### Title

Stereotypes as Bayesian Inferences: Hierarchical Computations Underlying Belief Formation and Maintenance




### Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)



### Authors

Lu, Junsong
Lin, Chujun



### Publication Date

2026

# Stereotypes as Bayesian Inferences: Hierarchical Computations Underlying Belief Formation and Maintenance

**Junsong Lu (jul140@ucsd.edu)[1] & Chujun Lin[2]**
[1]Department of Psychology, University of California San Diego
[2]Department of Psychology, University of Columbia

**Abstract**

Stereotypes are pervasive and rigid. We propose that stereotyping arises because perceivers solve a hierarchical inference problem in which traits are inferred simultaneously at the individual and group levels. Using hierarchical Bayesian models, we derive normative predictions about stereotype formation and updating and test them in three experiments. Study 1 shows that evidence pooling in hierarchical inference enables group-level impressions to form faster than individual-level impressions when perceivers observe limited information of many group members, allowing strong stereotypes without strong individual impressions. Study 2 shows that stereotypes resist updating when stereotype-inconsistent group members behave heterogeneously, making perceivers attribute counterevidence to individual differences rather than updating group stereotypes. Study 3 shows that when stereotype-inconsistent individuals also belong to a second, unfamiliar group, counterevidence is attributed to this second group, leaving the original stereotype intact. Together, these findings provide a computational account of how hierarchical inference produces and sustains rigid group beliefs.

**Keywords:** stereotypes; Bayesian cognition; person perception

## Introduction

Stereotypes are pervasive and resistant to change (Lippmann, 1922). Classic accounts attribute their persistence to ingroup favoritism (Tajfel et al., 1971) and cognitive shortcuts that simplify social judgments (Fiske et al., 2007; Fiske & Neuberg, 1990), but they fail to address when stereotypes emerge and why they are so rigid. Recent information-sampling accounts show that bias can arise even in the absence of true group differences or cognitive limitations: early impressions shape subsequent behavior by discouraging interaction with certain groups, creating biased samples that sustain stereotypes (Allidina & Cunningham, 2021; Bai et al., 2022; Denrell, 2005; Harris et al., 2020; Le Mens & Denrell, 2011). Despite their contributions, they depend on specific assumptions about reward structures and do not fully explain the rapid speed or ubiquity of stereotype acquisition (Chua & Freeman, 2022; Cuddy et al., 2009). Moreover, they primarily attribute the source of bias to self-reinforcing interaction preference and thus limited exposure to counterevidence. However, stereotypes may persist even when disconfirming information is readily available (Schultner, Lindström, et al., 2024).

We proposed a computational theory of stereotype formation and maintenance, explaining why stereotypes are acquired rapidly and persist despite disconfirming evidence. Specifically, we argue that social interaction with members of different groups (e.g., races) poses a hierarchical inference problem: perceivers must infer or update trait inferences about individuals and their group as a whole to guide subsequent social interactions. Like other forms of human knowledge, social inferences operate at multiple levels of abstractions: group-level representations, or stereotypes, guides inferences about individuals, while individual-level inferences in turn shape group stereotypes. Hierarchical Bayesian models (HBMs) provide a normative solution to this problem and have been widely used to explain categorical cognition and causal inference (Anderson, 1991; Kemp et al., 2007; Kemp & Tenenbaum, 2008). Two computational advantages of HBMs make them well suited for modeling stereotyping. First, despite the apparent complexity of hierarchical models, HBMs are statistically parsimonious due to pooling: data from different clusters (e.g., each group member) are combined to estimate higher-level group estimations like the group average, which induce dependencies among individual estimations. This reduces the number of effective parameters and computational costs relative to single-level models that represent all individual parameters independently (McElreath, 2015). Second, group-level representations serve as informative priors for attributing newly encountered individuals, enabling faster individual impression formation than flat priors that require more data. Together, these properties mirror how stereotypes emerge from learning about individual group members and suggest that the mind may exploit hierarchical inference as a computationally efficient strategy.

We formalized this account in three studies. In each study, we derived the hypotheses from simulations and then tested them in human experiments. We define stereotypes as probabilistic representations (e.g., posterior distributions) of traits regarding a social group; stronger stereotypes correspond to more confident and certain group impressions. In Study 1, we addressed the question of why stereotypes are formed rapidly. We then investigated how stereotypes are maintained in Study 2 and 3 as initially formed stereotypes can be challenged by counter-stereotypical group members. Different from prior work assuming that perceivers have strong control over whom to interact with (Allidina & Cunningham, 2021; Bai et al., 2022), we argue that exposure



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

to counterevidence in daily life is inevitable. We elaborate on these hypotheses in detail in their corresponding section.

