Asking the right questions? What people learn about strangers in conversation

Erik Brockbank[1], **Nora Dee**[1], **Misha O'Keeffe**[1,2], **Wasita Mahaphanit**[3], **Tobias Gerstenberg**[1], **Judith E. Fan**[1], **& Robert D. Hawkins**[2]
[1] Department of Psychology, Stanford University, USA
[2] Department of Linguistics, Stanford University, USA
[3] Department of Psychological and Brain Sciences, Dartmouth College, USA

Abstract

When meeting somebody for the first time, how do we get to know them? In the current work, we investigate how people learn about others' personalities through the questions they ask in conversation. Across two studies, participants completed a personality inventory then were paired with an online partner for a ten-minute chat. They were either instructed to get to know their partner in freeform conversation or were provided questions to discuss. The questions were either informative or uninformative for getting to know a stranger. Participants completed the same personality inventory about their partner afterwards. We test whether choosing from informative questions enabled participants to form a more accurate impression of their partner. We find that freeform conversation improved personality predictions overall, but differences in the informativeness of the questions discussed had minimal effects on accuracy; deep questions may only be as good as the disclosures they elicit.

Keywords: active learning; personality; conversation

Introduction

People mingle at dinner parties, strike up conversations at bars and bus stops, and introduce themselves to new co-workers, sometimes all in the same day. In these settings, we participate in a complex human ritual: interacting with people we've never met before. One of the core challenges faced by social humans is learning about the people around us, even those with whom we have no prior history. When interacting with somebody for the first time, we may want to know what they are like, to form impressions that will help us predict their behavior, decide whether to trust them, and determine whether we might get along. How do we arrive at this knowledge?

People know a lot about the people they know best; acquiring this knowledge about someone involves filling in a mental model that captures the ways they differ from other people we know (Anzellotti & Young, 2020; Tamir & Thornton, 2018; Vélez et al., 2019). Our mental models of others may include a broad swath of knowledge about them; however, it's been proposed that abstract traits or dispositions are a critical component of how we represent others, whether close friends or those we're meeting for the first time (Anzellotti & Young, 2020; Tamir & Thornton, 2018). For example, first impressions may be highly sensitive to people's *warmth* and *competence* (Fiske et al., 2007) or their *trustworthiness* and *social dominance* (Oosterhof & Todorov, 2008). Beyond first impressions, *power*, *sociality*, and *valence* may underlie many of the judgments we make about others (Thornton & Mitchell, 2018). Regardless of the exact format of our representations, learning about others seems to involve inferring abstract, trait-like dimensions (van Baar et al., 2022).

To learn this information, people rely on a diverse set of cues during interaction (FeldmanHall & Nassar, 2021; Zheng & Lin, 2024). For instance, people make rapid, spontaneous judgments about traits like *generosity* after observing others' actions (Uleman et al., 2008; Winter & Uleman, 1984); they infer *trustworthiness* from faces (Oosterhof & Todorov, 2008; Peterson et al., 2022); even the sound of someone's voice may carry signal about their personality (Lavan et al., 2024). Despite this array of social cues, this work largely overlooks a critical aspect of human social learning—in addition to *passive* inferences about others from their actions or appearance, people may *actively seek out* information about them.

The ability to engage in active learning through intervention and exploration has been critical to explaining how people acquire useful information about their environment (Coenen et al., 2019; Gureckis & Markant, 2012). In particular, recent work has emphasized people's ability to pose *informative questions* which maximize expected information gain about a task (Grand et al., 2024; Rothe et al., 2018) and to reuse questions that have been useful in the past (Liquin & Gureckis, 2022). Learning about other people through exploration—active *social* learning—requires a parallel capacity to ask informative questions. Recent work suggests that people have reliable intuitions about what makes a good question when getting to know somebody; further, these intuitions were broadly consistent with information about a speaker's personality that could be inferred from their answers (Brockbank et al., 2025). Thus, beliefs about what makes a good question may reflect the information it reveals about the respondent's personality.

