TESTING IMPRESSION FORMATION FROM A BAYESIAN PERSPECTIVE

By

Prachi Sudhir Solanki

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology – Doctor of Philosophy

2024

ABSTRACT

Can people combine various sources of information when forming impressions of others? Past social cognition research identified two broad information types used during impression formation: individuating and categorical (stereotype) information. Individuating information is about specific individuals' past behaviours or hobbies. Stereotype information is about one's social or demographic classifications like race or age. I aimed to move beyond this traditional distinction — which assumed stereotype information to be the base rate — by exploring if both information types can be used within a Bayesian framework (see McCauley & Stitt, 1978). Across two experiments, I presented categorical information only vs. categorical and individuating information together. Results suggested people's judgements were less aligned with Bayes' theorem when given social categorical information only than when given both, categorical and individuating information together (N = 1130). Broadly, this project aimed to link impression formation work with the larger literature on Bayesian decision-making in cognitive psychology.

ACKNOWLEDGEMENTS

Many people have my eternal gratitude for helping me in this marathon journey. My parents, for believing in me and allowing me to make them proud, no matter the cost. My brother, for always being the wind beneath my wings. My advisor, Joe Cesario, for letting me walk into his office and bother him with questions any time I needed. My committee members, Drs. Bill Chopik, Richard Lucas, and Erik Altmann, for their time, patience, and guidance throughout my dissertation. My lifelong advisor and mother hen, Zach Horne, for setting me up for success – I would be lost without your support.

Special thanks to the following people for their friendship and generosity, for reassuring me, and troubleshooting with me when things got rough: Alejandro and Mariah, my Struggle Bus companions for partaking in venting sessions at the drop of a hat; Nicole, Hunter, Patricia, Sara, Daisy, and Olivia, the OG CCD lab members and my very first friends in USA; Kenya, Ata, Hyewon, Brian, Katie, Jeewon, Jenna, Victor, David, and Abhilasha, my graduate school fam whose presence was invaluable and made the last five years bearable; Anand, Akshay, Azhar, Eishaan, Harris, Heenal, Neha, Shuchi, Utkarsh, and Varun, for making me feel less homesick and forcing me to take breaks; Siddhi, Rucha, Abbas, Aaron, Adwaita, Damini, Mugdha, and Seerat – some of my oldest friends who cheered me on through the years.

And finally, Karan – for being my partner, my rock, my emotional support animal, and the most annoying human being on this planet only to help me through the countless slumps. I'm thankful to have 8,000 layers of inyeon with you.

TABLE OF CONTENTS

INTRODUCTION

How do people combine various sources of information when forming impressions of others? Past social cognition research has identified two broad types of information that can impact impression formation: individuating information (personality or behavioral information about an individual) and categorical information (stereotype information based on social groups; Ajzen & Fishbein, 1977; Brewer, 1988; Ginossar & Trope, 1987; Hilton & Fein, 1989; Kahneman & Tversky, 1973; Locksley et al., 1982; Nelson et al., 1990). Individuating information has been defined as information about a specific individual and can include their past behaviours, preferences, hobbies, appearances, and so on (Ginosar & Trope, 1980; Kunda & Sherman-Williams, 1993). Categorical information has been understood as information about a person's social or other demographic classifications, for example their race, sex, age, occupation, religion, and so on. The social cognition literature generally considers stereotypes as base rate information (i.e., as relative distributions of the occurrence of certain events among certain populations) and differentiates it from individuating information (e.g., Ginosar & Trope, 1980; Kunda & Sherman-Williams, 1993; Kutzner & Fiedler, 2017). From this view, stereotypes can also be thought of as information about the prior probability of certain traits existing among certain social groups (e.g., the likelihood that a person who is German is also efficient; Ginosar & Trope, 1980; Hinsz et al., 1988). Relatedly, the judgement and decision-making literature has long studied how people weigh individuating vs. base rate information in making predictions (e.g., Kahneman & Tversky's base rate fallacy).

As an illustration, consider a typical social-cognitive impression formation study, in which participants are given information about target person Tom and asked to render a judgement about him. Participants might be told that Tom, a German, was occasionally late,

often failed to complete his work before it was due, and sometimes procrastinated. Participants would then rate the extent to which they believed Tom was industrious. In such a study, researchers frame the task as one in which participants are presented with two *qualitatively distinct* pieces of information: categorical information (stereotype information of Germans being industrious) and individuating information (concrete pieces of evidence specific to Tom as a person). In this literature, categorical information is often considered a "prior" or base rate, indicating people's pre-existing beliefs before taking into consideration any new information when forming judgements. Individuating information is often considered to be the new information and is manipulated to determine its impact on people's impressions of a target in comparison to the categorical information. In the above example, to the extent that participants rate Tom as being relatively more industrious, they are said to be using categorical or stereotype information; to the extent they rate Tom as being relatively less industrious, they are said to be using individuating information. The current project aimed to begin moving beyond this traditional distinction between categorical vs. individuating information by investigating whether all types of information can be understood within the same Bayesian framework. Specifically, this work tested whether people integrate both categorical and individuating information in line with Bayes' theorem when forming impressions of a target person. In doing so, this work also aimed to link social-cognitive work in impression formation with the broader literature on Bayesian decision-making in cognitive psychology.

Researchers have often presumed that categorical and individuating information impact impression formation via separate cognitive processes (Brewer, 1988; Fiske & Neuberg, 1990; Gilbert et al., 1988; Hugenberg et al., 2010; Trope, 1986; but see Pennycook et al., 2014). When processing categorical information, individuals are thought to use abstract or broad information

(e.g., “industrious” as derived from the stereotype of Germans) and apply it to the specific instance of a particular individual (e.g., Gilbert et al., 1988; Trope, 1986). This requires making an inference about a particular target person based on knowledge about others who are like them or who belong to the same categories or groups. Conversely, individuating information requires using concrete pieces of information about an individual (e.g., being late, procrastinating) and integrating those pieces to form conclusions about their personality and/or their possible group membership. Brewer’s (1998) Dual Process Model explicitly classifies these processes into top-down and bottom-up, respectively. More specifically, Brewer states that top-down processing occurs when there is low personal involvement and perceivers adopt an inter*group* orientation to form an impression of a target whereas bottom-up processing occurs when there is high personal involvement, and perceivers adopt an inter*personal* orientation to form an impression of a target. The model goes on to predict that stereotyping occurs when information about a target is aligned with category information, whereas individuation (i.e., forming a distinct mental representation of an individual based on specific information about them) occurs when information about a target is unaligned with category information (Brewer, 1988). This prediction assumes that stereotype information about a target (i.e., categorical information) is more easily integrated into an impression than non-stereotypic information. That is, this model assumes that categorical information is usually processed more easily than individuating information. This is partly because the model assumes individuating information is still somewhat influenced by categorical information – when exposed to a target, people form an initial impression using category information which later acts as a filter for category-related vs. category-unrelated information when forming judgements (Brewer, 1988).

The important aspect to appreciate of Brewer's model and others like it (e.g., Fiske & Neuberg, 1990; Rumelhart & Ortony, 1977) is that qualitatively different types of information are paired with qualitatively different cognitive processes to explain how these information types are used and how they influence impression formation. However, an alternative is to approach the question of impression formation through a Bayesian lens. The Dual Process Model and other approaches rely heavily on the premise that individuating information can be empirically differentiated from categorical information in some way, which has various limitations (a point I return to in the General Discussion). More important, however, are the potential advantages in explaining impression formation with the same model that can describe many other types of judgments, namely Bayes' theorem. I now turn to a discussion of the Bayesian approach to connect and reinterpret past work in terms of Bayes' Theorem.

Reinterpreting Individuating vs. Categorical Information Using Bayes' Theorem

Bayes' theorem is the normative model for integrating new information with existing beliefs to make future predictions. In general, it uses conditional probabilities to describe how people should integrate new information with their prior (existing) beliefs to generate optimal, updated beliefs. Under the Bayesian view, there is no need to explicitly define or differentiate information type, despite past impression formation work having differentiated between categorical and individuating information. For instance, in Fiske and Neuberg's (1990) Continuum Model, the authors assume that people give higher priority to category-information over individuating information because people are motivated to maintain their initial category-based impressions of a target. From a Bayesian perspective, this would be reinterpreted as people wanting to form judgements most aligned with their priors (e.g., information presented before forming a new judgement). Bayes' theorem does not require making an explicit distinction

between categorical and individuating information, but rather considers both to have the same potential to be considered either a prior or new information.

In order to connect the impression formation literature with a Bayesian approach, I first describe the research originating from McCauley and Stitt (1978), who introduced the possibility that judgements involving stereotypes can be understood as Bayesian probabilistic predictions that distinguish one group from another (though this line of reasoning has since been lost within the social-cognitive literature on stereotyping; Stangor, 2016). Bayes' Theorem is the normative model for finding conditional probabilities and is represented by the equation below:

$$p(\mathrm{A}|\mathrm{B}) = \frac{p(\mathrm{A}) \times p(\mathrm{B}|\mathrm{A})}{p(\mathrm{B})} \quad (1)$$

This equation contains 4 components:

1. p(A|B) = the posterior probability: probability of event A given event B
2. p(A) = the prior probability of event A
3. p(B|A) = the probability of event B given event A and
4. p(B) = the probability of event B

where #3 and #4 make up the likelihood ratio. To find the posterior probability of event A given event B, we would multiply the prior by the likelihood ratio. Stereotype application can also be understood as probabilistic predictions following Bayesian reasoning using conditional probabilities. This becomes clearer if we rewrite Bayes' theorem wherein the conditional probability signifies the probability of a certain trait occurring within a certain social group:

$$p(\mathrm{Trait}|\mathrm{Group}) = \frac{p(\mathrm{Trait}) \times p(\mathrm{Group}|\mathrm{Trait})}{p(\mathrm{Group})} \quad (2)$$

Thus, in Equation 2 the probability of A given B represents the probability of a certain trait being associated with a certain social category (e.g., the probability of an individual being

industrious given that they are German). Specifically, McCauley and Stitt (1978) wanted to test whether people's probabilistic predictions about social groups were appropriately Bayesian. They presented various stereotypic and non-stereotypic traits about Germans and asked participants to report probabilities corresponding to each of the four components of Bayes' theorem for each trait. For example, 1) what percent of Germans are industrious or p(Trait | Group), 2) what percent of all the world's people are industrious or p(Trait), 3) what percent of industrious people are German or p(Group | Trait), and 4) what percent of all the world's people are German or p(Group). They called people's responses to Question 1 (component #1 in Equation 1) a direct posterior judgement or "judged posterior" and compared this value to a calculated posterior (derived by using components #2, #3, and #4). Results showed that participants' judged and calculated posteriors were highly correlated (r = 0.91), and the authors interpreted this as people's judgements being aligned with the Bayesian normative standard. In a conceptual replication and extension of this work, Solanki and Cesario (2024) used the same methodology to test this question using multiple social groups and updated stereotypes. This research also found a large correlation between the judged and calculated posterior predictions (r = 0.79 to 0.89), suggesting people's predictions about social groups were aligned with Bayes' theorem.

In these studies, participants were essentially asked to rank order multiple traits (stereotypic and non-stereotypic) associated with a given social group to explore if their judgements generally aligned with Bayes' theorem across groups. The current work took a slightly different approach testing the extent to which people's judgements across multiple *individuals* aligned with Bayes' theorem. That is, participants made an inference about a specific target possessing a given trait across multiple targets. Further, participants had to form these

judgements based on different types of information (i.e., categorical, individuating) about each target. For instance, participants had to judge the probability that a specific individual (Tom) possesses a certain trait (industrious) given that he belongs to a certain social group (German). This probability can be measured just as easily using Equation 2 as the probability that a certain group (Germans) possess a certain trait (industrious). Using Bayes' theorem in this manner helped explore the main question of the current work: can people combine and use different types of information when forming impressions of others? Specifically, I tested the extent to which people's predictions aligned with Bayes' theorem when forming impressions about various targets given different types of information about each target.

Rasinski et al. (1985) attempted to answer a similar conceptual question within a Bayesian framework, though these researchers used a different computational method from the current work. The authors reexamined Locksley et al.'s (1980) experiment in which participants were asked to judge the likelihood that a target was assertive across one of three conditions: categorical information only vs. categorical plus non-diagnostic individuating information vs. categorical plus diagnostic individuating information.[1] For example, they were asked to rate if a target was assertive based on their sex (categorical information: male or female) and a description of them behaving in an assertive or unassertive way (individuating information: complaining to a store manager or buying a book). Locksley et al.'s results suggested that people relied on diagnostic individuating information to form impressions when such information was available, but used categorical information when the individuating information was non-

[1] The degree to which a certain piece of information is indicative or relevant to forming an impression has been generally labelled as information "diagnosticity" in past work. However, past work has not attempted to operationally define diagnosticity as a variable within their experiments. In the current proposal, I did not aim to define diagnosticity but simply determine whether a piece of information is indicative of a certain trait; I did so based on pretest findings.

diagnostic or unavailable. Based on this finding, Locksley et al. concluded that people do not neglect base rates and that even a minimal amount of diagnostic (individuating) information is sufficient to eliminate the effects of stereotypes on people's impressions. Rasinski et al. (1985) decided to reexamine this finding as they noted two limitations in Locksley et al.'s (1980) experiment.

First, Locksley et al. concluded that people commit base-rate fallacy because they rated both targets similarly on assertiveness when there was diagnostic information available, but they rated male targets higher on assertiveness than female targets when given only categorical information. That is, people *should* have been influenced to some degree by stereotypes (base rates) even in the presence of individuating information; the fact that stereotypes had no effect under these conditions led Locksley to conclude that participants were showing base-rate fallacy. Rasinski et al., however, contend that Locksley et al.'s conclusion of base-rate neglect was unwarranted because participants' prior beliefs (about assertiveness in males vs. females) were unknown and no normative standard was established against which to compare how people integrated categorical and individuating information. Without these, a proper test of base-rate neglect cannot be conducted. (After all, one cannot conclude that people are failing to incorporate base-rate information without first establishing the precise base rates). They claimed that in the studies where Locksley et al. did attempt to construct an explicit normative criterion, the criterion was based on the *authors'* prior probability estimates about the existence of assertiveness among males vs. females rather than the participants' prior estimates.[2] Second, Locksley et al. assumed that the individuating information presented in their studies would be

[2] However, Locksley et al. measured participants' priors in Experiment 2 by asking participants to report probabilities for assertive behaviour in general and assertive behaviour in males and females, respectively. They compared these priors to participants' responses predicting a specific target's assertiveness.

considered equally diagnostic for both male and female targets. Rasinski et al. instead argue that some people might interpret a trait such as assertiveness to be generally more normative among males than females (see Attribution Theory; Jones & Davis, 1965). That is, whether certain behaviours are diagnostic of certain traits is more likely to be based on people's subjective judgements of a target, and Rasinski et al. suggest that such subjectivity needs to be considered. Thus, they aimed to replicate Locksley et al.'s Study 2 by measuring people's prior beliefs and comparing participants' responses to a normative standard set using components of Bayes' theorem.