## Study 1

A key feature of HBMs is that learning occurs simultaneously at multiple levels of abstraction, enabling faster learning of group-level knowledge than individual-level knowledge in certain conditions (Kemp et al., 2007; Kruschke, 2015; McElreath, 2015). We propose and test the *interaction sparsity hypothesis*, which predicts that group-level impressions become more certain than individual impressions when information about each group member is sparse (Figure 1). Sparsity arises when a fixed amount of observed information is distributed across many group members, yielding only limited evidence about each individual. Consequently, perceivers can develop strong stereotypes about a group without strong beliefs toward any specific individual. We next simulate this process.

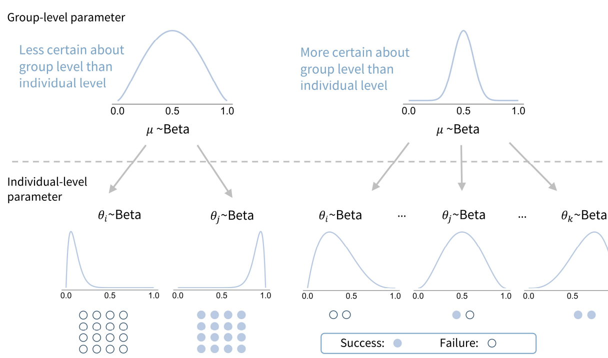


Figure 1: The interaction sparsity hypothesis. Left: high dispersion (e.g., variance) in individual behavior yields a weak group-level impression μ (e.g., one individual succeeds 0/16 times while another succeeds 16/16). Right: with few observations per individual, individual impressions are uncertain; however, because there is no strong evidence that individuals are highly distinctive**,** pooling across individuals produces a strong group-level impression.

### Simulations

Across the paper, each target's trait or the group-level trait was represented with a variable ranging from 0 to 1, reflecting underlying "competence" that is not directly observable and assumed to follow a beta distribution. Empirical observations were binary (0 = failure, 1 = success), reflecting underlying competence but subject to uncertainty. A hierarchical Bayesian perceiver had observations from multiple group members and inferred both individual and group-level competence by implementing a hierarchical beta-binomial model:

$$y_i \sim \text{Binomial}(n_i, \theta_i);\ \theta_i \sim \text{Beta}(\mu\kappa, (1-\mu)\kappa)$$
$$\mu \sim \text{Beta}(2,2);\ \kappa \sim \text{Gamma}(6,0.3)$$

where $y_i$ represents observations from the $i$th target, $\theta_i$ is their estimated competence. There are also group perceptions: $\mu$ is the group-level average competence, and $\kappa$ is the concentration parameter which characterizes how closely individual competence clusters around the group average $\mu$. We used weak priors for both group-level parameters.

We used a factorial design to study the effect of interaction sparsity. Total observations ($N_{total} \in \{8,16,32,64\}$) were crossed with different competence priors ($\alpha \in \{2,4,6,8\}$), and the number of encountered group members ($N_{target} \in \{N_{total}, N_{total}/2, 2\}$) to manipulate observations per target. Across these 48 primary conditions, we used integer partitioning to generate all valid distributions of mining successes across targets, yielding 177 unique conditions. For each condition, we calculated the posterior SD of $\mu$ as a measure of certainty (smaller SD = stronger stereotype) and a dispersion score (SD of success counts across individuals). Model parameters were estimated via Gibbs sampling.

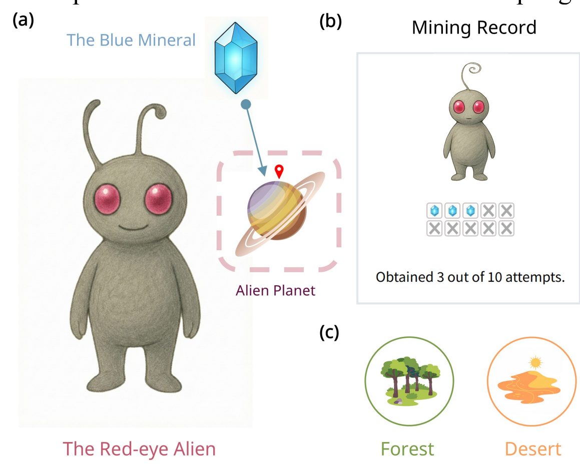


Figure 2: Experimental manipulations. (a) Participants played the role of explorers observing aliens mining blue minerals. Participants were told that the aliens' eye color indicated their group membership, and each participant only observed aliens from one group (red). (b) Participants infer mining capability from individuals' mining records (example records from Study 2). (c) In Study 3, an additional group membership (living area: forest vs. desert) was introduced.