Here, we build on these findings by investigating whether questions asked during conversation with a stranger help people learn about their personality, and how much it matters *which questions* they discuss. Across two studies, online participants were paired for a ten-minute conversation in which they were either instructed to simply *get to know* their partner, or were given different questions to guide the conversation (Figure 1). Participants completed a personality inventory before the conversation, then answered the same personality questions about their partner afterwards. We evaluate the relationship between the conversation setting in which participants got to know their partner, and the personality information they learned about them in the process.

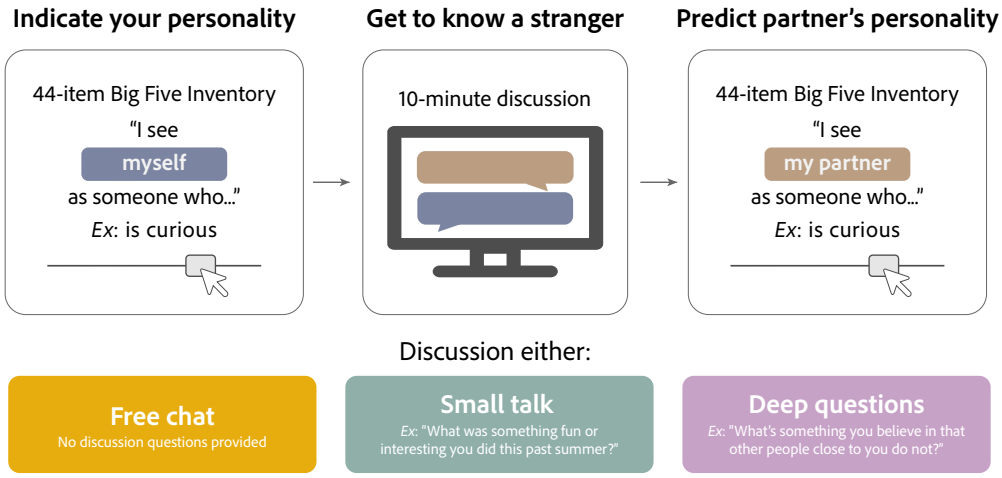

Figure 1: **Experiment overview**. Participants completed the 44-item Big Five personality inventory, then were matched with another online participant for a ten-minute conversation. Participants were either instructed to simply "get to know" their partner (*free chat*), or were prompted with questions meant to be informative (*deep questions*) or uninformative (*small talk*) for getting to know a stranger. After conversing, participants predicted their partner's responses on the same personality inventory.

Getting to know a partner in conversation

We conducted two closely related studies.[1] The first characterized what people learn about a stranger's personality during freeform conversation. The second tested whether prompting participants to discuss questions previously judged to be more or less informative modulates what they learn. Because the studies shared the same platform, procedure, and measures, we describe methods jointly below, noting where they differ.

Participants

We recruited 750 participants from Prolific. For Study 1 (*free chat*), we targeted 100 complete dyads; for Study 2 (*question prompting*), we targeted 100 dyads per condition (200 dyads total). Participants were excluded if they failed attention checks, failed to complete the full study or interact for the full duration of the chat, or experienced technical issues preventing dyad formation. When either member of a dyad was excluded, the full dyad was removed from analyses. Data collection and exclusion criteria were preregistered on the Open Science Framework. Of 250 individuals recruited for Study 1, 98 were excluded for the reasons above, leaving 76 complete dyads. Of the 500 individuals recruited for Study 2, 94 were excluded for the reasons above, leaving 203 complete dyads. Participants represented a mix of gender (50% female, 48% male), age (mean age: 41 years, SD = 13 years, range: 18–94 years), and race (77% White, 11% Black, 5% Asian).

Materials

Personality inventory Participants completed the 44-item Big Five Inventory (BFI; John & Srivastava, 1999), which measures five broad personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each item asks respondents to rate agreement with a self-descriptive statement (e.g., "*I see myself as someone who is talkative*") on a continuous slider ranging from 0 (strongly disagree) to 100 (strongly agree). The BFI is a widely used framework for characterizing individual differences in self-described personality (Goldberg, 1993) and has been proposed to align with abstract trait dimensions people use to reason about others (Tamir & Thornton, 2018); it is therefore an ideal candidate for distinguishing participants' inferences when discussing more and less informative questions. Two attention checks embedded in each BFI administration instructed participants to move the slider to an extreme endpoint. Participants whose response deviated more than 10 points from the target on any attention check were excluded.