Rasinski et al. (1985) used the same three conditions (categorical information only vs. non-diagnostic individuating information vs. diagnostic individuating information) and asked participants to rate a target's assertiveness. Again, the target belonged to one of two gender categories (male vs. female). The normative standard was set by measuring participants' stereotype-driven prior beliefs that a target 1) belonging to a certain group possessed a certain trait (e.g., the probability of males [vs. females] being assertive) and 2) engaged in a certain action (e.g., told someone who was bothering them to go away, got a haircut, etc.). In the condition where participants had to form impressions based only on group membership (categorical) information, the posterior probability was computed using Equation 2 above. When both group membership and behavioural information was presented, two separate posteriors were calculated for each category (i.e., male and female), using the equations below.

$$p(\text{Trait}|\text{Behaviour})_{Male} = \frac{p(\text{Trait}) \times p(\text{Behaviour}|\text{Trait})}{p(\text{Behaviour})} \tag{3.1}$$

$$p(\text{Trait}|\text{Behaviour})_{Female} = \frac{p(\text{Trait}) \times p(\text{Behaviour}|\text{Trait})}{p(\text{Behaviour})} \tag{3.2}$$

Results replicated Locksley et al., indicating no significant difference in participants' assertiveness ratings between males and females in the condition with diagnostic information and males being rated higher than females in the condition with only categorical information. However, people perceived assertiveness as differentially diagnostic for male targets compared to female targets, with behavioural information being judged as more diagnostic for females. When investigating how people integrated categorical and individuating information to form an impression, participants were overcautious in revising their stereotype-based judgements and contrary to Locksley et al.'s conclusions, did not appear to commit base-rate fallacy. That is, when people had to integrate categorical and individuating information, their judgements deviated systematically less than predicted by the Bayesian normative model, suggesting that people did not simply disregard their stereotypes when given diagnostic individuating information. Rasinski et al. concluded that the results showed no evidence that people ignored or underused stereotypes (as they would if they were committing base-rate fallacy).

Note that Rasinski et al. (1985) computed and compared two separate posterior probabilities for each group (male vs. female) when given categorical and individuating information (see Equations 3.1 and 3.2 above). However, here I tested whether people appropriately combined information when forming impressions of others using joint probabilities within Bayes' theorem. In addition to this difference from past work, the current work also addressed a particular limitation that Rasinski et al.'s work shares with other research in this area. Namely, past work assumes stereotype (categorical) information to be the base rate or prior whereas I argue that *any* information types can be combined to constitute a prior. For example, the probability that Tom possesses a certain trait (posterior prediction) given that he belongs to a certain social group (categorical information) and behaves in a certain way (individuating

information) can be estimated using conjoint probabilities in Bayes' theorem. The probability of Tom being smart can be estimated using conjoint probabilities if we know he is Asian *and* that he got an A in math in 5th grade. This can be generally represented using Equation 4.1 and/or Equation 4.2 below.

$$p(\text{Trait}|\text{Behaviour},\ \text{Group}) = \frac{p(\text{Trait}|\text{Group}) \times p(\text{Behaviour}|\text{Trait},\ \text{Group})}{p(\text{Behaviour}|\text{Group})} \quad (4.1)$$

Equation 4.1 can be rewritten to represent information type, as shown below.

$$p(\text{Trait}|\text{Individuating},\ \text{Categorical}) = \frac{p(\text{Trait}|\text{Categorical}) \times p(\text{Individuating}|\text{Trait},\ \text{Categorical})}{p(\text{Individuating}|\text{Categorical})} \quad (4.2)$$

Given this reinterpretation, we can then ask whether people's impressions are appropriately Bayesian—that is, do people update their beliefs about others as per Bayes' theorem when given multiple types of information (categorical vs. individuating)? The next section details how my reinterpretation (see Equations 4.1 and 4.2) can be applied in typical impression formation experiments.

Linking Past Impression Formation Work to Bayes' Theorem

From a Bayesian perspective using joint probabilities, Locksley et al.'s (1980) experiment would need to include the probability of a target being assertive given their sex *and* some assertive behavior exhibited by the target. This can be written out as follows (McCauley, 1994; McCauley & Stitt, 1978):

$$p(\text{Assertive}|\text{Behaviour}, \text{Sex}) = \frac{p(\text{Assertive}|\text{Sex}) \times p(\text{Behaviour}\,|\,\text{Assertive},\text{Sex})}{p(\text{Behaviour}|\text{Sex})} \quad (5)$$

where a probability conditioned on two cues is separated by commas, e.g., p(Behaviour| Assertive, Sex) and can be read as the probability of certain behaviour occurring given *both* the target's sex and that they possess traits associated with being assertive. In Equation 5 above, the probability that a target is assertive is determined by using their sex or category membership, such that p(Assertive|Sex) is the prior. In other words, the base rate of the target's assertiveness is conditional on their sex – the categorical information about them. However, as I have shown above, Bayes' theorem allows any information to be considered a prior. The behavioural description of the target (i.e., yelling at someone) could also be considered a prior such that p(Assertive|Behaviour) is the prior and the new information would be based on sex. The full Bayesian accounting for this case would be:

$$p(Assertive|Behaviour, Sex) = \frac{p(Assertive|Behaviour) \times p(Sex|Assertive, Behaviour)}{p(Sex|Behaviour)} \quad (6)$$

where the new information is a probability conditioned on two cues i.e., p(Sex| Assertive, Behaviour) and can be read as the probability of a target belonging to a certain sex given both, their specific behaviour and that the target possesses traits associated with being assertive; and the base rate of the target's assertiveness is conditional on the (individuating) information that they yelled at someone. Although the above formulas used different priors (one representing traditionally categorical information and the other representing traditionally individuating information) and different new information, the two Bayesian predictions are equivalent. If both are correct, the determination of what is a prior and what is new information is arbitrary. It follows, as shown, that *either* categorical or individuating information can serve as

the prior or as new information. Thus, studying impression formation from a Bayesian lens can supersede the need for a stark differentiation between information type.

Past work on impression formation has used a specific, standard methodology. First, participants were given categorical information (under the assumption they were aware of stereotypes associated with those categories). Next, they received individuating information which varied in terms of how diagnostic the behaviours were of a certain trait. For example, Krueger and Rothbart (1988) measured participants' ratings of a target's aggressiveness. The target was first described either as a construction worker or housewife (categorical information varying in the stereotype of aggressiveness; e.g., Denson et al., 2018; Frodi, McCauley, & Thome, 1977; Maccoby & Jacklin, 1974). Next, individuating information that varied in terms of aggressiveness (strong vs. moderate vs. neutral) was provided. For instance, the neutral individuating information was the target "recently bought the latest book of a bestselling author," the moderate individuating information was the target "complained to a store manager about the quality of a product," and the strong individuating information was the target "beat his/her child." Participants then rendered a judgement on the target's aggressiveness. Similarly, Hilton and Fein (1989) tested whether people's judgements of a target's assertiveness changed based on different individuating information. Again, the target belonged to one of two social categories (male vs. female) and participants were given behavioural information about the target varied based on three levels (irrelevant vs. pseudo-relevant vs. relevant). Assertiveness was used as a trait because it is also associated with sex stereotypes such that males in general have been typically rated as more assertive than females (e.g., Hentschel et al., 2019; but see Park et al., 2016). As a final example, Locksley et al. (1982) asked participants to provide ratings about targets based only on social category information versus two levels of individuating information

(less vs. more relevant). The category they chose was nocturnal versus diurnal and they asked participants to rate whether the target possessed (or did not possess) certain traits based on the presented information.

All the above experiments found a similar pattern of results: participants generally disregarded categorical information when any individuating information was available, and the authors ascribed this effect to the experimental manipulation of the diagnosticity of individuating information. However, these illustrative experiments also suffer from the same issues as most other work on this topic. Mainly, they differentiate starkly between individuating and categorical information. Instead, I proposed to understand these findings from a Bayesian view of belief updating where information type does not require such a differentiation.

One way to test this question and reconcile past issues is to present individuating and categorical information *simultaneously* in the same vignette to test whether people's explicit predictions across various targets follow the Bayesian instruction (i.e., whether people are using information as Bayes' theorem says they should). The current methods were inspired from past work (McCauley & Stitt, 1978; Solanki & Cesario, 2024) and tested whether people's posterior predictions matched what Bayes' theorem states their predictions should be when forming impressions about individuals. For example, if a participant directly reported their judgement about a target given stereotype information, "Tom is Asian. How likely is it that Tom is smart?" this would be termed their "judged posterior" prediction and signify the extent to which they believe Tom is smart (i.e., people directly reported responses to component #1 in Equation 1). Suppose the participant says it is 80% likely. Further, if the participant were also asked to report their estimations for the other components that map onto Bayes' theorem (i.e., the prior and likelihood ratio; components #2, #3, and #4 in Equation 1), one could also *compute* a posterior

prediction. Let's say the participant's responses to components 2, 3, and 4 of Equation 1 are 60, 70, 50 respectively. The calculated posterior would be 84, as per Equation 7 below.

$$\frac{60 \times 70}{50} = 84 \tag{7}$$

Thus, the participant's direct judgment (80) and their calculated judgement as per Bayes' Theorem (84) would be largely aligned. In this way, if people's direct posterior judgements (responses to component #1 in Equation 1 or their *judged* posteriors) were to correlate highly with their computed posteriors across different targets, this would indicate that people generally formed impressions aligned with Bayes' theorem (see McCauley & Stitt, 1978). Using conjoint probabilities, this reasoning can be extended to cases where people receive two types of information (individuating and categorical) about a given target (see Equations 4.1, 4.2, 5, and 6). Again, in contrast to previous studies where participants rank ordered the probability of a given trait being stereotypic of a certain social group, the current approach tests the probability of an individual possessing a certain trait across multiple target individuals. This method allowed me to consider 1) how people generally formed impressions of individuals and 2) whether people could combine different types of information to generate predictions across multiple individuals. The current work thus aimed to provide a preliminary, formal test of whether adopting a Bayesian perspective can help reconceptualize past work that starkly differentiates between categorical and individuating information during impression formation.

This reconceptualization would be consistent with the stereotyping literature and with a Bayesian view of belief updating thus providing norms against which the rationality (or lack thereof) of belief updating during impression formation can be tested. Below I elaborate on the methodology used to test these hypotheses. To summarize, the current work aimed to move

beyond traditional social-cognitive approaches, specifically the classification of information as individuating vs. categorical, in two ways: 1) to test the degree to which people's judgements about individuals followed Bayes' theorem and 2) to test the degree to which people were able to combine different types of information when forming impressions about individuals within this framework.

PRETEST

I conducted pretest experiments using participants from Michigan State University's human-subjects pool (N = ~200). There were two goals of the pretest tested across two separate experiments. The first experiment (N = 100) tested the degree to which certain traits were stereotypic of the social groups being used in the main experiments. For example, if being aggressive was a trait stereotypically associated with being a construction worker or if being greedy was stereotypic of being a politician. This was assessed using two questions, "What percent of all males do you think are aggressive?" and "What percent of all the world's people do you think are aggressive?" on a scale of 0 to 100 (0 = 0% and 100 = 100%). I used responses to these questions to calculate a *diagnostic ratio* (*DR*). McCauley and Stitt (1978) proposed that a diagnostic ratio is a stereotype measure to determine which traits are stereotypic of which social groups. *DR*s were calculated by dividing a participant's direct posterior prediction (component #1) by their prior (component #2) in Equation 1. *DR*s above 1.0 indicated that a trait was considered stereotypic of a certain social category and *DR*s below 1.0 indicated that a trait was considered non-stereotypic for that category. I have previously used this method to identify traits stereotypic of different social groups (Solanki & Cesario, *under review*). Here, this method again helped me determine which traits were considered stereotypic of which social categories. The *DR*s for all categories except "construction worker" were above 1, suggesting that participants considered most traits to be stereotypic of a given social group (see Table A1 and Figure A1 in Appendix A). The "construction worker" category was excluded from the main experiments.

The second goal of the pretest was to get behaviours linked to certain traits. For this, I asked participants to list two actions associated with a given trait, as shown in the example below:[3]

> Please list two behaviours or actions that indicate kindness. Kind (*adjective*) is defined as *caring about others*.
> **What does it mean for a behaviour to indicate kindness**? A behaviour that indicates kindness is **a behaviour that typically kind people would perform or a behaviour that makes you think a person is *DEFINITELY* a kind person**. For example, volunteering at an eldercare facility is a behaviour that *definitely* indicates kindness.

I used the responses provided to ask a different sample of participants (N = 100) to rate questions corresponding to the Diagnostic Ratio. I did this to determine whether a given behaviour/action was indicative of a given trait (e.g., aggressiveness). For example, if pushing someone was mentioned as indicative of aggressiveness, I asked participants to report, "What percent of all people who pushed someone do you think are aggressive?" and "What percent of all the world's people do you think are aggressive?". Responses were measured on a scale ranging from 0 to 100 (0 = 0%, 100 = 100%). *DR*s above 1.0 indicated that a behaviour was indicative of a certain trait and *DR*s below 1.0 indicated that a behaviour was not indicative of that trait. This method helped me determine which behaviours were considered indicative of which traits. Behaviours with the largest *DR*s were included in the main experiments (see Table A2 in Appendix A).

[3] A shorter version of the pretest (N = 30) was used to check whether these instructions were worded clearly for participants to report that a given behaviour was indicative of a certain trait.

GENERAL METHODOLOGY

My overall hypothesis was that when given different types of information (categorical vs. individuating), people's impressions about individuals will generally follow Bayes' theorem. That is, their judged and calculated posterior predictions will be highly correlated when making judgements about certain *individuals* possessing certain traits. This hypothesis was informed by past work which found that people's judged and calculated posteriors were highly correlated when making judgements about certain *social groups* possessing certain traits (McCauley & Stitt, 1978; Solanki & Cesario, 2024). I conducted two experiments. Participants from Michigan State University's (MSU) online participant pool received course credits in exchange for their participation. Their task was to make judgements about 12 unique targets in each experiment given either categorical information (Experiment 1) or both, categorical and individuating information (Experiment 2), about a target.