### Experiments

We recruited 365 participants from CloudResearch Connect (age: $M$ = 40.40, $SD$ = 12.90; 48.5% females). To align with the simulation, participants inferred the underlying competence from observed data, with the total number of observations across targets fixed at 32. We manipulated dispersion (low vs high) and the number of targets (2 vs 16), yielding 4 conditions (observations: low-2: [7, 9]; high-2: [4, 12]; low-16: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 0, 0]; high-16: [2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 2, 0, 1, 1]). For example, in the high-2 condition, participants observed 2 targets and for each target observed 16 mining records (32 ÷ 2). One target succeeded in 4 trials thus extracted 4 minerals while the other in 12 out of 16 trials. In contrast, in the high-16 condition, participants observed 16 targets, each with 2 trials.

Participants were randomly assigned to one of the 4 conditions. They read a cover story that they were investigating aliens in an alien planet (Figure 2). On the planet, aliens extract blue minerals. Participants needed to estimate the mining capability of alien groups defined by eye colors (e.g., red). The mention of different eye colors served only to activate group categorization, and participants only observed aliens belonging to the red-eye group. This setup was the same across all studies. Participants estimated the mining capability of red-eyed aliens by viewing mining records from multiple aliens. Each alien's mining record consisted of multiple attempts, with each attempt resulting in either one mineral (success) or nothing (failure). Regardless of the number of targets, two mining attempts were displayed per screen to equate task length across conditions. For example, in the 2-target condition, each alien's 16 attempts were shown across 8 screens.

After viewing all records, participants rated the competence of the last observed alien and the group as a whole, and indicated their confidence in each rating, all on 7-point Likert scales. The order of judgments was randomized.

## Results and Discussion

Simulation results demonstrated that the sparsity of social observations promotes stereotype formation. Specifically, increasing the number of targets observed ( $N_{target}$) significantly decreased the posterior uncertainty of group impression after controlling for the total number of observations ($N_{total}$), $\beta = -5.19 \times 10^{-4}, p < .001$. There was also a significant interaction effect between $N_{target}$ and dispersion, $\beta = 2.69 \times 10^{-4}, p = .01$. If more targets from the group are encountered and they are heterogeneous, the stereotype would be weakened.

We measured impression certainty in human subjects using the confidence ratings. A robust two-way ANOVA with number of targets (2 vs. 16) and dispersion (high vs. low) on group-level confidence revealed no significant main effects nor any interaction (all *p*s > .24). This null result may, however, reflect participants' use of idiosyncratic internal scales in Likert judgments, limiting interpretability. We thus derived a confidence difference score by subtracting the individual from the group-level confidence, with positive values indicate stronger and more certain group-level impressions (stereotypes). We performed a robust ANOVA on the confidence difference score. The number of targets significantly increased the gap between individual- and group-level certainty, Q = 32.87, $p$ = .001. Higher dispersion reduced the gap marginally, Q = 3.62, $p$ = .06. The main effects were qualified by a significant interaction effect, Q = 6.59, $p$ = .012, indicating that when only 2 heterogeneous targets were observed, perceivers felt more certain about individual-level impressions than group-level impressions. In all other conditions, group impressions were more confident. These results suggest that stereotypes are strengthened not through deep engagement with group members, but through thin slices of observations of many group members.

# Study 2

We next examine why initially formed stereotypes resist updating despite contradictory evidence. We propose and test the *attenuated contingency hypothesis* posits that hierarchical Bayesian reasoning discourages stereotypes from updating when perceivers observe multiple stereotype-inconsistent group members whose behaviors are heterogeneous. In the beta-binomial model, this process is governed by the concentration parameter $\kappa$: greater dispersion of individual impressions $\theta_i$ around the group mean $\mu$ yields lower $\kappa$ and weaker perceived group-individual contingency (e.g., how strong $\theta_i$ is dependent on $\mu$). When $\kappa$ is low and both individual-level impressions $\theta_i$ and the group mean $\mu$ are uncertain, new observations update both levels. However, when the group-level stereotype is strong, stereotype-inconsistent evidence preferentially updates the more uncertain individual-level parameter, leaving $\mu$ largely unchanged (Figure 3). In real-world settings, high heterogeneity among individuals naturally induces low $\kappa$, creating a paradox: established stereotypes are resistant to change when they are inaccurate—that is, when no single group mean adequately summarizes group member variation.