Question prompts In Study 2, participants were shown a list of questions to guide their conversation. These questions were sampled from a large set normed in prior work in which raters judged how informative each question would be for getting to know a stranger (Brockbank et al., 2025). In the *deep questions* condition, participants chose from questions rated as highly informative (e.g., "*What would your past self think of your current self?*"). In the *small talk* condition, participants chose from questions rated as uninformative (e.g., "*Describe the last time you went to the zoo.*"). There were 234 questions in total (125 *deep questions* and 109 *small talk* questions).

Procedure

In both studies, participants first completed the BFI about themselves. They were then directed to a virtual lobby where they waited up to five minutes to be matched with a part-

[1] All experiment and analysis code, stimuli, and data are available on Github at: https://github.com/erik-brockbank/traits_from_conversation_cogsci_2026.

Figure 2: **Sample chat segments from the *Small talk* and *Deep questions* conditions**. Chosen questions are shown at the top, with each participant's initial response below.

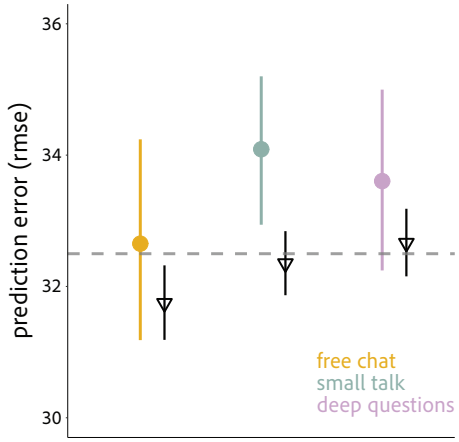

Figure 3: **Performance across experimental conditions**. Empirical and model-estimated root mean squared error (rmse) of personality predictions. Dashed line estimates the error that would be expected from predicting a *typical* response on each item. Error bars show 95% confidence intervals. Offset black points show model estimates.

ner. Participants who were not matched were excluded. Once matched, dyads proceeded to a text-based chat interface for a ten-minute conversation. In Study 1, participants were instructed to "get to know" their partner, with no further guidance about how to structure the interaction. In Study 2, dyads were randomly assigned to the *deep questions* or *small talk* condition. During the chat, participants took turns choosing a question to discuss. The question bank was populated with five questions sampled at random based on the dyad condition. A progress bar at the top of the chat indicated the recommended time (roughly two minutes) for each question. If participants discussed all five questions before the chat concluded, the question bank was refreshed with a new set of questions. After completing the chat, all participants predicted their partner's responses to the BFI. In a post-experiment survey, they were asked how well they felt they had gotten to know their partner and how accurate they believed their BFI predictions were.

Control study: predictions without conversation

As a comparison for participants' prediction accuracy about their partner, we also collected BFI predictions from participants who were never paired with a conversation partner. Instead, they were asked to give their best estimate of how a *typical person* would respond to each item ("*Imagine we selected a random person from the general population. Without knowing anything else about them, how do you think they would respond?*"). 50 participants were recruited from Prolific. We compare the accuracy of participants' BFI predictions in Study 1 and Study 2 to the accuracy that would be obtained by predicting the typical response on each item.

Results

If people use questions posed in conversation to learn what others are like, several predictions about such *active social learning* emerge. First, people should make more accurate predictions about a conversation partner's personality after talking to them than in the absence of conversation. Second, some questions should matter more than others when it comes to inferring a partner's personality. We test these hypotheses by comparing personality prediction accuracy across our studies. We fit multilevel Bayesian regression models (Bürkner, 2017) to participants' prediction error with random intercepts for variation in BFI item and participant within each dyad. We compare these models using *estimated log predictive density* (Δelpd) in cross-validated samples (Vehtari et al., 2017).

