273575*, 2022.

Nilsson, Håkan, Henrik Olsson, and Peter Juslin, "The cognitive substrate of subjective probability.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2005, *31* (4), 600.

Silveira, Rava Azeredo Da, Yeji Sung, and Michael Woodford, "Optimally imprecise memory and biased forecasts," Working Paper, National Bureau of Economic Research 2020.

Tversky, Amos and Daniel Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive psychology*, 1973, *5* (2), 207–232.

_ **and** _ , "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.," *Psychological review*, 1983, *90* (4), 293–315.

Wachter, Jessica A and Michael Jacob Kahana, "A retrieved-context theory of financial decisions," Technical Report, National Bureau of Economic Research 2019.

Wilson, Andrea, "Bounded memory and biases in information processing," *Econometrica*, 2014, *82* (6), 2257–2294.

A Proof of Theorem 1

Consider a general index $Ind = \{A, \tilde{A}\}$ and prior $\boldsymbol{x} \sim Dir(\alpha_1, ..., \alpha_{2^K})$. Denote $\alpha_A = \sum_{i \in A} \alpha_i$ and $\alpha_{\tilde{A}} = \sum_{i \in \tilde{A}} \alpha_i$. Let $g(i)$ be the indexed group of subgroup i, i.e., $g(i) = A$ if $i \in A$ and $g(i) = \tilde{A}$ if $i \in \tilde{A}$. After N observations, the posterior G is also Dirichlet, $\boldsymbol{x} \sim Dir(\alpha_1 + n_1, ..., \alpha_{2^K} + n_{2^K})$. Denote subgroup shares conditional on the indexed group $\tilde{x}_i = \frac{x_i}{P(g(i))}$. As a property of the Dirichlet distribution, $(\tilde{x}_i)_{i \in g} \sim Dir((\frac{\alpha_i + n_i}{\alpha_g + n_g})_{i \in g})$ for $g = A, \tilde{A}$. The expected values of $\tilde{x}_i$ are

$$\hat{\tilde{x}}_i = \mathbb{E}_G(\tilde{x}_i) = \frac{\alpha_i + n_i}{\alpha_{g(i)} + n_{g(i)}} \tag{35}$$

By LLN, the estimates of parameters $\hat{\boldsymbol{x}}$ (for which $n_{g(i)} \to \infty$) converge to true values $\boldsymbol{x}$ in probability:

$$\begin{aligned} \hat{x}_i = P(g(i)) \frac{n_i + \alpha_i}{n_{g(i)} + \alpha_{g(i)}} &\xrightarrow{p} P(g(i)) \tilde{x}_i = x_i && (36) \\ \Rightarrow \hat{\boldsymbol{x}} &\xrightarrow{p} \boldsymbol{x} && (37) \end{aligned}$$

Assume f_i' is continuous for each $i \in \{1, ..., 2^K\}$. The estimates of asymptotic variance Taylor approximation components for the two groups, A and $\tilde{A}$, converge to their true values in probability by Mann-Wald theorem:

$$\sum_{i \in A} (f_i'(\hat{\boldsymbol{x}})^2 \hat{x}_i (P(A) - \hat{x}_i) - 2 \sum_{j \in A : j > i} f_i'(\hat{\boldsymbol{x}}) f_j'(\hat{\boldsymbol{x}}) \hat{x}_i \hat{x}_j) \xrightarrow{p} \sum_{i \in A} (f_i'(\boldsymbol{x})^2 x_i (P(A) - x_i) - 2 \sum_{j \in A : j > i}^{4} f_i'(\boldsymbol{x}) f_j'(\boldsymbol{x}) x_i x_j) \tag{38}$$

$$\sum_{i \in \tilde{A}} (f_i'(\hat{\boldsymbol{x}})^2 \hat{x}_i (P(\tilde{A}) - \hat{x}_i) - 2 \sum_{j \in \tilde{A} : j > i} f_i'(\hat{\boldsymbol{x}}) f_j'(\hat{\boldsymbol{x}}) \hat{x}_i \hat{x}_j) \xrightarrow{p} \sum_{i \in \tilde{A}} (f_i'(\boldsymbol{x})^2 x_i (P(\tilde{A}) - x_i) - 2 \sum_{j \in \tilde{A} : j > i} f_i'(\boldsymbol{x}) f_j'(\boldsymbol{x}) x_i x_j) \tag{39}$$

Let $S_t^{target}(A)$ be the "target" share of A in the sample that minimizes the current estimate of the asymptotic variance Taylor approximation:

$$\begin{aligned} S_t^{target}(A) = \arg\min_{S(A)} \frac{1}{S(A)} & \left[\sum_{i \in A} (f_i'(\hat{\boldsymbol{x}})^2 \hat{x}_i (P(A) - \hat{x}_i) - 2 \sum_{j \in A : j > i} f_i'(\hat{\boldsymbol{x}}) f_j'(\hat{\boldsymbol{x}}) \hat{x}_i \hat{x}_j) \right] \\ & + \frac{1}{1 - S(A)} \left[\sum_{i \in \tilde{A}} (f_i'(\hat{\boldsymbol{x}})^2 \hat{x}_i (P(\tilde{A}) - \hat{x}_i) - 2 \sum_{j \in \tilde{A} : j > i} f_i'(\hat{\boldsymbol{x}}) f_j'(\hat{\boldsymbol{x}}) \hat{x}_i \hat{x}_j) \right] \end{aligned} \tag{40}$$

The target sample share $S_t^{target}(A)$ under the approximate strategy is continuous in the estimates $\hat{\boldsymbol{x}}$. By Mann-Wald theorem, it converges in probability to the sample share that minimizes the true asymptotic variance: $S_t^{target}(A) \xrightarrow{p} S(A)$. The agent's strategy is to sample from group