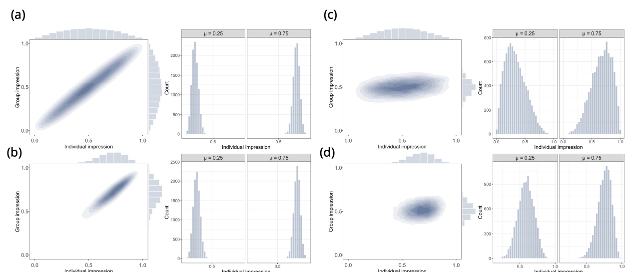


Figure 3: The attenuated contingency hypothesis. (a)–(b) Under a fixed high κ (strong group–individual contingency), initially uncertain group-level and individual-level impressions (a) are jointly updated when a group member shows strong competence (b), as reflected by tighter marginal distributions concentrated at the high end. Paired histograms (μ = 0.25 and μ = 0.75) show the conditional distribution of individual-level impressions given the group average. (c)–(d) Under low κ and certain group-level impression, the same evidence primarily updates individual-level impressions while the group-level impression remains stable. Here, the conditional distributions are wider than in (c)–(d), reflecting weaker group–individual contingency.

## Simulations

Using the same beta-binomial model as in Study 1, we simulated belief updating in response to stereotype-inconsistent evidence. A prior stereotype was specified by assigning the group average $\mu$ a Beta(10, 20) distribution, reflecting a strong belief of low competence (mean ≈ 0.33), and placing a Gamma(0.01, 0.01) prior on $\kappa$. The perceiver then observed new records from multiple group members and updated both individual- and group-level impressions. We used a factorial design, manipulating the number of group members ( $N_{target} \in \{5,10,15\}$ ), observations per targets ($N_{obs\_each} \in \{2,5,8\}$), and the distribution of ground-truth competence ($a \in \{30,60,90\}$ for Beta(a, 20) ). This resulted in 27 primary conditions. For each condition, we enumerated all possible evidence configurations via integer partitioning, spanning perfectly uniform to highly heterogeneous distributions. For example, if there were 5 extracted minerals under the condition $N_{target} = 5, N_{obs\_each} = 2$, a list of more concentrated observations would be [1,1,1,1,1], while a dispersed one would be [2,2,1,0,0]. This resulted in 11539 finer conditions. Dispersion was quantified as the variance of observations. The dependent measure was an update score, defined as the posterior minus prior mean of μ; smaller values indicate stronger stereotype persistence.

## Experiments

We recruited 160 participants from CloudResearch Connect (age: *M* = 41.60, *SD* = 11.90; 46% females). Participants were randomly assigned to either a concentrated or dispersed evidence condition. They read a cover story similar to Study 1 and estimated the mining capability of red-eyed aliens, defined as the average number of minerals collected in 10 attempts. To induce an initial stereotype, participants first viewed estimates given by three prior participants. They were told that each prior participant had viewed the 10-attempt mining records of multiple aliens, based on which they estimated the group-level mining ability. The three prior participants' estimates were 2.6, 2.8, and 3.0 minerals (out of 10), providing a relatively negative impression with room for updating given more positive evidence. Each estimate was displayed for 10 seconds.

After viewing these estimates, participants' group-level beliefs were measured using Markov Chain Monte Carlo with People (MCMCP), a method for eliciting mental representations in the form of probability distributions (Sanborn et al., 2010). Participants completed a two-alternative forced-choice task consisting of 20 trials. In each trial, they indicated which of two values better reflected the average number of minerals extracted by red-eye aliens in 10 attempts. The options were generated using the Metropolis-Hastings algorithm. On the first trial, one option was sampled from Normal(2.5, 0.7), and the other was sampled from a normal distribution centered on the value of the other option; on subsequent trials, the option selected in the previous trial was retained while the alternative option was resampled around it. All values were discretized to one decimal place, and the resulting choices approximate samples from participants' posterior belief over group-level competence.