Impact of conversation on personality predictions To understand whether conversation helped people learn about what their partner was like, we compare participants' prediction accuracy to that obtained by guessing *typical* personality responses (Figure 3). Participants in Study 1 and Study 2 were matched with participants from our control study to generate sample prediction errors from guessing typical responses on each BFI item. We averaged the sample prediction errors for subjects across 1000 bootstrapped estimates to obtain a prediction error distribution for BFI predictions that lack individual information. We compare this to model estimates of the average error (rmse) in each condition from a multilevel Bayesian regression model fit to participants' prediction error with condition as a factor. The posterior probability that participants made more accurate predictions about their

	Free chat	Small talk	Deep questions
Questions	–	4.5 (1.2)	4.3 (1.4)
Messages	28.8 (10.0)	19.4 (9.9)	17.4 (7.1)
Words	288 (95)	242 (101)	255 (86)

Table 1: Summary of chats by condition (mean and SD).

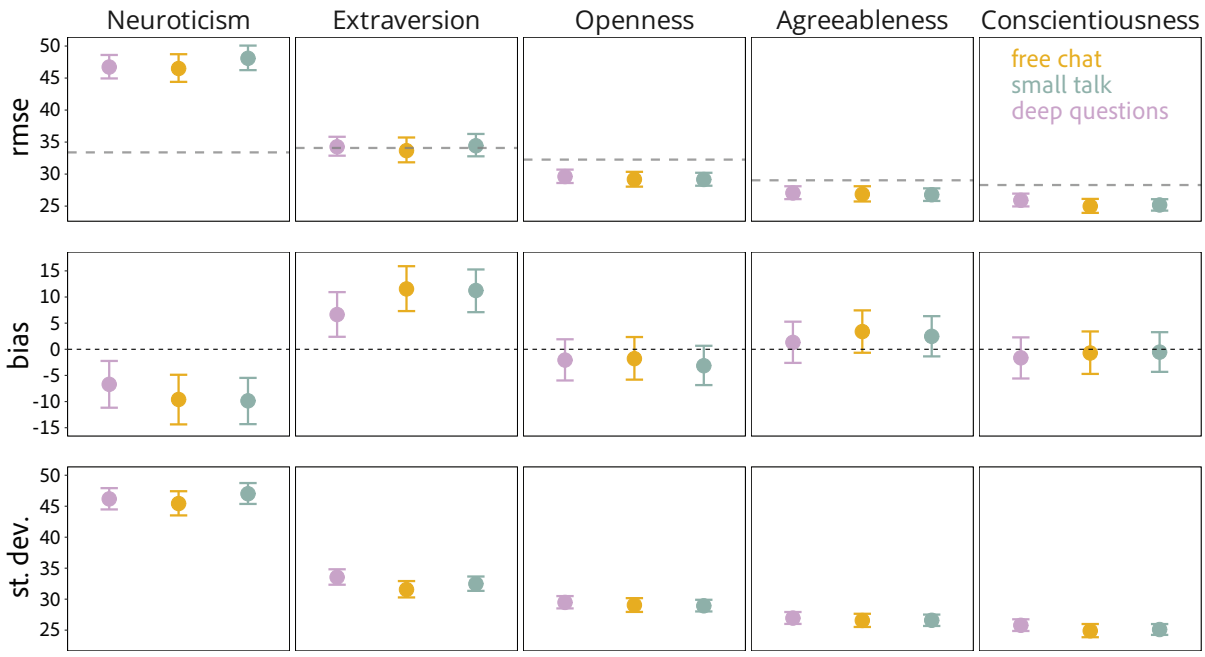

Figure 4: **Prediction accuracy by experimental condition for each Big Five personality trait**. (Top row): Prediction error (rmse), with dashed lines indicating the error that would be expected from predicting a *typical* response on the relevant personality items (lower values indicate more accurate predictions). (Middle row): Model estimates of the mean signed error (prediction *bias*) from predictions for each trait (values > 0 indicate overestimating the trait). (Bottom row): Model estimates of the uncertainty (*standard deviation*) from predictions for each trait. All error bars indicate 95% credible intervals.

partner than would be expected by guessing typical responses was largest in the *free chat* condition (98.1%, 95% credible interval: [97.8%, 98.4%]). This was followed by the *small talk* (67.3%, 95% CrI: [66.2%, 68.2%]) then *deep questions* (31.3%, 95% CrI: [30.3%, 32.4%]) conditions. Thus, participants' predictions about their partner after conversation were reliably more accurate than predictions about a typical person in the *free chat* condition, while accuracy improvements in the *deep questions* and *small talk* conditions were less clear.