A whenever the current sample share is below the target $S_t(A) < S_t^{target}(A)$, and to sample from group $\tilde{A}$ otherwise. Since the target sample share converges to a constant in the limit, the actual sample share $S_t(A)$ also converges to the same limit:

$$S_t(A) \xrightarrow{p} S(A) \tag{41}$$

Similarly, assuming continuity of f, the objective function for the optimal non-approximate strategy (multiplied by N) converges to the true asymptotic variance:

$$N \cdot E(\widehat{f(\boldsymbol{x})} - f(\boldsymbol{x}))^2 \xrightarrow{p} AVar(f(\boldsymbol{x}) \mid \boldsymbol{x}^*) \tag{42}$$

The sample share $S_t^*(A)$ under the optimal strategy is continuous in the sample (which is integer-valued), so it also converges in probability to the sample share that minimizes the true asymptotic variance:

$$S_t^*(A) \xrightarrow{p} S(A) \tag{43}$$

By Central Limit Theorem, $\sqrt{N}(\hat{\boldsymbol{x}} - \boldsymbol{x}) \xrightarrow{d} \mathcal{N}(0, AVar(\boldsymbol{x}))$. Applying the Delta method, sample shares converge at the rate of $\sqrt{N}$.

B Monte-Carlo Simulations: Exact vs. Approximate Sampling

We compare the approximate and optimal sampling strategies in finite samples using some problems, for which we can calculate variance exactly. Specifically, assume $\boldsymbol{x} = (x_1, ..., x_8) \sim Dir(1, ...,)$. Let the index be $Ind = \{A, \tilde{A}\}$, where $A = \{1, 2, 3, 4\}$, $\tilde{A} = \{5, 6, 7, 8\}$. Consider the following three statistics: $f^{lin}(\boldsymbol{x}) = x_1 + x_5$, $f^{prod}(\boldsymbol{x}) = (x_1 + x_2)(x_5 + x_6)$, $f^{rat}(\boldsymbol{x}) = \frac{x_1 + x_2}{x_5 + x_6}$.

For each of these statistics, we can calculate the exact variance as a function of the sample. Let n_i, $i \in \{1, ..., 8\}$ be the number of retrieved observations of subgroup i, and let $n_g = \sum_{i \in g} n_i$ for $g \in \{A, \tilde{A}\}$. The variances are

$$Var(f^{lin}(\boldsymbol{x})) = P(A)^2 \frac{(n_1 + 1)(n_A - n_1 + 3)}{(n_A + 4)^2 (n_A + 5)} + P(\tilde{A})^2 \frac{(n_5 + 1)(n_{\tilde{A}} - n_5 + 3)}{(n_{\tilde{A}} + 4)^2 (n_{\tilde{A}} + 5)} \tag{44}$$

$$\begin{aligned} Var(f^{prod}(\boldsymbol{x})) = & P(A)^2 P(\tilde{A})^2 \frac{(n_1 + n_2 + 2)(n_5 + n_6 + 2)}{(n_A + 4)(n_{\tilde{A}} + 4)} \\ & \times \left(\frac{(n_1 + n_2 + 3)(n_5 + n_6 + 3)}{(n_A + 5)(n_{\tilde{A}} + 5)} - \frac{(n_1 + n_2 + 2)(n_5 + n_6 + 2)}{(n_A + 4)(n_{\tilde{A}} + 4)} \right) \end{aligned} \tag{45}$$

$$Var(f^{rat}(\boldsymbol{x})) = \frac{P(A)^2}{P(\tilde{A})^2} \frac{(n_1 + n_2 + 2)(n_{\tilde{A}} + 3)}{(n_A + 4)(n_5 + n_6 + 1)} \tag{46}$$

$$\times \left(\frac{(n_1 + n_2 + 3)(n_{\tilde{A}} + 2)}{(n_A + 5)(n_5 + n_6)} - \frac{(n_1 + n_2 + 2)(n_{\tilde{A}} + 3)}{(n_A + 4)(n_5 + n_6 + 1)} \right) \tag{47}$$

We set the number of periods to be 100 and the number of simulations to be 10000 for each

	$\mid n_t(A) - n_t^*(A) \mid$	$\mid S_t(A) - S_t^*(A) \mid$
f^{lin}	0.13	0.004
f^{prod}	0.16	0.016
f^{rat}	1.01	0.053

Table 4: The simulated average difference between the approximate and optimal strategies for three problems.

of the three problems. For each simulation, we generate a random vector $\boldsymbol{x}$ from $Dir(1, ..., 1)$ and then generate a random database of observations according to the realized $\boldsymbol{x}$.

We compare two strategies. The approximate strategy solves problem (8) — minimizing the approximation of variance holding estimates fixed. The optimal strategy solves problem (5) — minimizing the exact variance, taking the next-period change in the estimates into account. We compare the number and the share of observations sampled from group A under both strategies. Table 4 summarizes the absolute differences in the strategies, averaged over simulations and time periods. The two strategies are very close across all three problems. The average absolute difference in the number of sampled observations from group A is between 0.13 and 1.01. The average absolute difference in the share of sampled observations from group A is between 0.004 and 0.053.