Then, participants gathered their own observations of 7 aliens, observing each alien's mining record out of 10 attempts. In the concentrated evidence condition, the 7 records were either [4, 6, 4, 6, 5, 5, 5] or [4, 6, 5, 5, 5, 5, 5]. These records exhibited low variance and were shown in random order for each participant. In the dispersed condition, the records showed high variance and were fixed as [0, 2, 4, 5, 7, 8, 9]. Because variance was the focal manipulation, trials were randomized with a constraint that adjacent records were not highly similar. Importantly, the mean number of extracted minerals was identical across conditions. After observing all seven aliens, participants completed two Likert-scale items assessing whether mining capability was driven by group (red eyed) or individual differences. They then completed the MCMCP task again to measure their updated estimate of the group's mining capability.

## Results and Discussion

The 11539 simulated experiments were analyzed using a linear mixed model. Observing more targets ($\beta = 0.008, p = .003$) and having more observations per target ($\beta = 0.025, p < .001$) increased stereotype updating. As predicted, dispersion of observations discouraged stereotype updating, $\beta = -0.020, p < .001$. There was also a significant interaction between $N_{target}$ and dispersion, $\beta = -0.002, p = .01$, suggesting that if more targets from the group were observed and they were heterogeneous, stereotype update would be further attenuated.

In the experiment, participants in the concentrated condition rated eye-color as more influential on capability than those in the dispersed condition, $t(158) = 5.77$, $p < .001$, Cohen's $d = 0.91$. The pattern was reversed for individual-difference rating, $t(158) = -6.27$, $p < .001$, Cohen's $d = -0.99$. These results provided initial evidence that stereotype-inconsistent information was preferentially attributed to individual differences when group members vary.

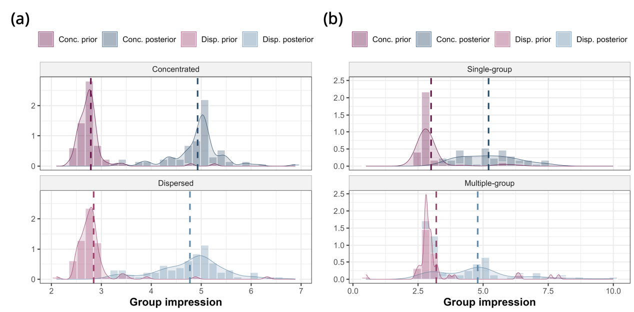


Figure 4: Distributions of responses in Studies 2 and 3. Prior and posterior response distributions are shown for the (a) concentrated and dispersed conditions in Study 2, and for the (b) single-group and multi-group conditions in Study 3.

Next, we calculated each participant's capability estimate as the mean of their 20 MCMCP choices. A robust mixed-effects ANOVA with the condition (concentrated vs. dispersed) as a between-subjects factor and the measurement time (prior vs. posterior) as a within-subjects factor revealed a significant main effect of measurement, $F(1, 70.69) = 2786.46$, $p < .001$, as well as a significant condition by measurement interaction, $F(1, 70.69) = 4.25$, $p = .043$. This interaction indicated that the posterior estimate of the alien groups' mining capability was lower in the dispersed condition (*M* = 4.77, *SD* = 0.68) than in the concentrated condition (*M* = 4.93, *SD* = 0.53), suggesting a greater stereotype rigidity under heterogeneous evidence.

# Study 3

Study 3 extends Study 2 by leveraging the same mechanism: once stereotypes are established through social learning, counterevidence is preferentially assigned to uncertain alternative parameters rather than revising established stereotypes. In Study 2, the uncertainty resided at the individual level, whereas in Study 3 it resides in alternative group memberships. We propose the *multi-group membership hypothesis*: when individuals belong to multiple social groups, stereotype-inconsistent evidence is attributed to the more uncertain group membership, thus shielding the established stereotypes from revision. For example, an individual with strong stereotypes about men but weak stereotypes about Californians may attribute a men-stereotype-inconsistent observation from a Californian man to "Californians" rather than updating their beliefs about men. This arises from the multiplicative nature of Bayesian inference. When the prior of males is concentrated on a narrow set of hypotheses and assigns near-zero probability to alternatives, it discounts even strong counterevidence (e.g., 0 times 20 = 0). In contrast, a diffuse prior over Californians remains responsive and is more likely to absorb new evidence.

## Simulations

When an individual shares multiple group memberships, individual impressions can be influenced by a mixture of multiple stereotypes. We parameterized a hierarchical beta-binomial model in which the individual impressions were determined by a mixture of two group-level beta distributions. For the first group (Group A), we specified a relatively strong prior to represent an established stereotype. For the second group (Group B), we specified a weak prior, Beta(1, 1), to capture a less established group impression.