Impact of deep questions and small talk on personality predictions If some questions reveal more about a conversation partner than others, participants' predictions about their partner's personality might have differed depending on whether they answered deep or small talk questions. We compare model estimates of prediction error (rmse) across conditions to determine whether error was lower among participants who discussed deep questions (Figure 3). Our multilevel Bayesian regression model fit to participants' signed prediction error estimated similar rmse values for the *free chat* (31.75, 95% credible interval: [31.19, 32.32]), *small talk* (32.35, 95% CrI [31.87, 32.84]), and *deep questions* (32.66, 95% CrI [32.15, 33.18]) conditions. This model did not show higher expected out-of-sample predictive accuracy than a null model that included only random effects (Δelpd = -0.2, *se* = 2.8). Different questions discussed across conditions did not enable more accurate overall predictions about one's conversation partner.

Impact of deep questions and small talk on individual trait predictions Participants might have exhibited differing levels of prediction accuracy about their partner for *individual traits* as a result of the questions they discussed (e.g., discussing deep questions might enable more accurate inferences about a partner's Conscientiousness than small talk). We fit a multilevel Bayesian regression model to participants' signed prediction error based on condition, Big Five personality dimension, and their interaction. The model allowed the mean and residual variance of prediction error to vary as a function of condition and personality dimension. We use model estimates of mean and variance to uncover whether differences in prediction error across traits and conditions were attributable to systematic biases or simply greater uncertainty (Figure 4).

Overall, prediction accuracy did not differ meaningfully across conditions for individual traits—the interaction between trait and condition modestly increased expected out-of-sample predictive accuracy (Δelpd = 2.9, *se* = 5.6). However, in all three conditions, participants' prediction accuracy varied substantially across traits (Figure 4, top row). In particular, predictions for Neuroticism (rmse: 47.16, 95% credible interval: [45.97, 48.46]) exhibited notably higher error than the other traits (*Extraversion*: 34.14, 95% CrI [33.01, 35.43]; *Openness*: 29.29, 95% CrI [28.66, 29.95]; *Agreeableness*: 26.85, 95% CrI [26.24, 27.55]; *Conscientiousness*: 25.35, 95% CrI [24.79, 25.96]). Pairwise differences between the other traits were smaller but credibly greater than zero.

Trait-level prediction error also exhibited substantial deviations from what would be expected by guessing typical responses to the BFI items for each trait (Figure 4, top row). Relative to this control, prediction error was *larger* for Neuroticism (mean difference in rmse: 13.08, 95% CrI [13.07, 13.10]) and reliably *smaller* for Agreeableness (mean difference: −7.23, 95% CrI [−7.24, −7.22]), as well as Openness (mean difference: −2.97, 95% CrI [−2.98, −2.96]) and Conscientiousness (mean difference: −2.94, 95% CrI [−2.94, −2.93]). Thus, conversations led participants to make more inaccurate predictions about their partner's Neuroticism, while also generating more accurate predictions about their Agreeableness, Openness, and Conscientiousness.

This pattern can be explained in part by trait-level *biases* in participants' predictions (Figure 4, middle row). Predictions for Neuroticism exhibited a strong negative bias (underattributing the trait to one's partner; $\mu = -8.71$, 95% CrI [−12.43, −4.98]), while predictions for Extraversion exhibited a positive bias ($\mu = 9.83$, 95% CrI [6.31, 13.33]). Bias for the remaining traits were relatively small and not credibly different from zero (*Openness*: $\mu = -2.41$, 95% CrI [−5.68, 0.76]; *Agreeableness*: $\mu = 2.33$, 95% CrI [−1.05, 5.86]; *Conscientiousness*: $\mu = -0.88$, 95% CrI [−4.34, 2.60]).