C Proof of Theorem 2

Consider a general index $Ind = \{A, \tilde{A}\}$. In the limit, the agent's sampling shares are characterized by problem (9). The solution and the minimized value of the asymptotic variance are

$$\frac{S(A)}{S(\tilde{A})} = \frac{\sqrt{\sum_{i \in A}(f_i'(\boldsymbol{x})^2 x_i(P(A) - x_i) - 2\sum_{i,j \in A: j>i} f_i'(\boldsymbol{x}) f_j'(\boldsymbol{x}) x_i x_j)}}{\sqrt{\sum_{i \in \tilde{A}}(f_i'(\boldsymbol{x})^2 x_i(P(\tilde{A}) - x_i) - 2\sum_{i,j \in \tilde{A}: j>i} f_i'(\boldsymbol{x}) f_j'(\boldsymbol{x}) x_i x_j)}} \tag{48}$$

$$\begin{aligned} AVar(f(\boldsymbol{x})) = &\Bigg(\sqrt{\sum_{i \in A}(f_i'(\boldsymbol{x})^2 x_i(P(A) - x_i) - 2 \sum_{i,j \in A: j>i} f_i'(\boldsymbol{x}) f_j'(\boldsymbol{x}) x_i x_j)} \\ &+ \sqrt{\sum_{i \in \tilde{A}}(f_i'(\boldsymbol{x})^2 x_i(P(\tilde{A}) - x_i) - 2 \sum_{i,j \in \tilde{A}: j>i} f_i'(\boldsymbol{x}) f_j'(\boldsymbol{x}) x_i x_j)} \Bigg)^2 \end{aligned} \tag{49}$$

Suppose $f(\boldsymbol{x})$ is a simple subset problem, so it can be represented as $\tilde{f}(P(I'_A \mid I_A), P(I'_{\tilde{A}} \mid I_{\tilde{A}}))$. For convenience, denote $\tilde{f}'_A = \frac{\partial \tilde{f}}{\partial P(I'_A|I_A)}(P(I'_A \mid I_A), P(I'_{\tilde{A}} \mid I_{\tilde{A}}))$, and $\tilde{f}'_{\tilde{A}} = \frac{\partial \tilde{f}}{\partial P(I'_{\tilde{A}}|I_{\tilde{A}})}(P(I'_A \mid$

$I_A), P(I'_{\tilde{A}} \mid I_{\tilde{A}}))$. Consider first $i \in A$ (everything is analogous for $i \in \tilde{A}$). The derivatives are

$$f'_i(\boldsymbol{x}) = \begin{cases} \tilde{f}'_A \dfrac{P(I_A \setminus I'_A)}{P(I_A)^2} \text{ for } i \in I'_A \\ \tilde{f}'_A \dfrac{-P(I'_A)}{P(I_A)^2} \text{ for } i \in I_A \setminus I'_A \\ 0 \text{ for } i \notin I_A \end{cases} \tag{50}$$

Plug in the values of $f'(\boldsymbol{x})$ into the formula above and simplify the expression under the square root:

$$\sum_{i \in A} (f'_i(\boldsymbol{x})^2 x_i (P(A) - x_i) - 2 \sum_{i,j \in A : j > i} f'_i(\boldsymbol{x}) f'_j(\boldsymbol{x}) x_i x_j) = \tilde{f}'^2_A \frac{P(I'_A \mid I_A)(1 - P(I'_A \mid I_A))}{P(I_A \mid A)} \tag{51}$$

The same steps apply for $i \in \tilde{A}$. Therefore, the ratio of sample shares and the asymptotic variance are

$$\frac{S(A)}{S(\tilde{A})} = \frac{\mid \tilde{f}'_A \mid \sqrt{P(I'_A \mid I_A)(1 - P(I'_A \mid I_A)) P(I_{\tilde{A}} \mid \tilde{A})}}{\mid \tilde{f}'_{\tilde{A}} \mid \sqrt{P(I'_{\tilde{A}} \mid I_{\tilde{A}})(1 - P(I'_{\tilde{A}} \mid I_{\tilde{A}})) P(I_A \mid A)}} \tag{52}$$

$$\begin{aligned} AVar(f(\boldsymbol{x})) = \Bigg(& \mid \tilde{f}'_A \mid \sqrt{\frac{P(I'_A \mid I_A)(1 - P(I'_A \mid I_A))}{P(I_A \mid A)}} \\ & + \mid \tilde{f}'_{\tilde{A}} \mid \sqrt{\frac{P(I'_{\tilde{A}} \mid I_{\tilde{A}})(1 - P(I'_{\tilde{A}} \mid I_{\tilde{A}}))}{P(I_{\tilde{A}} \mid \tilde{A})}} \Bigg)^2 \end{aligned} \tag{53}$$

D Proof of Theorem 3

Consider a general index $Ind = \{A, \tilde{A}\}$ and a linear problem $f(\boldsymbol{x})$. For a linear problem:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \sum_i f'_i(\boldsymbol{x}^*)(x_i - x^*_i) \tag{54}$$

Take expectations of both sides:

$$\widehat{f(\boldsymbol{x})} = f(\boldsymbol{x}^*) + \sum_i f'_i(\boldsymbol{x}^*)(\hat{x}_i - x^*_i) \tag{55}$$

The distortion for such a problem is

$$\begin{aligned}
Dist(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}) &= \mathbb{E}(\widehat{f(\boldsymbol{x})} - f(\boldsymbol{x}^*) \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}) && (56)\\
&= \mathbb{E}\left(\sum_i f_i'(\boldsymbol{x}^*)(\hat{x}_i - x_i^*)) \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}\right) && (57)\\
&= \sum_i f_i'(\boldsymbol{x}^*)\frac{\alpha_{g(i)}}{n_{g(i)} + \alpha_{g(i)}}\left(P(g(i))\frac{\alpha_i}{\alpha_{g(i)}} - x_i^*\right) && (58)\\
&= \frac{\alpha_A}{n_A + \alpha_A} Dist_A(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, 0, 0) + \frac{\alpha_{\tilde{A}}}{n_{\tilde{A}} + \alpha_{\tilde{A}}} Dist_{\tilde{A}}(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, 0, 0) && (59)
\end{aligned}$$

where $Dist_g(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*, 0, 0) = \sum_{i \in g} f_i'(\boldsymbol{x}^*)\left(P(g)\frac{\alpha_i}{\alpha_g} - x_i^*\right)$ for $g = A, \tilde{A}$ is the initial distortion at prior.