Experiment 1 was geared to provide foundational confirmation that people would judge specific individuals in accordance with Bayes' theorem within a very simplified decision task (i.e., only getting one piece of information). This experiment helped set up for Experiment 2, in which people's impressions of target individuals would be based on a combination of different information types (i.e., categorical, individuating). Both experiments used the same methodology to test if people integrated information as per Bayes' theorem and data was analyzed in the same way as past work (McCauley & Stitt, 1978; Solanki & Cesario, *under review*). Across both experiments, I measured two things: 1) participants' direct posterior judgements regarding a target individual or their "judged posteriors" and 2) separate components of Bayes' theorem i.e., the prior and likelihood ratio. For the judged posteriors, I directly asked participants to report their judgements about a certain target given only categorical information in Experiment 1 and

given two types of information (categorical, individuating) in Experiment 2. For example, in Experiment 2 I presented information that, “Tom is Asian and Tom got an A in math in undergrad” and asked participants, “Considering only this information, how likely is it that Tom is smart?” I measured participants’ prior beliefs by asking, “What percent of all Asian people do you think smart?” (component #2 in Equation 1; see McCauley & Stitt, 1978). I measured components of the likelihood ratio by asking questions like, “What percent of all Asians who are smart do you think got an A in math in undergrad?” and “What percent of all Asians do you think got an A in math in undergrad?” (components #3, #4 in Equation 1). These independent components (prior, likelihood ratio) were used to compute a posterior using Bayes’ theorem (see Equation 1). The computed or calculated posterior values were then correlated with the judged posterior values to test the degree to which they aligned. A large correlation between these two posteriors would suggest that participants’ impressions were highly aligned with the Bayesian normative criterion in this task. In this way, I tested whether people’s impressions of individuals possessing certain traits followed Bayes’ theorem given two different types of information.

EXPERIMENT 1

Here, I tested my general hypotheses that people's impressions will generally follow Bayes' theorem – their judged and calculated posterior predictions will be highly correlated. Participants were asked to form impressions of various targets given categorical information only. I did this by using materials from four classic social-cognition experiments: Krueger and Rothbart (1988), Hilton and Fein (1989), Kunda and Sherman-Williams (1993), and Locksley et al. (1982), my past work (Solanki & Cesario, 2024), and the pretest (see Table A1, Figure A1 in Appendix A). The classic studies were chosen as they have been widely cited in the literature linked to categorical vs. individuating information use and similar methods have often been used to study impression formation.

Methods

Here, each participant was presented 14 trials with categorical information about 14 unique targets within each trial. Participants' task was 1) to form impressions about each target possessing a certain trait given the categorical information and 2) to provide judgements about questions corresponding to Bayes' theorem in each trial (i.e., the prior, likelihood ratio). Specifically, I asked participants to provide judgements corresponding to each component of Bayes' theorem across target individuals given their social category. This method was similar to the one used in my past work but focused on measuring judgements about specific individuals rather than social groups (Solanki & Cesario, 2024).

Participants, Pre-registration, & Analysis Plan

For a within-subjects design and 90% power to detect a small effect ($\rho = 0.15$), I needed a sample of 462 participants (as per G*power 3.1)[4]. However, to have sufficient data after accounting for ~10% to 15% data loss due to incomplete responding or poor data quality, I collected data from 590 participants from Michigan State University's human-subjects pool (Md_{age} = 19.66, M_{age} = 19.96, SD_{age} = 1.91; Female = 74.4%, Male = 23.2%, Non-binary/Other = 1.86%, Prefer not to say = 0.51%; White/European American = 67.3%, Asian/Asian American = 13.7%, Black/ African American = 6.1%, Hispanic/Latino = 4.24%, Middle Eastern = 4.24%, American Indian/Alaska Native = 0.17%, Native Hawaiian/ Pacific Islander = 0.17%, Other = 3.05%, Prefer not to say = 1.02%). Data was analyzed using Pearson's product moment correlation. Following similar analysis techniques as past work (McCauley & Stitt, 1978; Solanki & Cesario, 2024), the correlation between judged and calculated posteriors was computed after aggregating responses across each participant.[5] This study was preregistered and all data, analysis scripts, manipulations, and measures have been made available on the Open Science Framework (https://osf.io/aqxjk/?view_only=dd5dc367fd024872b9e88b5b07335163).

Procedure

All participants were presented one piece of traditionally categorical information about 14 unique targets in a within-subjects design. Their task was to provide ratings of all the relevant component pieces of Bayes' theorem as they pertained to their impression of the given target. For instance, participants rated the target's aggressiveness given that the target was a construction worker (Krueger & Rothbart, 1988).

[4] Specifically, the information entered on the G*power app to calculate sample size was as follows: Under the Exact test family, bivariate normal model for a two-tailed test was chosen. Correlation ρ H1 = 0.15, α = 0.05, Correlation ρ H0 = 0.

[5] However, additional analyses on unaggregated data is presented in the Appendix for both experiments.

An example item from the survey is given below wherein the categorical information is being a construction worker and the trait is "aggressive" (see Appendix B for full set of items):

1. Person 1 is a construction worker. Considering only this information, how likely is it that Person 1 is aggressive?[6]
2. What percent of all the world's people do you think are aggressive?[7]
3. What percent of all aggressive people do you think are construction workers?
4. What percent of all the world's people do you think are construction workers?

where question 1 measured the judged posterior prediction and questions 2 to 4 measured the components of Bayes' theorem (prior, likelihood ratio; see Equations 4.1 and 4.2).[8] Participants responded to the above questions using a 0 to 100 scale.[7,7] Questions 2 to 4 were used to calculate a posterior and correlate this value to the judged posterior (i.e., responses to Q1 above). If the correlation between judged and calculated posteriors was large, it would suggest that participants' responses about specific individuals followed what Bayes' theorem says their responses should be. As exploratory analyses, I also 1) correlated each of the other questions (#2, #3, and #4 above) with the judged posterior to test the magnitude of these correlations, 2) computed difference scores between the judged and calculated values to more directly examine the extent to which participants' judgements differed from their judgements as per Bayes' Theorem, and 3) examined whether the correlation between judged and calculated posteriors would differ for individuals with relatively strong vs. weak stereotype beliefs.

Results

The means and standard deviations for all components corresponding to Bayes' Theorem for the overall sample are presented in Table 1. Response distributions across all components of

[6] Response scale for this question ranged from 0 to 100 where 0 = highly unlikely ("I'm certain this person is unaggressive"), 50 = uncertain ("I'm uncertain about this person's level of aggressiveness"), and 100 = highly likely ("I'm certain this person is aggressive").
[7] Response scale for questions 2, 3, and 4 ranged from 0 to 100 where 0 = 0%, 50 = 50%, and 100 = 100%.
[8] Where question 2 measured the prior probability, questions 3 and 4 measured the likelihood ratio.

Bayes' Theorem are displayed in Figure 1. Calculated posterior values had a relatively large range and standard deviation, which may be due to absolute errors in participants' judgements in this task. To account for extreme calculated posterior values, the correlation between judged and calculated posteriors is also reported after restricting the calculated posterior values to range between zero to 100.

Table 1

Means and standard deviations for all components of Bayes' Theorem in Experiment 1 using unaggregated (raw) data.

Questions	Means	SDs
Judged posterior (pTraitGroup)	60.001	21.259
Prior (pTrait)	43.360	21.002
Likelihood ratio numerator (pGroupTrait)	44.377	25.529
Likelihood ratio denominator (pGroup)	45.759	14.631
Calculated posterior (calc_pTraitGroup)	76.665	178.791
Calculated posterior (calc_pTraitGroup)*	44.672	25.670

Note. * indicates Means and SDs after restricting the calculated posterior values between 0 to 100.

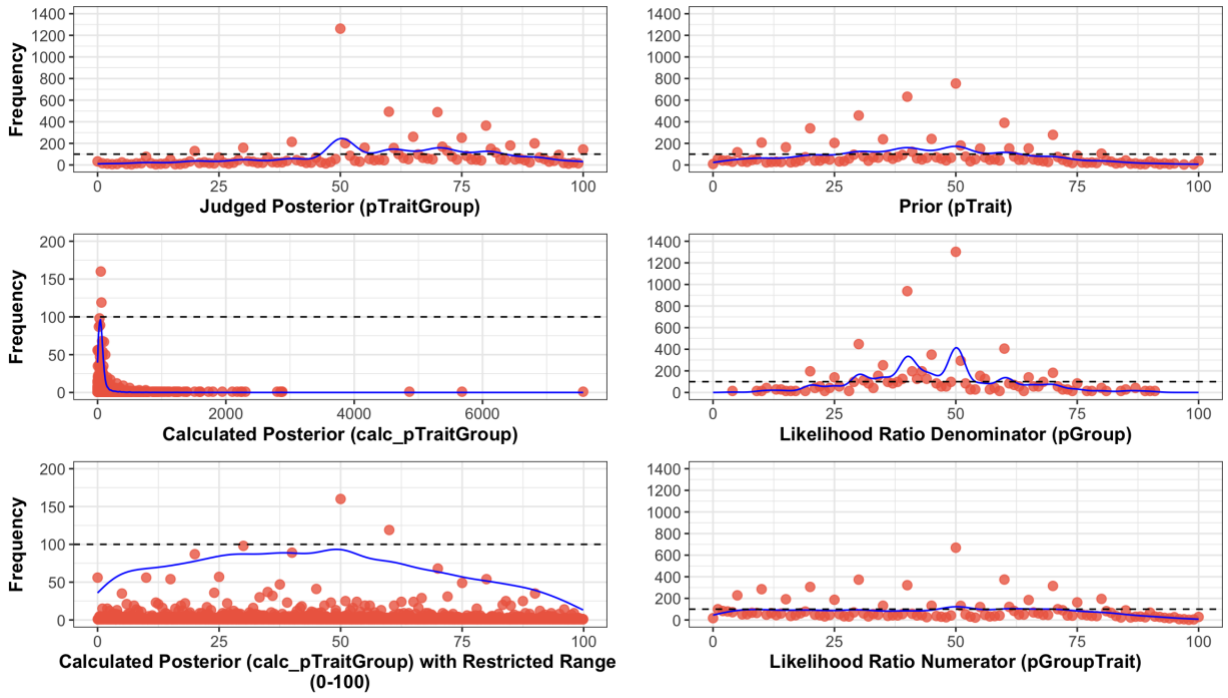

Figure 1
Scatter plots and kernel densities depicting frequencies of responses across all components of Bayes' Theorem in Experiment 1 using unaggregated (raw) data.

The main hypothesis focused on the correlation between judged and calculated posterior (henceforth referenced as r_{JC}) after aggregating observations across each participant.[5] A large r_{JC} correlation would suggest that participants' judgements about specific targets were highly aligned with what Bayes' theorem says their judgements should be. The r_{JC} correlation was small, $r_{JC} = 0.24$, $p < .001$, $t(588) = 5.98$, *95% CI* = [0.61, 0.31], suggesting weak alignment between participants' direct judgements and judgements prescribed by Bayes' theorem in this task.

Exploratory Analyses

I also checked 1) correlations between all other questions in this task (i.e., questions corresponding to the priors, likelihood ratio), 2) how closely participants' judged and calculated posteriors were aligned using a difference score, and 3) if the magnitude of the r_{JC} correlation

would be smaller for individuals with consistently held stereotype beliefs (i.e., low variability across responses to the component parts of Bayes' Theorem) compared to individuals with inconsistent stereotype beliefs (i.e., high variability across responses to the component parts of Bayes' Theorem). The first exploratory analysis tested whether other questions included in the task substantially impacted the r_{JC} correlation reported in the main analysis (e.g., particularly large *r* values). Table 2 lists the various combinations of correlations for the questions corresponding to the priors and likelihood ratio. Most of these correlations were slightly larger than or within the same range as the r_{JC} correlation, suggesting no particular impact of these effects on the r_{JC} correlation.

Table 2
Overall correlations between all combinations of questions corresponding to the prior, likelihood ratio, judged posterior, and calculated posterior probabilities in Experiment 1.

Questions	*r*	95% Lower CI	95% Upper CI	*p* value
pTraitGroup – pGroupTrait	0.560	0.502	0.613	< .001
Calculated_pTraitGroup – pGroup	-0.523	-0.583	-0.467	.049
pGroupTrait – Calculated_pTraitGroup	0.397	0.326	0.462	< .001
pTrait – pGroupTrait	0.323	0.249	0.393	< .001
pTraitGroup – Calculated_pTraitGroup*	0.310	0.233	0.381	< .001
pTrait – Calculated_pTraitGroup	0.294	0.219	0.366	< .001
pTrait – pGroup	0.291	0.215	0.363	< .001
pTrait – pTraitGroup	0.251	0.174	0.325	< .001
pTraitGroup – Calculated_pTraitGroup	**0.240**	**0.162**	**0.314**	**< .001**

Table 2 (cont'd)

pGroupTrait – pGroup	0.205	0.126	0.281	< .001
pTraitGroup – pGroup	0.154	0.075	0.232	< .001

Note. The r_{jc} correlation is in bold. * indicates the correlation after restricting the calculated posterior values between 0 to 100. pTraitGroup represents the judged posterior (question 1 in Equation 1), pTrait represents the prior (question 2 in Equation 1), pGroupTrait and pGroup represent components of the likelihood ratio (questions 3, 4 in Equation 1), Calculated_pTraiGroup represents the calculated posterior computed using the prior and likelihood ratio values.

The second exploratory analysis computed a difference score between calculated and judged posterior values across the overall sample to examine how closely people's judged and calculated posteriors were aligned. For example, imagine a participant is presented the question, "Person 1 is a construction worker. Considering only this information, how likely is it that Person 1 is aggressive?" and the participant's response was 80 on a scale of zero to 100 (i.e., their judged posterior = 80). Further, imagine that their calculated posterior (computed using prior and likelihood ratio values) was 60. There is a 20-point difference between the participant's direct judgement and what Bayes' theorem states their judgement *should be*. Computing a difference score in this manner would allow for a more direct test of the level of discrepancy between participants' judged and calculated posteriors in this task. Indeed, Jussim (2012; 2015) argued that simply testing statistical significance (e.g., using correlations) provided no information about the level of discrepancy between people's judgements and a certain criterion. He stated that focusing mainly on statistical significance can be misleading and uninformative. For instance, if people were asked, "What percent of people identify as lesbian or gay in USA?" and a Gallup poll found roughly 1.7% identify as gay or lesbian (Gates, 2011; Gates & Newport, 2012), there would be a non-zero discrepancy if participants made any judgement other than

1.7%. Nuanced discrepancies cannot be captured when only looking at statistical significance. Thus, to further explore the discrepancy between participants' judged and calculated posteriors, I computed a difference score.

Table 3 presents the differences between judged and calculated posterior values. I categorized these differences as "less than 10 points", "between 11 to 20 points", or "more than 20 points". These categories were chosen arbitrarily, assuming a less than 10-point difference suggests highest alignment between judged and calculated posterior predictions, a difference between 11 to 20 points suggests moderate alignment, and more than a 20-point difference suggests low alignment between participants' judged and calculated posterior predictions. In Experiment 1, most observations had more than a 20-point difference, suggesting low alignment between participants' direct judgements and what Bayes' theorem states their judgements should be. This finding aligns with the small r_{JC} correlation in the main analysis, indicating weak evidence for the claim that participants made judgements as per Bayes' Theorem in this task.