We used a factorial design manipulating the number of group members ($N_{target} \in \{5,10,15\}$), observations per targets ($N_{obs\_each} \in \{2,5,10\}$), and prior stereotype strength ($a \in \{1,2,5,13,30\}$), yielding 45 conditions. Prior beliefs about Group A were parameterized as Beta(a, 3a). In each condition, we also simulated a control model where individuals belonged to only one group. The dependent variable was the update difference score: the difference in the posterior mean of the group-level estimate between the multi-group and single-group models. Positive values indicate greater stereotype rigidity due to multi-group memberships.

## Experiments

We recruited 100 participants from CloudResearch Connect (age: $M$ = 41.30, $SD$ = 11.60; 56% females). Participants were randomly assigned to a multi-group or single-group condition. Participants read the same cover story as in Study 2. In the single-group condition, participants were told that they would observe the red-eye alien group. In the multi-group condition, participants were additionally told that the aliens lived in different areas (desert vs. forest), regardless of eye color. Thus, each alien belonged to two groups simultaneously (eye color and living area), allowing impressions of an alien to be decomposed into contributions from both group memberships. In the single-group condition, all aliens were described as inhabiting the same living area, rendering living area non-diagnostic.

Following the same setting in Study 2, participants viewed estimates given by three prior participants. Different from Study 2, rather than doing MCMCP, participants then simply provided an initial estimate of the group's average mining capability using a slider ranging from 0 to 10. This simpler measure was chosen for efficiency.

Next, participants observed mining records from five aliens ([3, 7, 6, 7, 5]). The first record (3) was always presented first and was consistent with the prior; the remaining four records were shuffled. In the single-group condition, the mining records were shown together with the alien. In the multi-group condition, the mining records were shown together with the alien and its living area, with the first alien always belonging to a different living area than the remaining four. Because living area is an artificial category only briefly introduced at the outset, we paired it with each observation and contrasted categories to increase its salience.

After observing all records, participants rated the extent to which mining capability was influenced by eye color, living area, and individual differences in a 7-point Likert scale. They then provided an updated estimate of the alien group's capability using the same 0-10 slider. Participants in the multi-group condition additionally rated the mining capability of aliens living in an area corresponding to the last four aliens they saw (e.g., forest aliens).

## Results and Discussion

For the simulation study, we first conducted a one-sample t-test on the update difference score. The update difference was significantly greater than zero, $t(44) = 14.60$, $p < .001$, Cohen's $d = 2.17$, indicating stronger stereotype rigidity when a second group membership was present. Correspondingly, the estimated mixing ratio assigned to the stereotyped group was significantly below 50%, $t(44) = -24.50$, $p < .001$, Cohen's $d = -3.65$. This indicated that impressions of new individuals were less influenced by the stereotyped group (e.g., less than 50%) due to the multiplicative nature mentioned above. A linear regression showed that prior strength, $N_{target}$, and $N_{obs\_each}$ significantly predicted the update difference score (all $ps < .025$), but their effects remained small (all $\beta s < .008$). The large intercept ($\beta = 0.227, p < .001$) indicated that stereotype persistence was primarily driven by the presence of the additional group membership.

Regarding the experiment, participants in the single-group condition attributed mining capability more strongly to eye color than those in the multi-group condition, $t(98) = 2.41$, $p = .011$, Cohen's $d = 0.48$, and significantly less to living area, $t(98) = -6.87$, $p < .001$, Cohen's $d = -1.37$. A robust mixed effect ANOVA using condition (single vs. multi-group) as between-subject factor and measure (prior vs. posterior) as the within-subject factor on the capability estimation indicated a significant main effect of measure, $F(1, 60.74) = 5.92$, $p < .001$, and a significant interaction effect between condition and measure, $F(1, 60.74) = 5.92$, $p = .018$. This interaction indicates that posterior estimates of the alien group' mining capability were lower in the multi-group condition ($M$ = 4.79, $SD$ = 1.62) than in the single-group condition ($M$ = 5.21, $SD$ = 1.09), replicating findings from the simulations that intersectional group memberships made stereotype updating difficult.