In the limit, $\frac{N}{n_{g(i)}+\alpha_{g(i)}} \xrightarrow[N\to\infty]{p} \frac{1}{S(g(i))}$. Therefore, the asymptotic distortion, decomposed by the index groups is

$$\begin{aligned}
ADist(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*) =& \frac{\alpha_A}{S(A)} \sum_{i \in A} f_i'(\boldsymbol{x}^*)\left(P(A)\frac{\alpha_i}{\alpha_A} - x_i^*\right)\\
&+ \frac{\alpha_{\tilde{A}}}{S(\tilde{A})} \sum_{i \in \tilde{A}} f_i'(\boldsymbol{x}^*)\left(P(\tilde{A})\frac{\alpha_i}{\alpha_{\tilde{A}}} - x_i^*\right) && (60)
\end{aligned}$$

Next, consider a general problem $f(\boldsymbol{x})$. Use Taylor approximation to derive distortion $\mathbb{E}(\widehat{f(\boldsymbol{x})} - f(\boldsymbol{x}^*) \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}})$:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*) + \sum_i f_i'(\boldsymbol{x}^*)(x_i - x_i^*) + \sum_{i,j} \frac{f_{i,j}''(\boldsymbol{x}^*)}{2}(x_i - x_i^*)(x_j - x_j^*) + h(\boldsymbol{x}, \boldsymbol{x}^*) \quad (61)$$

$$\Rightarrow \widehat{f(\boldsymbol{x})} - f(\boldsymbol{x}^*) = \sum_i f_i'(\boldsymbol{x}^*)(\hat{x}_i - x_i^*) + \sum_{i,j} \frac{f_{i,j}''(\boldsymbol{x}^*)}{2}(Cov_G(x_i, x_j) + (\hat{x}_i - x_i^*)(\hat{x}_j - x_j^*)) + \mathbb{E}_G(h(\boldsymbol{x}, \boldsymbol{x}^*)) \quad (62)$$

where $N \cdot \mathbb{E}_G(h(\boldsymbol{x}, \boldsymbol{x}^*)) \xrightarrow[N\to\infty]{p} 0$ and $N \cdot \mathbb{E}_G(\hat{x}_i - x_i^*)(\hat{x}_j - x_j^*) \xrightarrow[N\to\infty]{p} 0$. Take the expectation and limit to get the asymptotic distortion:

$$\begin{aligned}
N \cdot \mathbb{E}(\widehat{f(\boldsymbol{x})} - f(\boldsymbol{x}^*) \mid \boldsymbol{x}^*, n_A, n_{\tilde{A}}) \xrightarrow[N\to\infty]{p} & \sum_i \left(f_i'(\boldsymbol{x}^*)\frac{\alpha_{g(i)}}{S(g(i))}\left(P(g(i))\frac{\alpha_i}{\alpha_{g(i)}} - x_i^*\right)\right. && (63)\\
&\left. + \sum_{j \in g(i)} \frac{f_{i,j}''(\boldsymbol{x}^*)}{2} ACov(x_i, x_j \mid \boldsymbol{x}^*)\right) && (64)
\end{aligned}$$

Rearranging, the asymptotic distortion can be decomposed by the two index groups:

$$\begin{aligned}
&ADist(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*) \\
&= \frac{1}{S(A)} \sum_{i \in A} \left(f_i'(\boldsymbol{x}^*)\alpha_A \left(P(A)\frac{\alpha_i}{\alpha_A} - x_i^* \right) + \frac{f_{i,i}''(\boldsymbol{x}^*)}{2} x_i^*(P(A) - x_i^*) - \sum_{j \in A: j>i} f_{i,j}''(\boldsymbol{x}^*) x_i^* x_j^* \right) \\
&+ \frac{1}{S(\tilde{A})} \sum_{i \in \tilde{A}} \left(f_i'(\boldsymbol{x}^*)\alpha_{\tilde{A}} \left(P(\tilde{A})\frac{\alpha_i}{\alpha_{\tilde{A}}} - x_i^* \right) + \frac{f_{i,i}''(\boldsymbol{x}^*)}{2} x_i^*(P(\tilde{A}) - x_i^*) - \sum_{j \in \tilde{A}: j>i} f_{i,j}''(\boldsymbol{x}^*) x_i^* x_j^* \right)
\end{aligned} \tag{65}$$

The absolute value of a group-specific distortion is decreasing in the share of this group in the sample. The ratio of absolute values of group-specific distortions is proportional to the inverse ratio of the sample shares of those groups.

E Proof of Proposition 1

For conciseness, we refer to the various subgroups by their first letters: Good (G), Normal (N), Bad (B), non-Good (NG), non-Bad (NB), High outcome (H).