Table 3

Total number of observations categorized by point differences between participants' judged and calculated posterior probabilities across the overall sample and after restricting the range of calculated posteriors (between 0-100) in Experiments 1, 2.

	Number of Observations		
	Below 10-point difference	Between 11- to 20- point difference	Above 20-point difference
Experiment 1	1915	1558	4474
Experiment 1*	1887	2505	3226
Experiment 2	2139	1417	3706
Experiment 2*	2040	2245	2002

Note. * indicates number of observations after restricting the calculated posterior values between 0 to 100.

The third exploratory analysis checked whether there was a smaller r_{JC} correlation between participants with relatively strong, consistent stereotype beliefs compared to participants with weak, inconsistent stereotype beliefs. Participants with strong stereotype beliefs would have low variability in responses to the component parts of Bayes' Theorem (i.e., their priors, likelihood ratio estimates), resulting in a smaller r_{JC} correlation. In comparison, those with weak stereotype beliefs (i.e., high variability in prior, likelihood ratio estimates) would have a relatively larger r_{JC} correlation. This was considered an alternative way to check if people's direct judgements were aligned with judgements as per Bayes' Theorem.

Strong vs. weak stereotypes were determined using Diagnostic Ratios or *DR*s above and below 1.0 respectively. Like the pretest, *DR*s were computed using the judged posterior (pTraitGroup) value and dividing it to the prior (pTrait) value for each participant. Most participants (N = 549) had *DR*s equal to or above 1.0, suggesting homogeneity in participants' stereotype beliefs within the current sample. Still, looking at the r_{JC} correlation separately for participants with *DR*s above versus below 1.0 suggested the r_{JC} correlation was smaller among the subgroup of participants with low variability in their stereotype beliefs (i.e., strong, consistent stereotypes) compared to those with high variability in their stereotype beliefs (i.e., weak, inconsistent stereotypes) (*DR*s above 1: r_{JC} = 0.24, $p < .001$, $t(547) = 5.89$, *95% CI* = [0.16, 0.32]; *DR*s below 1: r_{JC} = 0.44, $p = .004$, $t(39) = 3.09$, *95% CI* = [0.16, 0.66]). That is, participants appeared to update their judgements as per Bayes' Theorem given the strength of their stereotype beliefs (with relatively stronger stereotype beliefs having lower variability and thus resulting in a lower r_{JC} correlation). However, the large correlation among participants with weak, inconsistent stereotypes could be due to Type 1 error, suggesting caution in drawing firm conclusions from this analysis.

Discussion

Experiment 1 tested whether participants' predictions about specific individuals possessing certain traits generally followed Bayes' theorem given only social category information about the targets. In general, participants' impressions about individuals were less correlated with what Bayes' theorem suggested their responses should be. In other words, people's impressions about specific individuals possessing certain traits were weakly aligned with Bayes' theorem.

Exploratory analyses suggested first that none of the correlations between the other questions were large enough to impact the observed correlation in the main analysis. Second, there were large differences between participants' judgements and the Bayesian normative criteria adopted in this task. Third, there was low variability in stereotype beliefs in this sample; but evaluating the observed correlation among people with strong vs. weak stereotype beliefs could help test if people update their beliefs as per Bayes' Theorem in the future. Overall, exploratory analyses provided weak support for the hypothesis that participants' judgements about individuals generally aligned with Bayes' theorem.

Still, this experiment did not test whether participants can combine both, categorical and individuating information, while forming impressions of individuals. To explore this question and the notion that researchers need not starkly differentiate between categorical and individuating information, I conducted a second experiment.

EXPERIMENT 2

Here, I tested if people's predictions about individuals followed Bayes' theorem when two types of information (categorical, individuating) were presented together. As before, I hypothesized that participants' judged and calculated posterior predictions would largely follow Bayes' theorem.

Methods

I used the same method as in Experiment 1 but simultaneously presented two different types of information – categorical and individuating information – about each target in the same vignette. Participants saw information about 14 unique targets across 14 trials total. The targets belonged to one of 10 social categories and participants saw 14 unique behaviours that the targets performed. For example, Tom was from the category "construction worker" and performed a behaviour such as "yelling at someone". Participants' task was to assess how likely it was that Tom possessed a certain trait given this information (e.g., "How likely is it that Tom is aggressive?"). The social groups and traits were the same as in Experiment 1. Behaviours associated with each trait were derived from a pretest and are listed in Table A2 of Appendix A. Past work typically presented categorical vs. individuating information to test which information type was more influential when forming impressions (Hilton & Fein, 1989; Krueger & Rothbart, 1988; Locksley et al., 1982). This research suggested mixed evidence wherein people either committed the base-rate fallacy (they generally ignored categorical information when individuating information was available; Fiske & Neuberg, 1990; Kahneman & Tversky, 1982; Kunda & Thagard, 1996; Nisbett & Ross, 1981) or used categorical information by default because it was easier to process (e.g., Bruner, 1957; Fiske & Neuberg, 1990; Rumelhart & Ortony, 1977). However, the current experiments wanted to shift focus from this debate and

were more concerned with 1) whether people could combine and use both information types jointly and 2) if people's judgements about individuals possessing certain traits aligned with Bayes' theorem.

Under the Bayesian perspective, the impact of *both* types of information was accounted for by incorporating joint probabilities (see Equation 4.2), providing a novel test and reinterpretation of past impression formation studies. Specifically, I asked participants to report the probability of an event given two events occurring simultaneously. Generally, a joint probability such as this can be expressed as $p(A|B,C)$ or the probability of event A occurring given events B *and* C occurring together. In an impression formation experiment, this could be expressed as $p(Assertive|Behaviour, Sex)$. So, for example, the probability of a target being assertive (event A) can be assessed based on some behaviour they performed (event B) *and* their sex (event C). In this way, participants had to combine both information types while forming impressions about multiple targets in Experiment 2. This provided a formal test of whether participants' impressions about individuals followed Bayes' theorem given two types of information simultaneously.

Participants, Pre-registration, & Analysis Plan

For a within-subjects design and 90% power to detect a small effect (ρ = 0.15) I needed a sample of 462 participants (as per G*power 3.1)[4]. However, to have sufficient data after accounting for a ~15% possible data loss due to incomplete responding or poor data quality, I collected data from 540 participants from Michigan State University's human-subjects pool (Md_{age} = 19.82, M_{age} = 20.00, SD_{age} = 1.95; Female = 77.8%, Male = 20%, Non-binary/Other = 1.30%, Prefer not to say = 0.93%; White/European American = 65.2%, Asian/Asian American = 13.1%, Black/ African American = 6.11%, Hispanic/Latino = 6.11%, Middle Eastern = 2.96%,

American Indian/Alaska Native = 0.19%, Native Hawaiian/ Pacific Islander = 0.19%, Other = 3.89%, Prefer not to say = 2.22%). Data was analyzed using Pearson's product moment correlation. The correlation between judged and calculated posteriors was computed after aggregating responses across each participant.[5] This study was preregistered and all data, analysis scripts, manipulations, and measures were made available on the Open Science Framework (https://osf.io/t3bky/?view_only=5b50e774d4f047b893810f5d85c5fc1e).

Procedure

All participants were presented one piece of traditionally categorical information and one piece of traditionally individuating information about 14 unique targets in a within-subjects design. Presentation order for each piece of information was randomized and counterbalanced to account for any order effects. Like Experiment 1, participants' task was to provide ratings corresponding to different components of Bayes' theorem. An example item from the survey is given below wherein the categorical information is being female, the individuating information is crying often, and the trait is "emotional" (see Appendix C for full set of items):

1. Person 2 is female and Person 2 cries often. Considering only this information, how likely is it that Person 2 is emotional?[9]
2. What percent of all females do you think are emotional?[7]
3. What percent of all females who are emotional do you think cry often?
4. What percent of all females do you think cry often?

where question 1 measured the judged posterior prediction of a target being emotional given a conjoint probability of being female *and* crying often. Questions 2 to 4 measured the prior and likelihood ratios (where Q3 measured the probability of crying often given a conjoint

[9] Response scale for this question ranged from 0 to 100 where 0 = highly unlikely ("I'm certain this person is unemotional"), 50 = uncertain ("I'm uncertain about this person's level of emotionality"), and 100 = highly likely ("I'm certain this person is emotional").

probability of being female *and* being emotional) which were the component parts of Bayes' theorem (see Equations 4.1 and 4.2).[8]

I used responses from the component parts to calculate a posterior and compared this "calculated" posterior to the judged posterior (i.e., responses to Q1 above). Like Experiment 1, if the correlation between judged and calculated posterior were large, it would suggest that participants' predictions about individuals were highly aligned with what Bayes' theorem says their judgements should be. In an exploratory analysis, I separately correlated each of the other questions (#2, #3, and #4) with the judged posterior to check the magnitude of these correlations and whether they may have been large enough to impact the correlation between judged and calculated posterior probabilities. Participants responded to the above questions using a 0 to 100 scale (see footnotes 7, 9 for specific interpretations of the scale).

Results

The means and standard deviations for all components corresponding to Bayes' Theorem for the overall sample are presented in Table 4. Response distributions across all components of Bayes' Theorem are displayed in Figure 2. Calculated posterior values had a similarly large range (but smaller standard deviation) compared to Experiment 1, which may be due to absolute errors in participants' judgements in this task. Again, to account for extreme calculated posterior values, the correlation between judged and calculated posteriors is also reported after restricting the calculated posterior values to range between zero to 100.

Table 4
Means and standard deviations for all components of Bayes' Theorem in Experiment 2 using unaggregated (raw) data.

Questions	Means	SDs
Judged posterior (pTraitGroup)	67.372	21.620

Table 4 (cont'd)

Prior (pTrait)	54.253	21.362
Likelihood ratio numerator (pGroupTrait)	62.332	25.366
Likelihood ratio denominator (pGroup)	45.319	23.579
Calculated posterior (calc_pTraitGroup)	86.982	87.934
Calculated posterior (calc_pTraitGroup)*	62.528	23.677

Note. * indicates Means and SDs after restricting the calculated posterior values between 0 to 100.

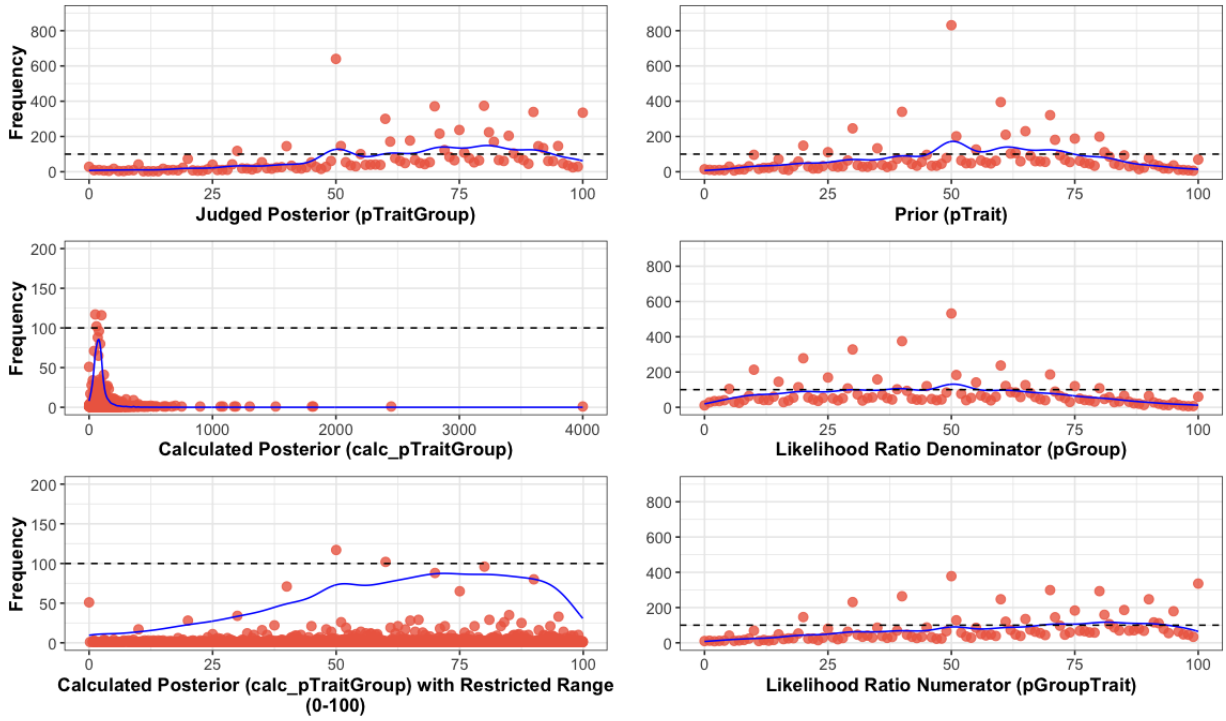

Figure 2
Scatter plots and kernel densities depicting frequencies of responses across all components of Bayes' Theorem in Experiment 2 using unaggregated (raw) data.

The main prediction again focused on the correlation between judged and calculated posteriors (r_{JC}) aggregated across each participant.[5] The correlation for the overall sample was large, $r_{JC} = 0.63$, $p < .001$, $t(538) = 18.72$, *95% CI* = [0.57, 0.68], suggesting people's judgements about specific individuals were highly aligned with what Bayes' theorem says their

judgements should be in this task. The correlation was larger than Experiment 1, implying people's judgements were more aligned with Bayes' Theorem when they had both, behavioural and social category information about a target rather than social category information only.

Exploratory Analyses

Again, I also looked at 1) correlations between the other questions (i.e., those corresponding to the priors, likelihood ratio), 2) how closely participants' judged and calculated posteriors were aligned using a difference score, and 3) if the magnitude of the r_{JC} correlation would be smaller for individuals with consistently held stereotypes (i.e., low variability across responses to the component parts of Bayes' Theorem) compared to individuals with inconsistent stereotypes (i.e., high variability across responses to the component parts of Bayes' Theorem).

First, I tested whether other questions included in the task substantially impacted the r_{JC} correlation reported in the main analysis (e.g., particularly large *r* values). Table 5 lists the various combinations of correlations for the questions corresponding to the priors and likelihood ratio. Most of these correlations were within the same range or slightly smaller than the r_{JC} correlation, suggesting no particular impact of these effects on the r_{JC} correlation.

Table 5
Overall correlations between all combinations of questions corresponding to the prior, likelihood ratio, judged posterior, and calculated posterior probabilities in Experiment 1.