Prediction error for individual traits can also be explained by the variability of participants' prediction error (Figure 4, bottom row). Model estimates for the standard deviation of prediction error were substantially larger for Neuroticism ($\sigma = 46.31$, 95% CrI [45.27, 47.38]) than all other traits (*Extraversion*: $\sigma = 32.65$, 95% CrI [31.96, 33.37]; *Openness*: $\sigma = 29.14$, 95% CrI [28.57, 29.72]; *Agreeableness*: $\sigma = 26.69$, 95% CrI [26.13, 27.28]; *Conscientiousness*: $\sigma = 25.28$, 95% CrI [24.74, 25.84]). Differences between the remaining traits were once again small but consistently greater than zero.

Impact of individual questions on trait predictions Participants may have learned about their conversation partner's individual traits through their responses to particularly *diagnostic* questions. To test this, we fit a Bayesian multilevel regression model predicting participants' trait inferences from the questions they discussed in Study 2. In particular, we modeled trait-level prediction accuracy using rmse computed separately for each Big Five trait and standardized across all observations. The model treated questions as "treatments" by including random intercepts for question, as well as question-by-trait, to capture trait-specific deviations for each question.

First, we assess whether individual questions improved or worsened prediction error based on the model-estimated standard deviation of question-level random effects and question-by-trait random effects. The standard deviation of overall question effects was small (0.012, 95% credible interval [0.0004, 0.034]), suggesting minimal differences between questions in their overall impact on prediction accuracy. The standard deviation of question-by-trait effects was similarly small (0.030; 95% credible interval [0.0015, 0.079]), indicating little evidence that particular questions systematically facilitated inference for specific personality traits.

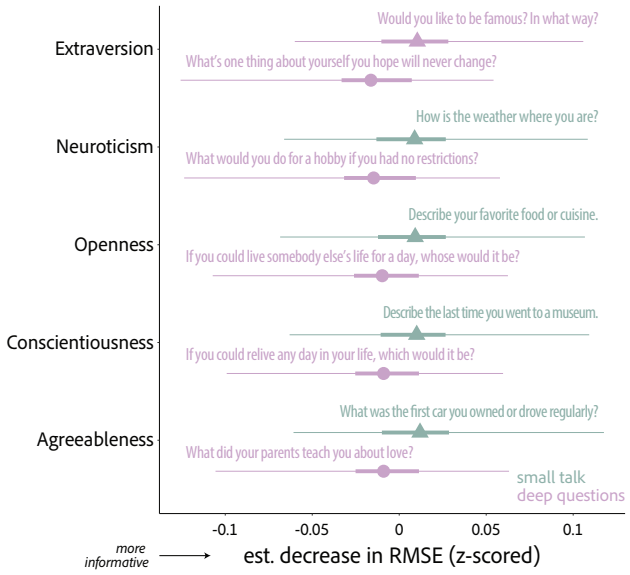

Figure 5: **Most and least informative questions**. Individual questions that were most (triangles) and least (circles) informative for each Big Five trait. Error bars show 50% and 95% credible intervals estimated from model posteriors.

Although overall variance attributable to individual questions was small, we examined posterior estimates of the largest and smallest individual question-by-trait effects (i.e., questions that most strongly reduced or increased prediction error for each trait). The *most informative* questions for each trait exhibited similar levels of prediction improvement (Figure 5); four of the five came from the *small talk* condition. Meanwhile, among the *least informative* questions for each trait, all of them came from the *deep questions* condition.

What makes a revealing conversation? Participants' conversations likely differed not only in the questions discussed, but in a host of ways arising from the participants themselves. In exploratory analyses, we investigate features of the participants and their conversations that might have impacted prediction accuracy. First, we examine whether dyads who were more similar made more accurate predictions about each other (Figure 6A). To estimate personality similarity, we calculated the Euclidean distance between participants' Big Five trait values in each dyad. In addition, we calculate the Euclidean distance between participants' Big Five trait values and their *predictions* to correct for higher prediction accuracy attributable to guessing values similar to one's own. After controlling for predictions similar to oneself, dyads who had more similar personalities made more accurate partner predictions ($\beta = 0.21$, 95% CI [0.18, 0.23], $p < .0001$); this effect was significant for correlation and cosine distance as well.