Consider an agent with an index for the Good state $Ind_G = \{G, NG\}$. Formulate the problem as a simple subset problem: $I_A = G$, $I'_A = G \cap H$, $I_{\tilde{A}} = B$, $I'_{\tilde{A}} = B \cap H$. Then, using Theorem 2, we derive the limiting optimal sampling shares:

$$\frac{S(G)}{S(NG)} = \frac{1/2}{1/2} \sqrt{\frac{P(H \mid G)(1 - P(H \mid G))P(B \mid NG)}{P(H \mid B)(1 - P(H \mid B))P(G \mid G)}} \tag{66}$$

Assume symmetry between Good and Bad states: $P(H \mid G) = 1 - P(H \mid B) > \frac{1}{2}$. Then, the ratio of optimal sampling shares is

$$\frac{S(G)}{S(NG)} = \sqrt{P(B \mid NG)} \tag{67}$$

Although the agent undersamples the Good state relative to non-Good, she oversamples the Good state relative to the Bad state:

$$\frac{S(G)}{S(B)} = \frac{S(G)}{S(NG)P(B \mid NG)} = \frac{1}{\sqrt{P(B \mid NG)}} > 1 \tag{68}$$

Using Theorem 3, we show that the belief distortion for the agent with the Good state index is positive:

$$ADist_G(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*) \propto \frac{1}{S(G)}\left(\frac{1}{2} - P(H \mid G)\right) + \frac{1}{S(B)}\left(\frac{1}{2} - P(H \mid B)\right) \tag{69}$$

$$\propto \frac{1}{2} - P(H \mid G) + \frac{1}{\sqrt{P(B \mid NG)}}\left(\frac{1}{2} - P(H \mid B)\right) \tag{70}$$

$$= \left(\frac{1}{\sqrt{P(B \mid NG)}} - 1\right)\left(\frac{1}{2} - P(H \mid B)\right) > 0 \tag{71}$$

Similarly, for an agent with the Bad state index, the optimal strategy oversamples the Bad state:

$$\frac{S(G)}{S(B)} = \frac{S(NB)P(G \mid NB)}{S(B)} = \sqrt{P(G \mid NB)} < 1 \tag{72}$$

As a result, the agent with the Bad state index underestimates the outcome:

$$ADist_B(\widehat{f(\boldsymbol{x})} \mid \boldsymbol{x}^*) \propto \frac{1}{S(G)}\left(\frac{1}{2} - P(H \mid G)\right) + \frac{1}{S(B)}\left(\frac{1}{2} - P(H \mid B)\right) \tag{73}$$

$$\propto \frac{1}{2} - P(H \mid G) + \sqrt{P(G \mid NB)}\left(\frac{1}{2} - P(H \mid B)\right) \tag{74}$$

$$= \left(\sqrt{P(G \mid NB)} - 1\right)\left(\frac{1}{2} - P(H \mid B)\right) < 0 \tag{75}$$

F Proof of Proposition 2

Agent F with the fine index can always reproduce the sampling strategy of agent C with the coarse index. Therefore, $AVar_F(f(\boldsymbol{x})) \leq AVar_C(f(\boldsymbol{x}))$.

Agent C can imperfectly reproduce the strategy of agent F by sampling from the respective coarser group. If agent F samples from group A_i, agent C can sample from group A and retrieve A_i with probability $P(A_i \mid A)$ and ignore the observation if it is not in A_i. Therefore, the ratio of effective sample sizes of agents C and F is bounded in the limit:

$$\frac{N_C}{N_F} \geq \min\{P(A_1 \mid A), ..., P(A_L \mid A), P(\tilde{A}_1 \mid \tilde{A}), ..., P(\tilde{A}_{\tilde{L}} \mid \tilde{A})\} \tag{76}$$

Since asymptotic variance is proportional to $\frac{1}{N}$, the lower bound on the ratio of asymptotic variances is

$$\frac{AVar_F(f(\boldsymbol{x}))}{AVar_C(f(\boldsymbol{x}))} \geq \min\{P(A_1 \mid A), ..., P(A_L \mid A), P(\tilde{A}_1 \mid \tilde{A}), ..., P(\tilde{A}_{\tilde{L}} \mid \tilde{A})\} \tag{77}$$

The following two examples show that the bounds are tight. Let $Ind_C = \{A, \tilde{A}\}$ and $Ind_F =$

$\{A_1, A_2, \tilde{A}\}$ Let k_1 be a value of non-indexed attribute K_1.

First, consider a question $P(k_1 \mid \tilde{A})$. The optimal strategy for both agents is to sample from group $\tilde{A}$, and hence in the limit $AVar_F(P(k_1 \mid \tilde{A})) = AVar_C(P(k_1 \mid \tilde{A}))$.

Second, suppose $P(A_1 \mid A) < P(A_2 \mid A)$ and consider a question $P(k_1 \mid A_1)$. Agent F can directly sample A_1. Agent C can only sample A, which yields A_1 with probability $P(A_1 \mid A)$. Therefore, in the limit

$$\frac{AVar_F(P(k_1 \mid A_1))}{AVar_C(P(k_1 \mid A_1))} = \min\{P(A_1 \mid A), ..., P(A_L \mid A), P(\tilde{A}_1 \mid \tilde{A}), ..., P(\tilde{A}_{\tilde{L}} \mid \tilde{A})\} \tag{78}$$

G Proof of Proposition 3

Suppose the agent uses an estimate based on the proxy index. After sampling N observations of A_1, her belief G^{Prox} about the distribution of $P(k_1 \mid \tilde{A}_1)$ is a mixture of two Dirichlet distributions: the prior distribution of $P(k_1 \mid \tilde{A}_1)$ with weight $1 - p$ and the posterior distribution of $P(k_1 \mid A_1)$ with weight p. Then, the expectation of $P(k_1 \mid \tilde{A}_1)$ is:

$$\mathbb{E}_{G^{Prox}} P(k_1 \mid \tilde{A}_1) = p\mathbb{E}_{G_N} P(k_1 \mid A_1) + (1-p)\mathbb{E}_{G_0} P(k_1 \mid \tilde{A}_1) \tag{79}$$

This is the proxy estimate after sampling N observations of A_1.