Questions	*r*	95% Lower CI	95% Upper CI	*p* value
pTrait – pGroup	0.745	0.705	0.780	< .001
pGroupTrait – Calculated_pTraitGroup	0.701	0.656	0.742	< .001
pTraitGroup – Calculated_pTraitGroup	**0.628**	**0.574**	**0.677**	**< .001**

Table 5 (cont'd)

pTraitGroup – pGroupTrait	0.627	0.562	0.667	< .001
pTrait – pTraitGroup	0.625	0.571	0.674	< .001
pTraitGroup – Calculated_pTraitGroup*	0.619	0.562	0.671	< .001
pTrait – pGroupTrait	0.519	0.454	0.578	< .001
pGroupTrait – pGroup	0.473	0.405	0.536	< .001
pTrait – Calculated_pTraitGroup	0.425	0.354	0.492	< .001
pTraitGroup – pGroup	0.342	0.266	0.415	< .001
Calculated_pTraitGroup – pGroup	-0.080	-0.163	-0.004	0.063

Note. The r_{jc} correlation is in bold. * indicates the correlation after restricting the calculated posterior values between 0 to 100. pTraitGroup represents the judged posterior (question 1 in Equation 1), pTrait represents the prior (question 2 in Equation 1), pGroupTrait and pGroup represent components of the likelihood ratio (questions 3, 4 in Equation 1), Calculated_pTraiGroup represents the calculated posterior computed using the prior and likelihood ratio values.

The second exploratory analysis computed a difference score between calculated and judged posterior values across the overall sample to examine the level of alignment between people's judged and calculated posteriors more directly (see Table 3 above). Again, I arbitrarily categorized these differences as "less than 10 points", "between 11 to 20 points", or "more than 20 points", assuming a less than 10-point difference suggests highest alignment between judged and calculated posterior predictions, a difference between 11 to 20 points suggests moderate alignment, and more than a 20-point difference suggests low alignment between participants' judged and calculated posterior predictions. Like Experiment 1, most observations had more than a 20-point difference, suggesting low alignment between participants' direct judgements and what Bayes' theorem states their judgements should be (although there were generally more

observations with a less than 10-point difference than in Experiment 1). This finding challenges the large r_{JC} correlation observed in the main analysis, cautioning readers from forming solid conclusions about the level of alignment between participants' judgements and judgements prescribed by Bayes' Theorem in this task.

The third exploratory analysis checked whether there was a smaller r_{JC} correlation between participants with relatively strong, consistent stereotype beliefs (low variability in responses) compared to participants with weak, inconsistent stereotype beliefs (high variability in responses). This was an alternative way to check if people's direct judgements were aligned with judgements as per Bayes' Theorem. Again, strong vs. weak stereotypes were determined using Diagnostic Ratios (*DR*s) above and below 1.0 respectively. Like Experiment 1, most participants (N = 505) had *DR*s equal to or above 1.0, suggesting homogeneity in participants' stereotype beliefs within the current sample. Still, looking at the r_{JC} correlation separately for participants with *DR*s above versus below 1.0 suggested the r_{JC} correlation was slightly smaller among participants with low variability in their stereotype beliefs (i.e., strong, consistent stereotypes) than those with high variability in their stereotype beliefs (i.e., weak, inconsistent stereotypes) (*DR*s above 1: r_{JC} = 0.62, p < .001, t(503) = 17.56, *95% CI* = [0.56, 0.67]; *DR*s below 1: r_{JC} = 0.65, p < .001, t(33) = 4.86, *95% CI* = [0.40, 0.81]). However, again the large correlation among participants with high variability in stereotype beliefs could be due to Type 1 error, suggesting caution in drawing firm conclusions from this analysis.

Discussion

Experiment 2 tested whether participants' predictions about specific individuals possessing certain traits would generally follow Bayes' theorem when given both, categorical and individuating information. Unlike Experiment 1, I found that participants' impressions about

individuals were highly correlated with responses as per Bayes' theorem. That is, people's impressions about specific individuals possessing certain traits were largely aligned with Bayes' theorem given both, categorical and individuating information about a target. It is possible that having more information (i.e., two pieces of information: social category and behavioural) in this experiment led to stronger predictions about a target possessing a certain trait than in Experiment 1 (which had only one piece of social category information). Another possibility is that behavioural information was perceived as more diagnostic than social category information during impression formation in general. Assessing how people update their beliefs by iteratively presenting information of varying strength could potentially help determine the extent to which these possibilities impact information use during impression formation. I breakdown this idea further in the General Discussion section below. However, overall, it appears people were able to combine and use both information types in Experiment 2, providing preliminary support for the claim that starkly differentiating between categorical and individuating information might be unnecessary.

Like Experiment 1, exploratory analyses suggested first, none of the correlations between the other questions were large enough to impact the correlation between judged and calculated posterior probabilities. Second, most responses indicated large differences between participants' judgements and the Bayesian normative criteria adopted in this task, although there were generally fewer discrepancies than in Experiment 1. Third, there was low variability in stereotype beliefs in this sample; in the future, evaluating the observed correlation among people with strong vs. weak stereotype beliefs could help test if people update their beliefs as per Bayes' Theorem. Overall, exploratory analyses again suggested weak support for the claim that participants' judgements about individuals aligned with Bayes' theorem in this task.

This experiment presented categorical and individuating information simultaneously, which has its limitations. For instance, this design precluded an iterative test of whether participants updated their posterior predictions given different types of information. I discuss potential explanations for the present findings, limitations, and future directions below.

GENERAL DISCUSSION

The present experiments adopted a Bayesian perspective to test whether people could combine two types of information (individuating, categorical) and jointly use them when forming impressions of others. Specifically, these experiments focused on testing whether participants' impressions of specific individuals aligned with Bayes' theorem given social category information only vs. social category plus behavioural information about a target. Observed results from the main experiments suggested participants' judgements about individuals possessing certain traits were generally aligned with Bayes' theorem. This alignment was weak when given social category information alone but strong when given social category plus behavioural information (Experiment 1: Pearson's $r = 0.24$; Experiment 2 Pearson's $r = 0.63$). That is, people appeared to be able to combine and use different types of information in this task, suggesting researchers can consider bypassing attempts to starkly distinguish information types during impression formation.

Exploratory analyses looked at three things: 1) the correlations between all other questions in this task, 2) difference scores between participants' judgements and judgements as per Bayes' Theorem, and 3) the magnitude of correlations among participants with strong vs. weak stereotype beliefs. First, no correlations appeared large enough to impact the observed correlation in the main analysis. Second, there were generally large differences in participants' judgements and judgements prescribed by Bayes' Theorem. This suggests caution in forming firm conclusions or generalizations based on the observed correlation in the main experiments. This also highlights the importance of exploring discrepancies between participants' direct judgements and judgements as per Bayesian normative criteria outside of correlation coefficients. Third, participants' stereotype beliefs were mostly homogenous, providing an

uncompelling test of whether correlations differed among people with strong vs. weak stereotype beliefs. However, probing whether people with strong stereotypes (i.e., low variability in responses) would have lower observed correlations than people with weak stereotypes (i.e., high variability in responses) presents a potential avenue for testing Bayesian updating during impression formation.

Together, this work suggests participants' judgements about *specific individuals* possessing certain traits were weakly aligned with Bayes' Theorem. Past work on *social groups* found a larger correlation between people's direct judgements and judgements prescribed by Bayes' Theorem (McCauley & Stitt, 1978; Solanki & Cesario, 2024; Pearson's *r* ranged between 0.79 to 0.89). One explanation for a smaller correlation for judgements about individuals versus social groups could be related to procedural differences in how people generally perceive individuals versus social groups. For instance, perceivers might expect different degrees of unity and coherence among groups as opposed to within an individual's personality, resulting in different impressions about groups vs. individuals (Hamilton & Sherman, 1996). People's impressions may also vary based on the specific traits being assessed. For example, impressions of individuals on warmth- and competence-related trait dimensions correlate positively (a halo effect) whereas impressions about social groups on these same trait dimensions correlate negatively (e.g., Asians are judged to have high competence and low warmth; Fiske et al., 2007).

Other explanations for the current results could be related to various methodological considerations. First, there was an important methodological difference between the current experiments and past work (McCauley & Stitt, 1978; Solanki & Cesario, 2024). Past work asked participants to rate the degree to which a certain trait was considered stereotypic or non-stereotypic of a given social group and did this across multiple traits within that group. That is,

participants had to rank the likelihood of specific traits being associated with certain social categories, focusing only on judgements about a single category (e.g., What percent of men are assertive? What percent of men are emotional?). In contrast, the current experiments had participants making judgements *across* multiple individuals, thereby comparing evaluations across a broader spectrum (e.g., How likely is it that Person 1 is assertive? How likely is it that Person 2 is emotional?). This methodological difference likely introduced additional variability in the current data, potentially reducing the strength of the observed correlations for specific individuals compared to social groups.

Second, people were always presented strong categorical and strong individuating information about a target in the current experiments; but manipulating information strength could provide an alternative test for how people assess likelihoods and update their posteriors in different contexts. Third, and relatedly, an artifact of presenting two pieces of information simultaneously in the same vignette is that it did not allow for an iterative test of 1) how either information type might constitute a prior, 2) how people might update their prior beliefs given different new information, and 3) whether people can combine different information types and update their posteriors accordingly. While individuals in real-life scenarios may frequently receive information from diverse sources, iterative testing remains crucial within this context to effectively address the question of whether both, categorical or individuating information, could be regarded as prior or new information in a Bayesian framework. Hence, it would be worthwhile to test whether people update their predictions rationally (as per Bayes' theorem) when iteratively given different prior information (e.g., strong vs. weak) and manipulating the information presented as the prior vs. new information. Such a manipulation would provide clearer support for the argument that people's impressions generally follow Bayes' theorem and

that impression formation researchers can sidestep attempts to starkly differentiate information types under a Bayesian framework.

A fourth methodological consideration and potential explanation for the current findings concerns the criteria used to assess discrepancy between participants' direct judgements and judgements as per Bayes' Theorem . Here, discrepancy was tested using a simple difference score to test the degree of alignment between people's judgements and adopted empirical criteria (i.e., Bayesian calculated posteriors). However, other criteria can be considered. For example, when forming an impression of a target's intelligence, an informative criterion to assess discrepancy could be a measure of intelligence obtained using a standardized test like Wechsler Adult Intelligence Scale. A combination of such relevant criteria may provide further empirical support in forming conclusions about judgement discrepancy during impression formation (see Jussim, 2012).

A fifth explanation and potential limitation of the current findings is related to the sample and specific items included. These experiments included a primarily undergraduate sample, which is known to be liberal and might have skewed results due to socially desirable responding (Hanel & Vione, 2016; Henry, 2008). Relatedly, the stereotypes and traits used in the current experiments were pretested on undergraduate samples and thus represent only a subset of population's views on the associations between these stereotypes and traits. Some traits might be judged as relatively more diagnostic of a given social group or a given individual's personality in the general public than among younger people. For example, there might be relatively more negative perceptions about gay people in the general population (Smith, 2011). Future work should test the same questions and pretest items in a more representative sample. This can help

identify any systematic biases in how people combine and use different types of information during impression formation.

To firmly establish that people are following Bayes' Theorem when forming impressions of others, future work can explore 1) information types (strong vs. weak), 2) iterative tests of how people update their judgements given different prior information, 3) different criterion measures, 4) representative samples and items, among others. Considering these factors, some future directions are discussed below.

Future directions

Testing Bayesian Updating Iteratively Using Different Information Types

Regarding the second and third points noted above (that only strong information was presented and both information types were presented simultaneously in the same vignette), an alternative method to study the current question would be to take a piecemeal approach of presenting different information as the prior.

Imagine if participants were presented two pieces of information sequentially and asked to make two judgements (one after each piece of information was presented) about the same target. When the first piece of information was strong categorical information, the second piece of information would always be weak individuating information; conversely, when the first piece of information was weak individuating information, the second piece would always be strong categorical information. This experiment can have two independent variables with two levels each (Information Type: Strong categorical vs. Weak individuating) × Time of Judgement (Time 1 vs. Time 2) in a within-subjects design. Note that all participants would see either strong categorical information or weak individuating information in place of the prior. This experiment can be done in two parts. For example, in Part I of the experiment, the strong categorical

information can be about a target being Asian, the weak individuating information can be that the target likes brunch, and the trait to be judged would be "smart". The questions would be as follows:

[Time$_1$] Person 3 is Asian. Considering only this information, how likely is it that Person 3 is smart?
[Time$_2$] Person 3 likes to eat brunch. Now, how likely is it that Person 3 is smart?

Or these questions can be presented in the reverse order, as follows:

[Time$_1$] Person 3 likes to eat brunch. Considering only this information, how likely is it that Person 3 is smart?
[Time$_2$] Person 3 is Asian. Now, how likely is it that Person 3 is smart?

In Part II, participants would always see the strong individuating information paired with weak categorical information. For example, the strong individuating information can be that the target got an A in math in undergrad and the weak categorical information can be about the target being European. The questions would be as follows:

[Time$_1$] Person 4 is European. Considering only this information, how likely is it that Person 4 is smart?
[Time$_2$] Person 4 got an A in math in undergrad. Now, how likely is it that Person 4 is smart?

Or these questions can be presented in the reverse order, as follows:

[Time$_1$] Person 4 got an A in math in undergrad. Considering only this information, how likely is it that Person 4 is smart?
[Time$_2$] Person 4 is European. Now, how likely is it that Person 4 is smart?

The effect of interest in such a design would be the interaction between Information Type and Time of Judgement. Specifically, one could have three separate predictions. First, average responses would be higher in trials where the prior was based on strong information compared to when the prior was based on weak information (see t1_rating in Figure F1, Appendix F). Second, because participants would have the same total aggregate information at Time 2, average difference in responses between trials at Time 2 should be smaller than average difference in responses between trials at Time 1 (see difference in t2_ratings vs. t1_rating in Figure F1,

Appendix F). Third, average difference in responses from Time 1 to Time 2 should be smaller for trials where strong information was followed by weak information rather than when weak information was followed by strong information (see red vs. blue box plots in Figure F1, Appendix F).

Bayes' Theorem does not require making an explicit distinction between categorical and individuating information, but rather considers both to have the same potential to be considered either a prior or new information. Thus, using the above design one could test 1) the degree of difference in participants' overall prior predictions (i.e., Time 1 ratings) given strong versus weak *base rate* information, 2) the degree of difference in participants' posterior judgements (i.e., Time 2 ratings) given strong versus weak *new* information, and 3) the degree of difference across participants' Time 1 ratings (based on strong vs. weak *base rate* information) versus Time 2 ratings (based on strong vs. weak *new* information). Mixed effects modeling using maximum likelihood estimation can be used to predict participants' responses (measured on a 0 to 100 scale) as a function of Information Type and Time of Judgement. Conceptually, this experiment can present relatively stronger evidence for the claim that *any* information type can be considered a prior vs. new information and people can combine different types of information when forming impressions.