## General Discussion

Across three studies with both simulations and human experiments, we investigated the computational principles underlying stereotyping. We formalize stereotyping as a hierarchical inference problem arising when perceivers encounter multiple individuals with known group memberships. Holding the total number of records observed constant, stereotypes were acquired more rapidly when perceivers observed thin slices of information of many group members than when they observe rich information of a few group members (Study 1); stereotypes were harder to update when perceivers observe group members that are heterogenous (Study 2) and when the individual belongs to multiple groups (Study 3). Together, these findings provide a computational explanation for why stereotypes are both easily formed and hard to change despite counterevidence.

Prior research has documented stereotype contents and stereotyping processes like automatic categorization (Fiske et al., 2002; Fiske & Neuberg, 1990), but a complete account also requires specifying why these representations are formed and how they influence social inference. We argued that stereotype representations are hierarchical: individual-level impressions are not only constrained by but also update group-level beliefs. Crucially, this hierarchy characterizes the perceiver's mental representation rather than the true generative structure of social traits. In real social groups, many of which are culturally constructed rather than biologically grounded, such hierarchical structure may not exist. Applying an otherwise efficient hierarchical inference in these contexts can therefore produce systematic biases.

Prior Bayesian accounts of stereotyping proposed that counterevidence can be accommodated through subgrouping (Gershman & Cikara, 2023). We extend this view by showing that even when groups are fixed, hierarchical inference can facilitate stereotype formation and maintenance. A key implication of HBMs is the interaction sparsity hypothesis in Study 1: stereotype acquisition can outpace individual impression formation when perceivers interact superficially with many targets. Statistically, sparse observations within a cluster (e.g., an individual) yield unreliable estimates. HBMs mitigate this by pooling information across individuals to estimate a more precise group-level parameter, which in turn facilitates individual-level learning (McElreath, 2015). This pooling naturally accelerates group-level learning. Consistent with this prediction, participants formed more confident group-level impressions in three of four conditions in Study 1; only when there were two highly heterogeneous individuals did individual-level impressions dominate. These results help explain why stereotypes are readily acquired.

Regarding stereotype maintenance, Studies 2 and 3 explain why stereotypes persist in the presence of counterevidence. In both cases, a shared Bayesian principle applies: to maximize information gain, new evidence is preferentially attributed to update less certain parameters. As a result, counterevidence is attributed to individual-level traits in Study 2 and to alternative, unfamiliar group memberships in Study 3, thereby slowing updates to group-level beliefs. The contingency in Study 2 aligns with findings on perceived category variability, where atypical members reduce stereotypes when they are perceived as typical (Hewstone & Hamberger, 2000). Moreover, because people typically hold multiple group memberships as in Study 3, counterevidence is diluted across representations, slowing belief revision. This relates to subtyping (Maurer et al., 1995), where atypical members are treated as outliers and thus preserve the original belief; however, Study 3 instead emphasizes competing attribution channels rather than the distribution of atypical members. Computationally, this parallels feature competition in instrumental learning (Rescorla, 1972; Schultner et al., 2025), where shared outcomes across predictors weaken learned contingencies..

However, our account is theoretically distinct from an instrumental learning perspective on stereotypes (Schultner, Stillerman, et al., 2024), which characterizes social learning as reward-contingent and conceptualizes stereotypes as estimates of state values (e.g., group traits) from feedback. As shown in Study 2, holding total extracted minerals constant, learning depends on how evidence is distributed across group members and on the perceiver's group-level priors—patterns that cannot be captured by reward-based updating alone. Although instrumental learning studies have demonstrated that prior beliefs modulate learning rates (Schultner, Lindström, et al., 2024), the key contribution of a Bayesian account is the assumption that perceivers form structured representations that support flexible inference and accelerated learning (Griffiths et al., 2024). Without hierarchical representations, instrumental learning cannot readily account for pooling or attenuated contingency effects. Also, despite their key roles in social learning, they are less efficient than Bayesian learning in dynamic environments (Qi & Vul, 2022), where the absence of explicit mental representations limits generalization across contexts and delays updating when conditions change.

The present study also has limitations. First, hierarchical inference alone cannot explain culturally shared stereotypes; a full account requires social transmission and environmental structure, with transmitted beliefs maintained by the computational principles we propose. Second, we used a simplified paradigm in which evidence is numerically represented and directly mapped to inference. In reality, stereotypes are multidimensional, and how evidence is interpreted involves additional inferential processes (Reader & Brewer, 1979). Third, our analysis operates at the computational level and does not specify the underlying psychological processes or neural implementations. Despite these limitations, our work provides a computational account of stereotyping, revealing why and when stereotypes form and persist.