The variance of the posterior under the proxy index is the variance of the mixture of the two distributions:

$$\begin{aligned} Var_{G^{Prox}}(P(k_1 \mid \tilde{A}_1)) &= pVar_{G_N}(P(k_1 \mid A_1)) + (1-p)Var_{G_0}(P(k_1 \mid \tilde{A}_1)) \\ &\quad + p(\mathbb{E}_{G_N} P(k_1 \mid A_1))^2 + (1-p)(\mathbb{E}_{G_0} P(k_1 \mid \tilde{A}_1))^2 \\ &\quad - (p\mathbb{E}_{G_N} P(k_1 \mid A_1) + (1-p)\mathbb{E}_{G_0} P(k_1 \mid \tilde{A}_1))^2 && (80) \\ &= pVar_{G_N}(P(k_1 \mid A_1)) + (1-p)Var_{G_0}(P(k_1 \mid \tilde{A}_1)) \\ &\quad + p(1-p)(\mathbb{E}_{G_0} P(k_1 \mid \tilde{A}_1) - \mathbb{E}_{G_N} P(k_1 \mid A_1))^2 && (81) \end{aligned}$$

If the agent estimates $P(k_1 \mid \tilde{A}_1)$ directly, her posterior variance is $Var_{G^{Dir}}(P(k_1 \mid \tilde{A}_1))$ under posterior G^{Dir}. We assume that $\tilde{A}_1$ is sufficiently rare so that after sampling N observations from $\tilde{A}$, the agent does not encounter any $\tilde{A}_1$ observations. In this case, the posterior variance is equal prior variance: $Var_{G^{Dir}}(P(k_1 \mid \tilde{A}_1)) = Var_{G_0}(P(k_1 \mid \tilde{A}_1))$.

Then, the proxy index is better if it gives lower posterior variance:

$$p(1-p)(\mathbb{E}_{G_0} P(k_1 \mid \tilde{A}_1) - \mathbb{E}_{G_N} P(k_1 \mid A_1))^2 < p(Var_{G_0}(P(k_1 \mid \tilde{A}_1)) - Var_{G_N}(P(k_1 \mid A_1))) \tag{82}$$

In the limit, as $N \to \infty$, $Var_{G^{Dir}}(P(k_1 \mid \tilde{A}_1)) \xrightarrow[N\to\infty]{p} 0$, while $Var_{G^{Prox}}(P(k_1 \mid \tilde{A}_1)) \xrightarrow[N\to\infty]{p}$ $const > 0$ because $P(k_1 \mid \tilde{A}_1) \neq P(k_1 \mid A_1)$ with probability $1-p$. Thus, in the limit, $Var_{G^{Dir}}(P(k_1 \mid \tilde{A}_1)) < Var_{G^{Prox}}(P(k_1 \mid \tilde{A}_1))$ — the direct estimate is better.

H Proof of Proposition 4

Agent S starts with $n_S = N_0$ immediate observations of scientists and $n_{NS} = 0$ observations of non-scientists. Her initial belief distortion is positive:

$$
\begin{aligned}
Dist(\widehat{P(G)} \mid \boldsymbol{x}^*, N_0, 0) &= \mathbb{E}(\widehat{P(G)} - P(G) \mid \boldsymbol{x}^*, N_0, 0) && (83)\\
&= \mathbb{E}(\widehat{P(G \mid S)}P(S) + \widehat{P(G \mid NS)}P(NS) - P(G) \mid \boldsymbol{x}^*, N_0, 0) && (84)\\
&= \mathbb{E}\left(\frac{n_{G,S}+1}{N_0+2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} - P(G \mid S)\frac{1}{2} - P(G \mid NS)\frac{1}{2}\middle|\boldsymbol{x}^*, N_0, 0\right) && (85)\\
&= \frac{1}{2}\left(\frac{N_0 P(G \mid S)+1}{N_0+2} + \frac{1}{2} - P(G \mid S) - P(G \mid NS)\right) && (86)\\
&= \frac{1}{2}\left(\frac{N_0(2P(G \mid S)-1)}{2(N_0+2)}\right) > 0 && (87)
\end{aligned}
$$

Thus, agent S has a positive belief distortion about $P(G)$ from the start.

Similarly, agent NS starts with $n_{NS} = N_0$ immediate observations of non-scientists and $n_S = 0$ observations of scientists. Her initial belief distortion is negative:

$$
\begin{aligned}
Dist(\widehat{P(G)} \mid \boldsymbol{x}^*, 0, N_0) &= \mathbb{E}(\widehat{P(G)} - P(G) \mid \boldsymbol{x}^*, 0, N_0) && (88)\\
&= \mathbb{E}(\widehat{P(G \mid S)}P(S) + \widehat{P(G \mid NS)}P(NS) - P(G) \mid \boldsymbol{x}^*, 0, N_0) && (89)\\
&= \mathbb{E}\left(\frac{1}{2}\frac{1}{2} + \frac{n_{G,NS}+1}{N_0+2}\frac{1}{2} - P(G \mid S)\frac{1}{2} - P(G \mid NS)\frac{1}{2}\middle|\boldsymbol{x}^*, 0, N_0\right) && (90)\\
&= \frac{1}{2}\left(\frac{1}{2} + \frac{N_0 P(G \mid NS)+1}{N_0+2} - P(G \mid S) - P(G \mid NS)\right) && (91)\\
&= \frac{1}{2}\left(\frac{N_0(2P(G \mid NS)-1)}{2(N_0+2)}\right) < 0 && (92)
\end{aligned}
$$

Thus, agent NS has a negative belief distortion about $P(G)$ from the start.

I Proof of Proposition 5

To prove the proposition, we consider all possible problems and compare the performance of each index.