Normative Criteria Beyond Bayesian Calculated Posteriors

Jussim (2012) has noted various criteria against which the discrepancy (or accuracy) of social perceptions can be tested. Chief among these are agreement with other perceivers (e.g., experts, theoretical models, independent judges, non-independent judges), agreement with the target (e.g., self-reports, self-perceptions), and hybrid criteria which are a combination of the

above.[10] These criteria can be informative in assessing accuracy during impression formation when accuracy is conceptualized as the degree of alignment between people's judgements and certain normative criteria. For instance, self-perceptions can be used as criteria to assess accuracy in perceivers' beliefs. The following example makes this clear:

> "Bertha may think she is a great athlete and Nyesha may think she is a good athlete. Both may be overestimates (Bertha may only be pretty good and Nyesha may be pretty average). But if their degree of self-inflation is similar, it may be true that Bertha is more athletic than Nyesha. So, a coach who views Bertha as more athletic than Nyesha would be correct (and the coach's views would correlate well with Bertha's and Nyesha's self-perceptions; Jussim, 2012, p. 189)."

Thus, Bayesian calculated posterior are only one among many criteria which can be used to test whether people rationally update their beliefs during impression formation. Understanding and using different criteria to supplement a Bayesian calculated posterior can arguably provide a clearer picture against which to assess accuracy (or lack thereof) during impression formation.

[10] Although each of the criteria come with their own limitations (see Jussim, 2012). For example, self-perceptions might be warped and lead to socially desirable responding (Paulhus, 1991; 1998).

CONCLUSION

The current work tested if people's judgements were aligned with Bayes' Theorem when given different types of information (categorical, individuating) during impression formation. Across two experiments, people's judgements about specific individuals were generally aligned with judgements prescribed by Bayes' Theorem, although exploratory analysis using difference scores suggested this alignment was weak. Overall, it appears people were able to combine and use different information types during impression formation, providing initial support for the claim that researchers need not starkly differentiate between categorical vs. individuating information when studying impression formation from a Bayesian lens.

REFERENCES

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918. https://doi.org/10.1037/0033-2909.84.5.888

Brewer, M. B. (1988). A dual process model of impression formation. In R. S. Wyer & T. K. Srull (Eds.), *Advances in social cognition* (Vol. 1). Lawrence Erlbaum Associates.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *64*(2), 123–152. https://doi.org/10.1037/h0043805

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., & 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 M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 1–74). Academic Press. https://doi.org/10.1016/S0065-2601(08)60317-2

Gates, G. J. (2011). How many people are lesbian, gay, bisexual, and transgender? JSTOR.

Gates, G. J., & Newport, F. (2012). Special report: 3.4% of us adults identify as LGBT. Washington, DC: Gallup.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*, 733–740. https://doi.org/10.1037/0022-3514.54.5.733

Ginosar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgments about another person. *Journal of Experimental Social Psychology*, *16*(3), 228–242. https://doi.org/10.1016/0022-1031(80)90066-9

Ginossar, Z., & Trope, Y. (1987). Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology*, *52*(3), 464–474. https://doi.org/10.1037/0022-3514.52.3.464

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review*, *103*(2), 336–355. https://doi.org/10.1037/0033-295X.103.2.336

Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). The Multiple Dimensions of Gender Stereotypes: A Current Look at Men's and Women's Characterizations of Others and Themselves. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00011

Hilton, J. L., & Fein, S. (1989a). The role of typical diagnosticity in stereotype-based judgments. *Journal of Personality and Social Psychology*, *57*(2), 201–211. https://doi.org/10.1037/0022-3514.57.2.201

Hilton, J. L., & Fein, S. (1989b). The role of typical diagnosticity in stereotype-based judgments. *Journal of Personality and Social Psychology*, *57*(2), 201–211. https://doi.org/10.1037/0022-3514.57.2.201

Hinsz, V. B., Tindale, R. S., Nagao, D. H., Davis, J. H., & Robertson, B. A. (1988). The influence of the accuracy of individuating information on the use of base rate information in probability judgment. *Journal of Experimental Social Psychology*, *24*(2), 127–145. https://doi.org/10.1016/0022-1031(88)90017-0

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–1187. https://doi.org/10.1037/a0020463

Jussim, L., Crawford, J. T., Anglin, S., Chambers, J., Stevens, S. T., & Cohen, F. (2015). Stereotype Accuracy: One of the largest and most replicable effects in all of Social Psychology. In *Handbook of Prejudice, Stereotyping, and Discrimination* (2nd ed., pp. 31–63). Psychology Press.

Jussim, L. J. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy.* (1st ed.). Oxford University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Krueger, J., & Rothbart, M. (1988). Use of categorical and individuating information in making inferences about personality. *Journal of Personality and Social Psychology*, *55*(2), 187–195. https://doi.org/10.1037/0022-3514.55.2.187

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the Construal of Individuating Information. *Personality and Social Psychology Bulletin*, *19*(1), 90–99. https://doi.org/10.1177/0146167293191010

Kutzner, F., & Fiedler, K. (2017). Stereotypes as Pseudocontingencies. *European Review of Social Psychology*, *28*(1), 1–49. https://doi.org/10.1080/10463283.2016.1260238

Locksley, A., Hepburn, C., & Ortiz, V. (1982a). Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology*, *18*(1), 23–42. https://doi.org/10.1016/0022-1031(82)90079-8

Locksley, A., Hepburn, C., & Ortiz, V. (1982b). Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology*, *18*(1), 23–42. https://doi.org/10.1016/0022-1031(82)90079-8

McCauley, C. R. (1994). Stereotypes as Base Rate Predictions: Commentary on Koehler on Base-Rate. *Psycoloquy*, *5*.

McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology*, *36*(9), 929–940. https://doi.org/10.1037/0022-3514.36.9.929

Nelson, T. E., Biernat, M. R., & Manis, M. (1990). Everyday base rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology*, *59*(4), 664–675. https://doi.org/10.1037/0022-3514.59.4.664

Paulhus, D. L. (1991). Measures of personality and social psychological attitudes. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measurement and control of response bias* (pp. 17–59). Academic Press.

Paulhus, D. L. (1998). Interpersonal and intrapsychic adaptiveness of trait self-enhancement: A mixed blessing? *Journal of Personality and Social Psychology*, *74*(5), 1197–1208. https://doi.org/10.1037/0022-3514.74.5.1197

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 544–554. https://doi.org/10.1037/a0034887

Rasinski, K. A., Crocker, J., & Hastie, R. (1985). Another look at sex stereotypes and social judgments: An analysis of the social perceiver's use of subjective probabilities. *Journal of Personality and Social Psychology*, *49*(2), 317–326. https://doi.org/10.1037/0022-3514.49.2.317

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Erlbaum.

Smith, T. W. (2011). *Public Attitudes toward Homosexuality* (pp. 1–4) [General Social Survey]. NORC, University of Chicago.

Solanki, P. & Cesario, J. (2024). Stereotypes as Bayesian Judgements of Social Groups. *The Journal of Social Psychology*. http://dx.doi.org/10.1080/00224545.2024.2368017.

Stangor, C. (2016). The study of stereotyping, prejudice, and discrimination within social psychology: A quick history of theory and research. In *Handbook of prejudice, stereotyping, and discrimination, 2nd ed* (pp. 3–27). Psychology Press.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*, 239–257. https://doi.org/10.1037/0033-295X.93.3.239

APPENDIX A: PRETEST MATERIALS & RESULTS

Table A1

List of social categories and associated traits used in Experiments 1 and 2.

Study	Group	Trait
Krueger & Rothbart, 1988	construction worker	aggressive
Locksley et al., 1982	morning person	healthy, dependable
Hilton & Fein, 1989	male	assertive
Kunda & Sherman-Williams, 1993	car salesperson	extroverted
	male	unemotional, dominant
	female	emotional
	gay people	flamboyant
Solanki & Cesario, 2024	lesbians	activists
	Asian	smart
	politician	greedy
	lawyer	rich

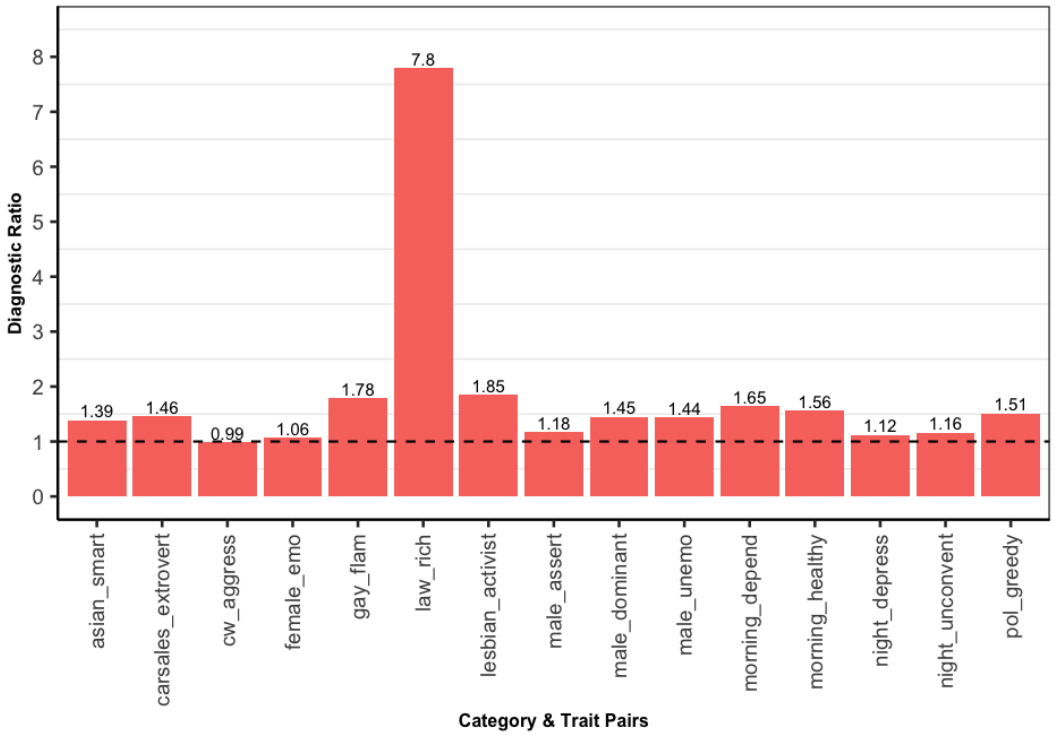

Figure A1

Diagnostic ratios for all trait and social group pairings in pretest Study 1 (also see Table A1 above). The dotted line indicates DR = 1.

Table A2

Diagnostic ratios for all traits and their associated behaviours in pretest Study 2.

Traits	Behaviours	Diagnostic Ratios
Aggressive.1	Yell at others	2.2
Aggressive.2	Scream during a disagreement	2.44
Aggressive.3	Hit someone else	2.02
Aggressive.4	Get into physical fights	1.89
Healthy.1	Exercise regularly	1.52
Healthy.2	Be physically active	1.88
Healthy.3	Eat nutritious food	1.80
Healthy.4	Do not eat fast food	1.84
Dependable.1	Always meet deadlines	3.03
Dependable.2	Always on time	2.32
Dependable.3	Always help someone else	2.34
Dependable.4	Always keep promises	2.25
Unconventional.1	Did not finish high school	2.83
Unconventional.2	Got driver's license after 30 years of age	1.91
Unconventional.3	Eat steak for breakfast	2.22
Unconventional.4	Use a unicycle as form of transportation	2.18
Depressed.1	Stay in bed all day	1.37
Depressed.2	Isolate themselves from family and friends	2.23
Depressed.3	Constantly sad	2.03
Depressed.4	Very moody	1.84
Assertive.1	Express their opinions freely	1.89
Assertive.2	Talk to customer care representatives when having a problem	1.68
Assertive.3	Take initiative on a project	1.46
Assertive.4	Able to say no to authority figures at work	1.43
Dominant.1	Take lead on a group project	1.48
Dominant.2	Take charge of planning a group vacation	1.94
Dominant.3	Try to control others	1.62
Dominant.4	Talk over someone in a group discussion	1.69
Unemotional.1	Have no reaction to a traumatic event	4.38
Unemotional.2	Do not openly express emotions	3.96
Unemotional.3	Lack empathy	1.61
Unemotional.4	Do not care about others' feelings	2.48
Emotional.1	Cry easily	1.10
Emotional.2	Cry often	1.06
Emotional.3	Openly express their feelings	1.34
Emotional.4	Talk a lot about their feelings	1.34
Flamboyant.1	Wear bright-colored clothes	2.07
Flamboyant.2	Have a unique dressing sense	1.84
Flamboyant.3	Speak loudly in public	2.22
Flamboyant.4	Sing in public	1.99

Table A2 (cont'd)

Activist.1	Publicly protest	1.75
Activist.2	Fight for their rights	2.33
Activist.3	Sign petitions for social change	4.62
Activist.4	Raise awareness about issues by posting on social media	4.69
Smart.1	Correctly apply calculus in real-world problems	1.64
Smart.2	Have a large vocabulary	1.40
Smart.3	Get good grades in school	1.55
Smart.4	Tutor others in academics	1.74
Greedy.1	Find it difficult to share	1.82
Greedy.2	Hoarding things	1.67
Greedy.3	Only care about oneself	0.85
Greedy.4	Manipulate others to get what one wants	1.12
Rich.1	Own a mansion	9.26
Rich.2	Spend money on designer brands	12.12
Rich.3	Invest most of one's income	8.47
Rich.4	Donate a lot of money to charity	13.27
Extrovert.1	Talking to strangers easily	1.44
Extrovert.2	Comfortable speaking in front of an audience	1.40
Extrovert.3	Make small talk easily	1.39
Extrovert.4	Start conversations at parties	1.44

Note. For each trait, four behaviours were included in the pretest but only the behaviour with the highest *DR* was chosen for Experiment 2. In cases where multiple behaviours had the same *DR* value (e.g., extrovert), the first behaviour in the list was picked.

APPENDIX B: EXPERIMENT 1 MATERIALS

Item developed based on Krueger and Rothbart's (1988) materials:

1. Category: construction worker, trait: aggressive
    - Person 1 is a construction worker. Considering only this information, how likely is it that Person 1 is aggressive?
    - What percent of all the world's people do you think are aggressive?
    - What percent of all aggressive people do you think are construction workers?
    - What percent of all the world's people do you think are construction workers?

Items developed based on Locksley et al.'s (1982) materials:

2. Category: Morning person, trait: healthy
    - Person 2 is a morning person. Considering only this information, how likely is it that Person 2 is healthy?
    - What percent of all the world's people do you think are healthy?
    - What percent of all healthy people do you think are morning people?
    - What percent of all the world's people do you think are morning people?

Item developed based on Hilton and Fein's (1989) materials:

3. Category: male, trait: assertive
    - Person 3 is a male. Considering only this information, how likely is it that Person 3 is assertive?
    - What percent of all the world's people do you think are assertive?
    - What percent of assertive people do you think are male?
    - What percent of all the world's people do you think are male?