Participants also varied in their subjective sense of having gotten to know their partner. In a post-experiment survey, they were asked: *How well do you think you got to know your partner?* (mean = 54.6, SD = 25.5) and *How accurate do*

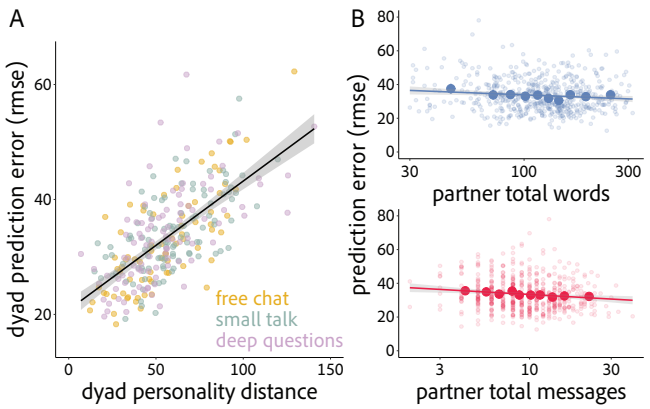

Figure 6: **Additional accuracy indicators**. (A) Prediction accuracy (rmse) and personality similarity (Euclidean distance) in each dyad. (B) Relationship between prediction error (rmse) and features of partner messages. Top: total words written by the partner (two participants who wrote fewer than 15 words are excluded). Bottom: total messages sent by the partner. Large points show mean and 95% CI by decile.

you think your predictions were? (mean = 58.3, SD = 21.1). Responses did not differ by condition (*Get to know partner*: $F(2, 555) = 1.75$, $p = .17$; *Predict partner*: $F(2, 555) = 0.64$, $p = .53$) and did not reliably relate to prediction accuracy (*Get to know partner*: $\beta = -0.03$, 95% CI $[-0.06, 0.003]$, $p = .08$; *Predict partner*: $\beta = -0.02$, 95% CI $[-0.06, 0.02]$, $p = .31$).

Finally, we test whether features of the conversation itself impacted what people learned about their partner. We evaluate whether participants whose partner sent *more* or *longer* messages made more accurate personality predictions. We fit linear regressions to each participant's signed prediction error (rmse), with the log-transformed total word count of their partner's messages and the log-transformed total number of messages sent by their partner as predictors (Figure 6B). We find that prediction error decreased significantly as a function of the log_{10}-transformed number of words sent by the partner ($\beta = -4.44$, 95% CI $[-8.10, -0.78]$, $p = .02$). Meanwhile, we observe an even stronger negative relationship between participants' prediction error and the log_{10}-transformed total number of messages sent by their partner ($\beta = -5.71$, 95% CI $[-9.38, -2.05]$, $p = .002$). These relationships did not differ reliably across conditions (*total words*: $F(2, 552) = 0.22$, $p = .80$; *total messages*: $F(2, 552) = 0.33$, $p = .72$).

General Discussion

From a preschooler's first time being let loose at a playground to an adult's first day at a new job, our social lives often involve interacting with people we've never met. In these settings, how do we learn the kind of information about them that helps us decide if we can trust them or want to spend more time with them? Across two studies, we found that freeform conversation helped participants draw more accurate inferences about their partner's personality than would be expected from "typical" responses, yet prediction accuracy did not vary as a function of the questions participants discussed. Instead, across conversation settings, we observed similar trait effects, with highly inaccurate under-attributions of Neuroticism, and accuracy improvements for Agreeableness, Openness, and Conscientiousness. These impacts likely canceled each other out in aggregate. Thus, while participants' social inferences showed little sensitivity to the questions discussed, conversation systematically impacted their trait learning.

People have strong intuitions that some questions tell us more about a stranger than others (Brockbank et al., 2025). Why might participants in the *deep questions* condition have failed to make more accurate personality inferences? One possibility is that in brief conversations with a stranger, pragmatics dominate. Participants may have been less willing to disclose personal information in response to more probative questions without sufficient time to build trust with their partner out of fear that they would be judged (Kardas et al., 2022). Questions posed in freeform conversation may have allowed them to achieve greater mutual disclosure more naturally.