First, consider the unconditional problem $P(R)$. It is a simple problem, so Theorem 2 applies. Consider the gender index $Ind_G = \{F, M\}$. The asymptotic variance is

$$AVar_G(P(R)) = (\sqrt{P(R \mid F)(1-P(R \mid F))}P(F) + \sqrt{P(R \mid M)(1-P(R \mid M))}P(M))^2 \quad (93)$$

The same expression holds for the nationality index. Use the ex-post uninformativeness assumption to simplify the expression:

$$AVar_G(P(R)) = AVar_N(P(R)) = P(R)(1 - P(R)) \tag{94}$$

Thus, both indices lead to the same asymptotic variance.

Next, consider a conditional problem $P(R \mid S)$. Note that the problem conditions on a non-indexed attribute. This is not a simple problem for either index, but we can still derive the asymptotic variance.

Lemma 1. *Consider the problem $f(\boldsymbol{x}) = P(R \mid S)$. If the index attribute is ex-post uninformative, the asymptotic variance is*

$$AVar(P(R \mid S)) = \frac{P(R \mid S)(1 - P(R \mid S))}{P(S)} \tag{95}$$

Proof. Without loss of generality, consider the gender index. Note that $f_i'(\boldsymbol{x}) = \frac{P(NR \cap S)}{P(S)^2}$ for $i \in R \cap S$ and $f_i'(\boldsymbol{x}) = -\frac{P(R \cap S)}{P(S)^2}$ for $i \in NR \cap S$. Thus, the asymptotic variance from equation (9) is

$$\begin{aligned} AVar_G(P(R \mid S)) = \min_{S(F)} \sum_{g \in F,M} \frac{P(g)^2}{S(g)} \Big(& \frac{P(NR \cap S)^2}{P(S)^4} P(R \cap S \mid g)(1 - P(R \cap S \mid g)) \\ & + \frac{P(R \cap S)^2}{P(S)^4} P(NR \cap S \mid g)(1 - P(NR \cap S \mid g)) \\ & + 2\frac{P(NR \cap S)P(R \cap S)}{P(S)^4} P(R \cap S \mid g)P(NR \cap S \mid g)\Big) \end{aligned} \tag{96}$$

Solving for the optimal sampling shares $S(F)$ yields

$$\begin{aligned} AVar_G(P(R \mid S)) = \Big(\sum_{g \in F,M} P(g) \Big(& \frac{P(NR \cap S)^2}{P(S)^4} P(R \cap S \mid g)(1 - P(R \cap S \mid g)) \\ & + \frac{P(R \cap S)^2}{P(S)^4} P(NR \cap S \mid g)(1 - P(NR \cap S \mid g)) \\ & + 2\frac{P(NR \cap S)P(R \cap S)}{P(S)^4} P(R \cap S \mid g)P(NR \cap S \mid g)\Big)^{\frac{1}{2}}\Big)^2 \end{aligned} \tag{97}$$

Next, use ex-post uninformativeness to simplify the expression:

$$AVar_G(P(R \mid S)) = \frac{P(R \mid S)(1 - P(R \mid S))}{P(S)} \tag{98}$$

□

Since both indices are ex-post uninformative, Lemma 1 shows that both indices result in the same asymptotic variance for problems $P(R \mid S)$ and $P(R \mid NS)$:

$$AVar_G(P(R \mid S)) = AVar_N(P(R \mid S)) = \frac{P(R \mid S)(1 - P(R \mid S))}{P(S)} \tag{99}$$

$$AVar_G(P(R \mid NS)) = AVar_N(P(R \mid NS)) = \frac{P(R \mid NS)(1 - P(R \mid NS))}{P(NS)} \tag{100}$$

Next, consider a problem conditioning on one of the indexable attributes, $P(R \mid k)$, $k \in \{F, M, A, \tilde{A}\}$. If the agent has the corresponding index, she can sample observations from k_i directly. Under the uninformativess assumption, this leads to the following asymptotic variance:

$$AVar(P(R \mid k)) = P(R)(1 - P(R)) \tag{101}$$

If the agent does not have the corresponding index, then it is not a simple problem, and Lemma 1 applies. Under the uninformativess assumption, this leads to the following asymptotic variance:

$$AVar(P(R \mid k)) = \frac{P(R)(1 - P(R))}{P(k)} \tag{102}$$

Suppose each problem $P(R \mid k)$, $k \in \{F, M, A, \tilde{A}\}$ arises with probability q_k. Assuming that gender and nationality are ex-post balanced, the expected asymptotic variance for these problems for each index is

$$\mathbb{E}_q AVar_G(P(R \mid k)) = P(R)(1 - P(R)) \left(q_F + q_M + 2(q_A + q_{\tilde{A}})\right) \tag{103}$$

$$\mathbb{E}_q AVar_N(P(R \mid k)) = P(R)(1 - P(R)) \left(q_A + q_{\tilde{A}} + 2(q_F + q_M)\right) \tag{104}$$

Therefore, the gender index has a lower expected asymptotic variance than the nationality index if and only if $q_F + q_M > q_A + q_{\tilde{A}}$, i.e., the probability of problems conditioning on gender is higher than the probability of problems conditioning on nationality.