Item based on Kunda & Sherman-Williams (1993):

4. Category: car salesman, trait: extroverted
    - Person 4 is a car salesman. Considering only this information, how likely is it that Person 4 is extroverted?
    - What percent of all the world's people do you think are extroverted?
    - What percent of all extroverted people do you think are car salesmen?
    - What percent of all the world's people do you think are car salesmen?

Items based on Solanki & Cesario (*unpublished manuscript*):

5. Category: male, trait: unemotional
    - Person 5 is a male. Considering only this information, how likely is it that Person 5 is unemotional?
    - What percent of all the world's people do you think are unemotional?
    - What percent of all unemotional people do you think are male?
    - What percent of all the world's people do you think are male?

6. Category: male, trait: dominant
    - Person 6 is a male. Considering only this information, how likely is it that Person 6 is unemotional?

- What percent of all the world's people do you think are dominant?
- What percent of all dominant people do you think are male?
- What percent of all the world's people do you think are male?

7. Category: female, trait: emotional
    - Person 7 is a female. Considering only this information, how likely is it that Person 7 is emotional?
    - What percent of all the world's people do you think are emotional?
    - What percent of all emotional people do you think are female?
    - What percent of all the world's people do you think are female?
8. Category: gay, trait: flamboyant
    - Person 8 is gay. Considering only this information, how likely is it that Person 8 is flamboyant?
    - What percent of all the world's people do you think are flamboyant?
    - What percent of all flamboyant people do you think are gay?
    - What percent of all the world's people do you think are gay?
9. Category: lesbian, trait: being an activist
    - Person 9 is lesbian. Considering only this information, how likely is it that Person 9 is an activist?
    - What percent of all the world's people do you think are activists?
    - What percent of all people who are activists do you think are lesbian?
    - What percent of all the world's people do you think are lesbian?
10. Category: Asian, trait: smart
    - Person 10 is Asian. Considering only this information, how likely is it that Person 10 is smart?
    - What percent of all the world's people do you think are smart?
    - What percent of all smart people do you think are Asian?
    - What percent of all the world's people do you think are Asian?
11. Category: politician, trait: greedy
    - Person 11 is a politician. Considering only this information, how likely is it that Person 11 is greedy?
    - What percent of all the world's people do you think are greedy?
    - What percent of all greedy people do you think are politicians?
    - What percent of all the world's people do you think are politicians?
12. Category: lawyer, trait: rich
    - Person 12 is a lawyer. Considering only this information, how likely is it that Person 12 is rich?
    - What percent of all the world's people do you think are rich?
    - What percent of all rich people do you think are lawyers?
    - What percent of all the world's people do you think are lawyers?

APPENDIX C: EXPERIMENT 2 MATERIALS

Note that all behaviours in the below examples are randomly chosen for the purpose of providing sample items. The behaviours for the main experiments will be based on pretest results. All traits and all behaviours will be chosen based on whether they have diagnostic ratios above 1.0.

1. Category: construction worker, behaviour: hitting someone, trait: aggressive
    - Person 1 is a construction worker and Person 1 hit someone who annoyed them. Considering only this information, how likely is it that Person 1 is aggressive?
    - What percent of all construction workers do you think are aggressive?
    - What percent of all construction workers who are aggressive do you think hit someone who annoyed them?
    - What percent of all construction workers do you think hit someone who annoyed them?
2. Category: Morning person, behaviour: cycling to work, trait: healthy
    - Person 2 is a morning person and Person 2 cycles to work. Considering only this information, how likely is it that Person 2 is healthy?
    - What percent of all morning people do you think are healthy?
    - What percent of all morning people who are healthy do you think cycle to work?
    - What percent of all morning people do you think cycle to work?
3. Category: male, behaviour: interrupting a group conversation to meet someone important, trait: assertive
    - Person 3 is a male and Person 3 interrupted a group conversation to meet someone important. Considering only this information, how likely is it that Person 3 is assertive?
    - What percent of all males do you think are assertive?
    - What percent of all males who are assertive do you think interrupted a group conversation to meet someone important?
    - What percent of all males do you think interrupted a group conversation to meet someone important?
4. Category: car salesman, behaviour: liking office parties, trait: extroverted
    - Person 4 is a car salesman and Person 4 likes to attend office parties. Considering only this information, how likely is it that Person 4 is extroverted?
    - What percent of all car salesmen do you think are extroverted?
    - What percent of all car salesmen who are extroverted do you think like to attend office parties?
    - What percent of all car salesmen do you think like to attend office parties?
5. Category: male, behaviour: not getting upset after serious arguments, trait: unemotional
    - Person 5 is a male and Person 5 does not get upset after serious arguments. Considering only this information, how likely is it that Person 5 is unemotional?
    - What percent of all males do you think are unemotional?

- What percent of all males who are unemotional do you think don't get upset after serious arguments?
- What percent of all males do you think don't get upset after serious arguments?

6. Category: male, behaviour: not letting anyone enter their lane while driving, trait: dominant
    - Person 6 is a male and Person 6 does not let anyone enter their lane while driving. Considering only this information, how likely is it that Person 6 is dominant?
    - What percent of all males do you think are dominant?
    - What percent of all males who are dominant do you think don't let anyone enter their lane while driving?
    - What percent of all males do you think don't let anyone enter their lane while driving?
7. Category: female, behaviour: crying often, trait: emotional
    - Person 7 is a female and Person 7 cries often. Considering only this information, how likely is it that Person 7 is emotional?
    - What percent of all females do you think are emotional?
    - What percent of all females who are emotional do you think are cry often?
    - What percent of all females do you think are cry often?
8. Category: gay, behaviour: wearing colourful clothes, trait: flamboyant
    - Person 8 is gay and Person 8 wears colourful clothes. Considering only this information, how likely is it that Person 8 is flamboyant?
    - What percent of all gay people do you think are flamboyant?
    - What percent of all gay people who are flamboyant do you think wear colourful clothes?
    - What percent of all gay people do you think wear colourful clothes?
9. Category: lesbian, behaviour: actively participating in an NGO, trait: being an activist
    - Person 9 is a lesbian and Person 9 actively participates in an NGO. Considering only this information, how likely is it that Person 9 is an activist?
    - What percent of all lesbian people do you think are activists?
    - What percent of all lesbian people who are activists do you think actively participate in an NGO?
    - What percent of all lesbian people do you think actively participate in an NGO?
10. Category: Asian, behaviour: getting an A in math in undergrad, trait: smart
    - Person 10 is Asian and Person 10 got an A in math in undergrad. Considering only this information, how likely is it that Person 10 is smart?
    - What percent of all Asians do you think are smart?
    - What percent of all Asians who are smart do you think got an A in math in undergrad?
    - What percent of all Asians do you think got an A in math in undergrad?
11. Category: politician, behaviour: discouraging donating money, trait: greedy
    - Person 11 is a politician and Person 11 discourages donating money. Considering only this information, how likely is it that Person 11 is greedy?
    - What percent of all politicians do you think are greedy?

- What percent of all politicians who are greedy do you think discourage donating money?
- What percent of all politicians do you think are discourage donating money?

12. Category: lawyer, behaviour: owning a summer home, trait: rich
    - Person 12 is a lawyer and Person 12 owns a summer home. Considering only this information, how likely is it that Person 12 is rich?
    - What percent of all lawyers do you think are rich?
    - What percent of all lawyers who are rich do you think own a summer home?
    - What percent of all lawyers do you think own a summer home?

APPENDIX D: EXPERIMENT 1 SUPPLEMENTARY ANALYSES

The correlation using unaggregated data was smaller than the observed correlation reported in the main analysis (which aggregated across participants), $r_{JC} = 0.15$, $p < .001$, $t(8258) = 14.07$, *95% CI* = [0.17, 0.13]. This suggests weak alignment with Bayes' theorem for judgements about specific individuals overall. When exploring the r_{JC} correlation, I also checked correlations between all other questions in this task (i.e., questions corresponding to the priors, likelihood ratio) and to test whether any other questions in the task substantially impacted the r_{JC} correlation reported in the main analysis (e.g., particularly large *r* values). Table D1 lists the various combinations of correlations for the questions corresponding to the priors, likelihood ratio in unaggregated data. Most of these correlations were within the same range or smaller than the r_{JC} correlation, suggesting no particular impact of these effects on the r_{JC} correlation.[11]

Table D1

Overall correlations between all combinations of questions corresponding to the prior, likelihood ratio, judged posterior, and calculated posterior probabilities using unaggregated data in Experiment 1.

Questions	*r*	95% Lower CI	95% Upper CI	*p* value
pTraitGroup – Calculated_pTraitGroup*	0.363	0.343	0.384	< .001
pTrait – pTraitGroup	0.227	0.206	0.247	< .001
pTrait – Calculated_pTraitGroup	0.213	0.192	0.233	< .001
pTrait – pGroupTrait	0.200	0.179	0.221	< .001

[11] A partial correlation showed $r_{JC} = 0.10$, $p < .001$, $t(8255) = 9.50$, *95% CI* = [0.08, 0.13], suggesting a relatively smaller r_{JC} correlation after controlling for responses to the other questions (i.e., questions corresponding to the priors, likelihood ratio) in Experiment 1.

Table D1 (cont'd)

pTraitGroup – Calculated_pTraitGroup	**0.153**	**0.132**	**0.174**	**< .001**
pTraitGroup – pGroupTrait	0.141	0.120	0.162	< .001
pGroupTrait – Calculated_pTraitGroup	0.131	0.110	0.152	< .001
pTrait – pGroup	0.116	0.094	0.137	< .001
pGroupTrait – pGroup	0.083	0.061	0.104	< .001
pTraitGroup – pGroup	0.070	0.049	0.092	< .001
Calculated_pTraitGroup – pGroup	-0.021	-0.043	-0.000	.049

Note. The r_{jc} correlation is in bold. * indicates the correlation after restricting the calculated posterior values between 0 to 100. pTraitGroup represents the judged posterior (question 1 in Equation 1), pTrait represents the prior (question 2 in Equation 1), pGroupTrait and pGroup represent components of the likelihood ratio (questions 3, 4 in Equation 1), Calculated_pTraiGroup represents the calculated posterior computed using the prior and likelihood ratio values.

When looking at the same correlation in a by-participant analysis, on average participants' judged posteriors were not significantly correlated with their calculated posteriors, $r_{JC} = 0.33$, $p = .39$, $t(12) = 1.36$, *95% CI* = [-0.19, 0.71] (see Figure D1 below).

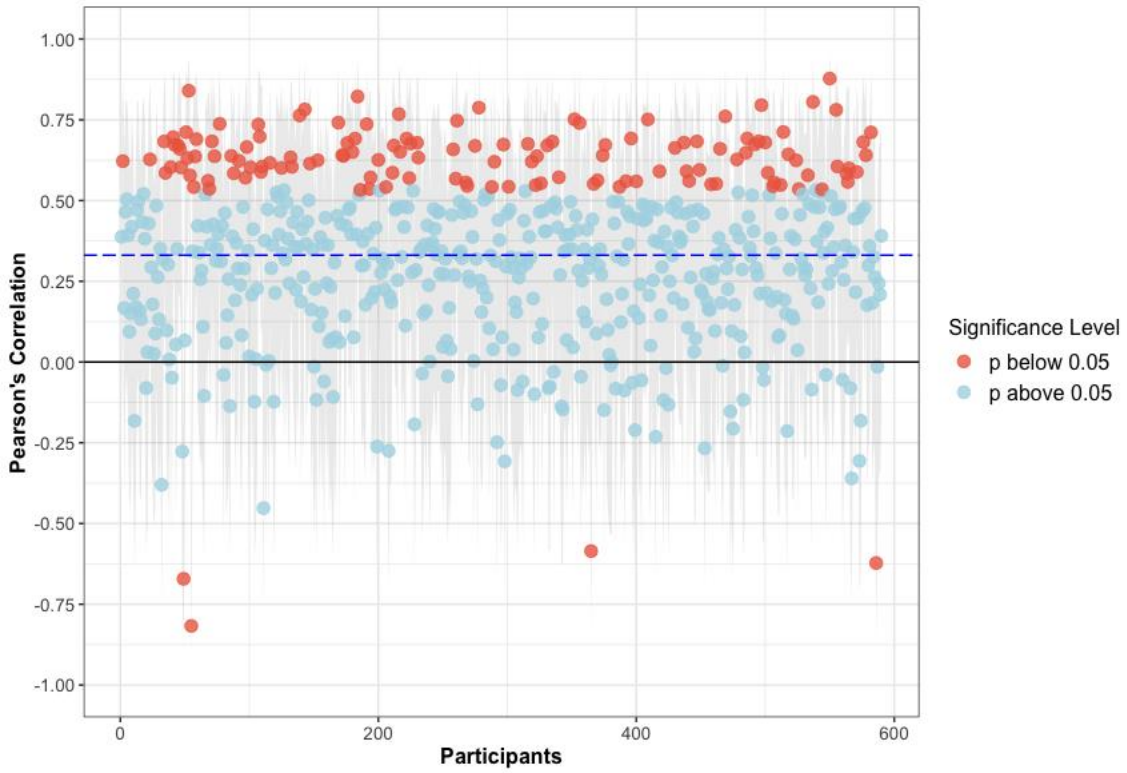

Figure D1
Scatter plot depicting the r_{JC} correlation in a by-participant analysis in Experiment 1. Each dot represents the r_{JC} correlation for a single participant. The dotted blue line indicates the mean correlation across participants. Grey ribbons in the background represent confidence intervals around the estimates.

When the above analyses were broken down by social category, the r_{JC} correlation ranged from small to moderate, with the correlation for the 'car salesman' category being non-significant in the overall sample; 'lawyer' and 'car salesman' being non-significant in the by-participant analysis (see Tables D2, D3 respectively). When the analyses were broken down by traits, the r_{JC} correlation again ranged from small to moderate with the 'extrovert' trait being non-significant in the overall sample; 'rich' and 'extrovert' being non-significant in the by-participant analysis (see Tables D4, D5 respectively).

Table D2

Overall correlations between the judged and calculated posterior probabilities across social categories in Experiment 1.

Category	*r*	95% Lower CI	95% Upper CI	*p* value
Male	0.567	0.534	0.597	< .001
Female	0.488	0.424	0.548	< .001
Night person	0.385	0.336	0.433	< .001
Asian	0.265	0.188	0.338	< .001
Morning person	0.226	0.171	0.279	< .001
Lesbian	0.210	0.131	0.286	< .001
Gay	0.200	0.121	0.276	< .001
Lawyer	0.181	0.101	0.258	< .001
Politician	0.129	0.049	0.208	.002
Car salesman	0.067	-0.013	0.146	.103

Table D3

Correlations between the judged and calculated posterior probabilities across social categories for the by-participant analysis in Experiment 1.