## Acknowledgments

The authors thank Tim Brady, Eshin Jolly, Piotr Winkielman, and Lucy Lai for their constructive feedback. The first author also extends special thanks to Ruoying Zheng, Fangxu Cai, Kangrui Liu, Jiabi He, Mei Yang, Tianxiang Cui, Mitchell Landers, and Max Yang for their valuable feedback and encouragement on this project. The authors used AI assistance (ChatGPT and Claude) to polish the writing this manuscript and to assist in the creation of figures for the experiments.

## References

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: A computational account of negative stereotype persistence. *Journal of Experimental Psychology: General*, *150*(10), 2078–2099. https://doi.org/10.1037/xge0001037

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Bai, X., Fiske, S. T., & Griffiths, T. L. (2022). Globally Inaccurate Stereotypes Can Result From Locally Adaptive Exploration. *Psychological Science*.

Chua, K.-W., & Freeman, J. B. (2022). Learning to judge a book by its cover: Rapid acquisition of facial stereotypes. *Journal of Experimental Social Psychology*, *98*, 104225. https://doi.org/10.1016/j.jesp.2021.104225

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J.-P., Bond, M. H., Croizet, J.-C., Ellemers, N., Sleebos, E., Htun, T. T., Kim, H.-J., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodríguez-Bailón, R., Morales, E., Moya, M., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1–33. https://doi.org/10.1348/014466608X314935

Denrell, J. (2005). Why Most People Disapprove of Me: Experience Sampling in Impression Formation. *Psychological Review*, *112*(4), 951–978. https://doi.org/10.1037/0033-295X.112.4.951

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology* (Vol. 23, pp. 1–74). Elsevier.

Gershman, S. J., & Cikara, M. (2023). Structure learning principles of stereotype change. *Psychonomic Bulletin & Review*, *30*(4), 1273–1293. https://doi.org/10.3758/s13423-023-02252-y

Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (2024). *Bayesian models of cognition: Reverse engineering the mind*. MIT Press. https://www.google.com/books?hl=zh-CN&lr=&id=FEb6EAAAQBAJ&oi=fnd&pg=PR7&dq=bayesian+models+of+cognition+book&ots=RHyzrOHPA8&sig=0c6Ei9msUGg4bOIVA54I43PmHVU

Harris, C., Fiedler, K., Marien, H., & Custers, R. (2020). Biased preferences through exploitation: How initial biases are consolidated in reward-rich environments. *Journal of Experimental Psychology: General*, *149*(10), 1855. https://doi.org/10.1037/xge0000754

Hewstone, M., & Hamberger, J. (2000). Perceived Variability and Stereotype Change. *Journal of Experimental Social Psychology*, *36*(2), 103–124. https://doi.org/10.1006/jesp.1999.1398

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Edition 2). Academic Press.

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379–392. https://doi.org/10.1037/a0023010

Lippmann, W. (1922). *Public opinion*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315127736/public-opinion-walter-lippmann

Maurer, K. L., Park, B., & Rothbart, M. (1995). Subtyping versus subgrouping processes in stereotype representation. *Journal of Personality and Social Psychology*, *69*(5), 812.

McElreath, R. (2015). *Statistical Rethinking*. Chapman and Hall/CRC. https://book.douban.com/subject/26607925/

Qi, W., & Vul, E. (2022). The evolution of theory of mind on welfare tradeoff ratios. *Evolution and Human Behavior*, *43*(5), 381–393. https://doi.org/10.1016/j.evolhumbehav.2022.06.003

Reader, G. D., & Brewer, M. (1979). A Schematic Model of Dispositional Attribution in Interpersonal Perception. *Psychological Review*, *86*(1), 61–79.

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical Conditioning, Current Research and Theory*, *2*, 64–69.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, *60*(2), 63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001

Schultner, D., Lindström, B. R., Cikara, M., & Amodio, D. M. (2024). Transmission of social bias through observational learning. *Science Advances*.

Schultner, D., Molleman, L., & Lindström, B. (2025). Feature-based reward learning shapes human social learning strategies. *Nature Human Behaviour*, (9), 2183–2198. https://doi.org/10.31234/osf.io/zm7ta

Schultner, D., Stillerman, B. S., Lindström, B. R., Hackel, L. M., Hagen, D. R., Jostmann, N. B., & Amodio, D. M. (2024). Transmission of societal stereotypes to individual-level prejudice through instrumental learning. *Proceedings of the National Academy of Sciences*, *121*(45), e2414518121. https://doi.org/10.1073/pnas.2414518121

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202