Alternatively, deep questions may have elicited more disclosure, but the Big Five Inventory may not fully capture what participants learned about their partner. Prior work suggests that discussing questions meant to elicit greater personal disclosure leads people to *feel closer* to a stranger than small talk (Aron et al., 1997). Recent work has explored the relationship between feelings of closeness and inferences about shared knowledge, beliefs, and values (Mahaphanit et al., 2024; Rossignac-Milon et al., 2021). However, while personality traits such as the Big Five may be critical to our representations of others, the precise relationship between our intuitive theory of informative questions and the way such questions support social learning remains poorly understood.

What, then, does it mean to *actively* learn about another person through conversation? Work on active learning in other domains has emphasized the selection of informative queries whose possible answers reduce uncertainty. Our findings suggest that active *social* learning may operate differently. In physical domains, the learner controls what queries to pose and the environment responds deterministically. Conversation, by contrast, is a joint activity in which both parties must coordinate not only what is said but how it is understood (Clark & Brennan, 1991). The answer to a question depends on the respondent's willingness to engage. Our results suggest that this latter challenge may be half the battle. First, participants in the *free chat* condition, who could adaptively steer the conversation, showed the strongest evidence of learning. Next, the higher prediction accuracy by participants who had a more "chatty" partner suggests that more disclosive conversation is more informative *when it is achieved*. Finally, participants' outsize underestimation of Neuroticism (and overestimation of Extraversion) may reflect efforts by their partner to manage how these traits came across. Future work on active social learning may broaden our understanding of how people learn about the environment when the environment talks back.

Acknowledgments

E.B. is supported by NSF SBE Postdoctoral Research Fellowship #2404706. T.G. is supported by grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and from the Cooperative AI Foundation. J.E.F. is supported by NSF CAREER Award #2436199, NSF DRL #2400471, and awards from the Stanford Human-Centered AI Institute (HAI) and Stanford Accelerator for Learning.

References

Anzellotti, S., & Young, L. L. (2020). The acquisition of person knowledge. *Annual Review of Psychology*, *71*(1), 613–634.

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and social psychology bulletin*, *23*(4), 363–377.

Brockbank, E., Gerstenberg, T., Fan, J., & Hawkins, R. (2025). How do we get to know someone? diagnostic questions for inferring personal traits. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149).

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, *25*(12), 1045–1057.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77–83.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist*, *48*(1), 26.

Grand, G., Pepe, V., Andreas, J., & Tenenbaum, J. B. (2024). Loose lips sink ships: Asking questions in battleship with language-informed program sampling. *arXiv preprint arXiv:2402.19471*.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research*. Guilford Press.

Kardas, M., Kumar, A., & Epley, N. (2022). Overly shallow?: Miscalibrated expectations create a barrier to deeper conversation. *Journal of personality and social psychology*, *122*(3), 367.

Lavan, N., Rinke, P., & Scharinger, M. (2024). The time course of person perception from voices in the brain. *Proceedings of the National Academy of Sciences*, *121*(26), e2318361121.

Liquin, E. G., & Gureckis, T. M. (2022). Where questions come from: Reusing old questions in new situations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Mahaphanit, W., Welker, C., Schmidt, H., Chang, L., & Hawkins, R. (2024). When and why does shared reality generalize? *Proceedings of the annual meeting of the cognitive science society*, *46*.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17), e2115228119.

Rossignac-Milon, M., Bolger, N., Zee, K. S., Boothby, E. J., & Higgins, E. T. (2021). Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology*, *120*(4), 882.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.

Tamir, D. I., & Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in cognitive sciences*, *22*(3), 201–212.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral cortex*, *28*(10), 3505–3520.

Uleman, J. S., Adil Saribay, S., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, *59*, 329–360.

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, *6*(3), 404–414.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vélez, N., Bridgers, S., & Gweon, H. (2019). The rare preference effect: Statistical information influences social affiliation judgments. *Cognition*, *192*, 103994.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? evidence for the spontaneousness of trait inferences. *Journal of personality and social psychology*, *47*(2), 237.

Zheng, R., & Lin, C. (2024). Cues driving traitimpressions in naturalistic contexts are sparse. https://doi.org/10.31234/osf.io/v39tk