Category	*r*	95% Lower CI	95% Upper CI	*p* value
Male	0.358	0.286	0.427	< .001
Female	0.348	0.275	0.417	< .001
Night person	0.347	0.274	0.416	< .001
Morning person	0.321	0.247	0.391	< .001
Lesbian	0.213	0.135	0.289	< .001
Gay	0.179	0.100	0.256	< .001

Table D3 (cont'd)

Politician	0.174	0.095	0.251	< .001
Asian	0.157	0.078	0.235	< .001
Lawyer	0.082	0.001	0.161	.045
Car salesman	0.048	-0.032	0.128	.239

Table D4

Overall correlations between the judged and calculated posterior probabilities across traits in Experiment 1.

Trait	*r*	95% Lower CI	95% Upper CI	*p* value
Unemotional	0.576	0.627	0.519	< .001
Emotional	0.489	0.548	0.425	< .001
Dominant	0.476	0.536	0.411	< .001
Assertive	0.456	0.518	0.390	< .001
Unconventional	0.408	0.473	0.338	< .001
Depressed	0.378	0.446	0.307	< .001
Healthy	0.371	0.438	0.299	< .001
Smart	0.265	0.339	0.189	< .001
Activist	0.210	0.286	0.132	< .001
Flamboyant	0.200	0.277	0.122	< .001
Rich	0.181	0.258	0.102	< .001
Dependable	0.154	0.231	0.074	< .001
Greedy	0.130	0.208	0.050	< .001
Extrovert	0.067	0.147	-0.014	.104

Table D5

Correlations between the judged and calculated posterior probabilities across traits for the by-participant analysis in Experiment 1.

Trait	*r*	95% Lower CI	95% Upper CI	*p* value
Unemotional	0.461	0.395	0.522	< .001
Depressed	0.372	0.299	0.434	< .001
Healthy	0.371	0.298	0.438	< .001
Dominant	0.357	0.284	0.425	< .001
Emotional	0.348	0.275	0.417	< .001
Unconventional	0.301	0.226	0.373	< .001
Assertive	0.275	0.198	0.348	< .001
Activist	0.214	0.135	0.289	< .001
Flamboyant	0.18	0.1	0.256	< .001
Greedy	0.174	0.095	0.251	< .001
Smart	0.158	0.078	0.235	< .001
Dependable	0.153	0.072	0.231	< .001
Rich	0.082	0.002	0.162	0.046
Extrovert	0.049	-0.032	0.129	0.239

APPENDIX E: EXPERIMENT 2 SUPPLEMENTARY ANALYSES

The correlation using unaggregated data was smaller than the correlation reported in the main analysis (which aggregated across participants), $r_{JC} = 0.17$, $p < .001$, $t(7558) = 14.54$, *95% CI* = [0.14, 0.19]. This suggests weak alignment with Bayes' theorem for judgements about specific individuals overall. When exploring the r_{JC} correlation, I also checked correlations between all other questions in this task (i.e., questions corresponding to the priors, likelihood ratio) and to test whether any other questions in the task substantially impacted the r_{JC} correlation reported in the main analysis (e.g., particularly large *r* values). Table E1 lists all the correlations for all combinations of questions corresponding to the priors, likelihood ratio. Most of these correlations were moderate-sized and larger than the r_{JC} correlation, suggesting some caution in interpreting the r_{JC} correlation within this experiment.[12] For example, there was a large effect of relationship between people's prior beliefs and their posterior predictions ($r = 0.47$, $p < .001$, *95% CI* = [0.45, 0.48]), which may have influenced calculated posterior computations and in turn the r_{JC} correlation.

Table E1

Overall correlations between all combinations of questions corresponding to the prior, likelihood ratio, judged posterior, and calculated posterior probabilities using unaggregated (raw) data in Experiment 2.

Questions	*r*	95% Lower CI	95% Upper CI	*p* value
pGroupTrait – Calculated_pTraitGroup	0.577	0.562	0.592	< .001
pTrait – Calculated_pTraitGroup	0.531	0.515	0.547	< .001
pTraitGroup – Calculated_pTraitGroup*	0.471	0.451	0.492	< .001

[12] The partial correlation after controlling for all other variables (i.e., questions corresponding to the priors, likelihood ratio) was $r_{JC} = 0.01$, $p = .63$, $t(7555) = 0.48$, *95% CI* = [-0.02, 0.03].

Table E1 (cont'd)

pTrait – pTraitGroup	0.469	0.451	0.486	< .001
pTraitGroup – pGroupTrait	0.333	0.312	0.353	< .001
pTraitGroup – pGroup	0.282	0.261	0.303	< .001
pTrait – pGroupTrait	0.261	0.204	0.282	< .001
Calculated_pTraitGroup – pGroup	-0.203	-0.225	-0.181	< .001
pTrait – pGroup	0.193	0.171	0.215	< .001
pTraitGroup – Calculated_pTraitGroup	**0.165**	**0.143**	**0.187**	**< .001**
pGroupTrait – pGroup	0.157	0.135	0.179	< .001

Note. The r_{jc} correlation is in bold. * indicates the correlation after restricting the calculated posterior values between 0 to 100. pTraitGroup represents the judged posterior (question 1 in Equation 1), pTrait represents the prior (question 2 in Equation 1), pGroupTrait and pGroup represent the likelihood ratio (questions 3, 4 in Equation 1), Calculated_pTraiGroup represents the calculated posterior computed using the prior and likelihood ratio values.

When looking at the same correlation in a by-participant analysis, the average correlation was non-significant, $r_{JC} = 0.23$, $p = .35$, $t(12) = 0.97$, *95% CI* = [-0.28, 0.64] (see Figure E1 below).

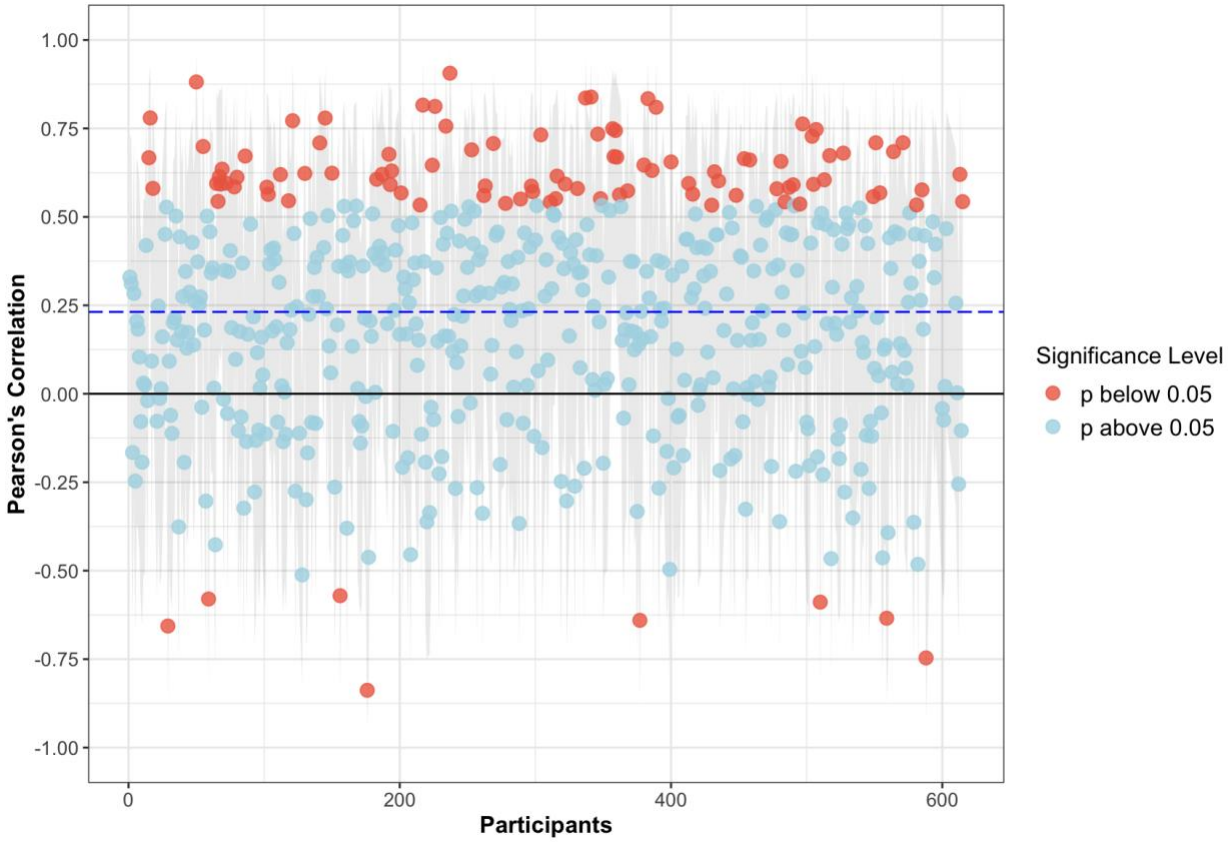

Figure E1
Scatter plot depicting the r_{JC} correlation from a by-participant analysis in Experiment 2 Each dot represents the r_{JC} correlation for a single participant. The dotted blue line indicates the mean correlation across participants. Grey ribbons in the background represent confidence intervals around the estimates.

When these analyses were broken down by social categories, the r_{JC} correlation ranged from small to moderate in both, the overall sample and by-participant analysis (see Tables E2, E3 respectively). When the analyses were broken down by traits, the r_{JC} correlation again ranged from small to moderate with the 'assertive' and 'dominant' traits being non-significant in the overall sample and in the by-participant analysis (see Tables E4, E5 respectively).

Table E2

Overall correlations between the judged and calculated posterior probabilities across social categories in Experiment 2.

Category	*r*	95% Lower CI	95% Upper CI	*p* value
Car salesman	0.447	0.376	0.512	< .001
Lawyer	0.345	0.268	0.416	< .001
Lesbian	0.343	0.266	0.415	< .001
Female	0.286	0.206	0.361	< .001
Morning person	0.283	0.226	0.336	< .001
Gay	0.220	0.138	0.298	< .001
Night person	0.194	0.135	0.250	< .001
Asian	0.193	0.111	0.273	< .001
Politician	0.145	0.061	0.227	< .001
Male	0.124	0.076	0.172	< .001

Table E3

Correlations between the judged and calculated posterior probabilities across social categories for the by-participant analysis in Experiment 2.

Category	*r*	95% Lower CI	95% Upper CI	*p* value
Car salesman	0.447	0.376	0.512	< .001
Morning person	0.344	0.267	0.416	< .001
Lesbian	0.343	0.266	0.415	< .001
Lawyer	0.302	0.224	0.377	< .001
Male	0.274	0.194	0.350	< .001
Female	0.273	0.193	0.349	< .001
Gay	0.208	0.125	0.287	< .001

Table E3 (cont'd)

Asian	0.193	0.111	0.273	< .001
Night person	0.171	0.088	0.252	< .001
Politician	0.139	0.056	0.221	.001

Table E4

Overall correlations between the judged and calculated posterior probabilities across traits in Experiment 2.

Trait	*r*	95% Lower CI	95% Upper CI	*p* value
Extrovert	0.447	0.512	0.377	< .001
Unemotional	0.358	0.429	0.282	< .001
Rich	0.345	0.417	0.268	< .001
Activist	0.343	0.415	0.266	< .001
Dependable	0.297	0.372	0.218	< .001
Emotional	0.286	0.362	0.207	< .001
Healthy	0.271	0.347	0.191	< .001
Flamboyant	0.220	0.299	0.138	< .001
Unconventional	0.213	0.292	0.131	< .001
Smart	0.194	0.274	0.111	< .001
Depressed	0.183	0.263	0.100	< .001
Greedy	0.145	0.227	0.062	.001
Dominant	0.060	0.143	-0.025	.166
Assertive	0.019	0.103	-0.066	.665

Table E5

Correlations between the judged and calculated posterior probabilities across traits for the by-participant analysis in Experiment 2.

Trait	*r*	95% Lower CI	95% Upper CI	*p* value
Extrovert	0.447	0.376	0.512	< .001
Activist	0.343	0.266	0.415	< .001
Unemotional	0.338	0.261	0.411	< .001
Rich	0.303	0.224	0.377	< .001
Dependable	0.286	0.206	0.361	< .001
Emotional	0.273	0.193	0.349	< .001
Healthy	0.271	0.191	0.347	< .001
Unconventional	0.207	0.124	0.286	< .001
Flamboyant	0.207	0.125	0.287	< .001
Smart	0.193	0.111	0.273	< .001
Depressed	0.178	0.095	0.259	< .001
Greedy	0.139	0.056	0.221	0.001
Dominant	0.058	-0.025	0.142	0.171
Assertive	0.011	-0.072	0.096	0.784

APPENDIX F: FUTURE DIRECTIONS PLOT

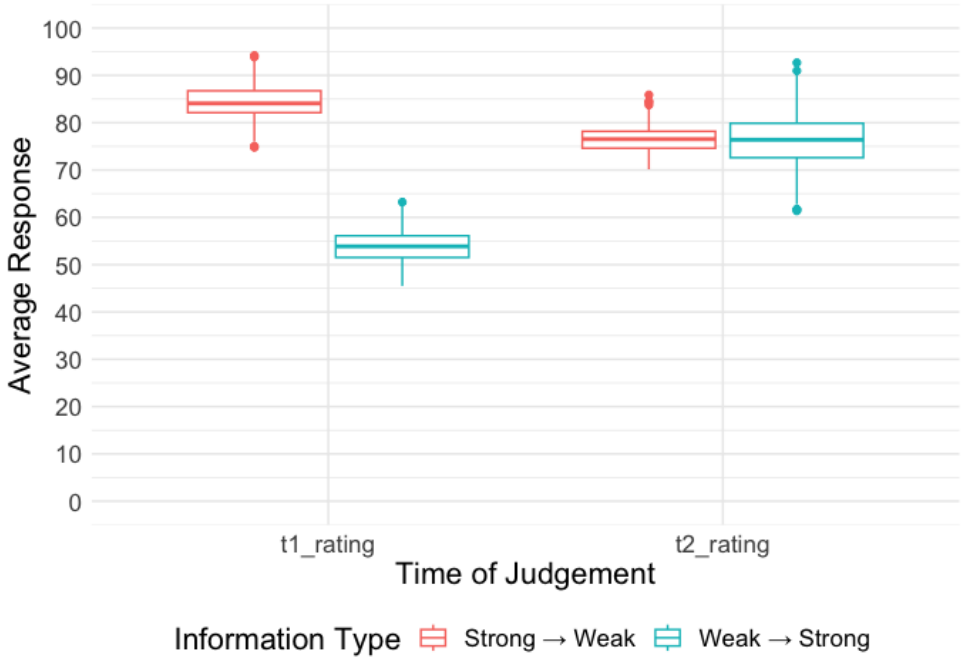

Figure F1
Box plot indicating the predicted difference in average participant responses at different time points — when strong information is presented as a prior followed by weak information or vice versa.