J Proof of Proposition 6

First, consider a problem conditional on the non-indexable attribute, $P(R \mid S)$ (equivalently, $P(R \mid NS)$). Using Lemma 1, the asymptotic variance with the ex-post uninformative nationality index is

$$AVar_N(P(R \mid S)) = \frac{P(R \mid S)(1 - P(R \mid S))}{P(S)} \tag{105}$$

$$AVar_N(P(R \mid NS)) = \frac{P(R \mid NS)(1 - P(R \mid NS))}{P(NS)} \tag{106}$$

If the agent instead has the fully informative gender index, i.e. $P(M \mid S) = 1$ and $P(M \mid$

$NS) = 0$, then the asymptotic variance for problem $P(R \mid S)$ from equation (97) simplifies to

$$\begin{aligned} AVar_G(P(R \mid S)) =& P(M)^2\Big(\frac{P(NR \cap S)^2}{P(S)^4}P(R \mid S)(1-P(R \mid S)) \\ &+ \frac{P(R \cap S)^2}{P(S)^4}P(NR \mid S)(1-P(NR \mid S)) \\ &+ 2\frac{P(NR \cap S)P(R \cap S)}{P(S)^4}P(R \mid S)P(NR \mid S)\Big) \\ =& P(R \mid S)(1-P(R \mid S)) \end{aligned} \tag{107}$$

Analogously, for problem $P(R \mid NS)$:

$$AVar_G(P(R \mid NS)) = P(R \mid NS)(1-P(R \mid NS)) \tag{108}$$

Therefore, having an informative index reduces the asymptotic variance for problems that condition on the non-indexable attribute.

Next, consider the unconditional problem $P(R)$. With the uninformative nationality index, the asymptotic variance is the same as without any index:

$$AVar_N(P(R)) = P(R)(1-P(R)) \tag{109}$$

Consider now the gender index. It is a simple problem, so based on Theorem 2, the asymptotic variance is

$$\begin{aligned} AVar_G(P(R)) =& \Big(P(F)\sqrt{P(R \mid F)(1-P(R \mid F))} + P(M)\sqrt{P(R \mid M)(1-P(R \mid M))}\Big)^2 \\ =& \Big(P(F)\sqrt{P(R \mid F)(1-P(R \mid F))} \\ &+ \sqrt{(P(R)-P(R \mid F)P(F))(1-P(R)-(1-P(R \mid F))P(F))}\Big)^2 \end{aligned} \tag{110}$$

The expression above for $AVar_G(P(R))$, as a function of $P(R \mid F)$, is maximized at $P(R \mid F) = P(R)$, with the maximum value of $P(R)(1-P(R))$. Therefore, $AVar_G(P(R)) \leq AVar_N(P(R))$, with strict inequality unless $P(R \mid F) = P(R)$. The informative gender index leads to a lower asymptotic variance than the uninformative nationality index.

K Proof of Proposition 7

We proceed in the same manner as in Appendix I. Since both indexable attributes are ex-post uninformative, their asymptotic variances for problems $P(R)$, $P(R \mid S)$, and $P(R \mid NS)$ are the same:

$$AVar_G(P(R)) = AVar_N(P(R)) = P(R)(1 - P(R)) \tag{111}$$

$$AVar_G(P(R \mid S)) = AVar_N(P(R \mid S)) = \frac{P(R \mid S)(1 - P(R \mid S))}{P(S)} \tag{112}$$

$$AVar_G(P(R \mid NS)) = AVar_N(P(R \mid NS)) = \frac{P(R \mid NS)(1 - P(R \mid NS))}{P(NS)} \tag{113}$$

Next, consider problems conditioning on the indexable attributes $P(R \mid k)$, $k \in \{F, M, A, \tilde{A}\}$. Suppose they all arise with the same probability, $q_F = q_M = q_A = q_{\tilde{A}} = q$. Then, from the results in Appendix I, the expected asymptotic variance for these problems with each index is

$$\mathbb{E}_q AVar_G(P(R \mid k)) = P(R)(1 - P(R))q\left(1 + 1 + \frac{1}{P(A)} + \frac{1}{P(\tilde{A})}\right) \tag{114}$$

$$\mathbb{E}_q AVar_N(P(R \mid k)) = P(R)(1 - P(R))q\left(1 + 1 + \frac{1}{P(F)} + \frac{1}{P(M)}\right) \tag{115}$$

An expression of the form $\frac{1}{x} + \frac{1}{1-x}$ with $x \in (0, 1)$ is minimized at $x = \frac{1}{2}$ and is increasing towards both endpoints. Thus, if $\mid P(F) - P(M) \mid < \mid P(A) - P(\tilde{A}) \mid$, then $\mathbb{E}_q AVar_N(P(R \mid k)) < \mathbb{E}_q AVar_G(P(R \mid k))$ for $k \in \{F, M, A, \tilde{A}\}$.

L Proof of Proposition 8

Use the ex-post uninformativeness assumption to simplify problem (96):

$$\begin{aligned} AVar_G(P(R \mid S)) = \sum_{g \in F,M} \frac{P(g)^2}{S(g)} \Big(& \frac{P(NR)^2 P(S)^2}{P(S)^4} P(R)P(S \mid g)(1 - P(R)P(S \mid g)) \\ & + \frac{P(R)^2 P(S)^2}{P(S)^4} P(NR)P(S \mid g)(1 - P(NR)P(S \mid g)) \\ & + 2\frac{P(NR)P(S)^2 P(R)}{P(S)^4} P(R)P(S \mid g)^2 P(NR)\Big) \end{aligned} \tag{116}$$

The variance-minimizing share of men in the sample is

$$S(M) = \frac{P(M)\sqrt{P(S \mid M)}}{P(M)\sqrt{P(S \mid M)} + P(F)\sqrt{P(S \mid F)}} \tag{117}$$

Multiplying by $P(S \mid M)$, get the share of male scientists in the sample:

$$S(\text{Scientist, Male}) = P(\text{Scientist, Male}) \times \frac{\sqrt{\frac{P(S|M)}{P(S|F)}}}{P(M)\sqrt{\frac{P(S|M)}{P(S|F)}} + P(F)} \tag{118}$$

The sample share S(Scientist, Male) is increasing in the representativeness of scientists for men,

$\frac{P(S|M)}{P(S|F)}$, and is equal the population share $P(\text{Scientist, Male})$ if $\frac{P(S|M)}{P(S|F)} = 1$.