Who Should Be My Teammates: Using A Conversational Agent to Understand Individuals and Help Teaming

Ziang Xiao
University of Illinois at Urbana-Champaign
Urbana, IL, U.S.A
zxiao5@illinois.edu

Michelle X. Zhou
Juji. Inc.
San Jose, CA, U.S.A
mzhou@acm.org

Wat-Tat Fu
University of Illinois at Urbana-Champaign
Urbana, IL, U.S.A
wfu@illinois.edu

ABSTRACT

We are building an intelligent agent to help teaming efforts. In this paper, we investigate the real-world use of such an agent to understand students deeply and help student team formation in a large university class involving about 200 students and 40 teams. Specifically, the agent interacted with each student in a text-based conversation at the beginning and end of the class. We show how the intelligent agent was able to elicit in-depth information from the students, infer the students' personality traits, and reveal the complex relationships between team personality compositions and team results. We also report on the students' behavior with and impression of the agent. We discuss the benefits and limitations of such an intelligent agent in helping team formation, and the design considerations for creating intelligent agents for aiding in teaming efforts.

CCS CONCEPTS

• **Human-centered computing** → **Human-Computer Interaction**; • **Computing methodologies** → **Intelligent agents**;

KEYWORDS

Teaming; Team Formation; Team Companion; Personality Inference; Conversational Agent; Chatbot

ACM Reference Format:
Ziang Xiao, Michelle X. Zhou, and Wat-Tat Fu. 2019. Who Should Be My Teammates: Using A Conversational Agent to Understand Individuals and Help Teaming. In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Rey, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3301275.3302264

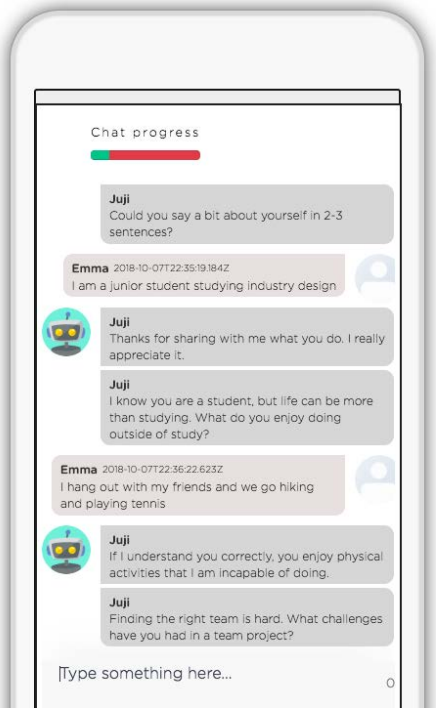

Figure 1: A screenshot of an example conversation that INDIGO supports with a student named Emma.

1 INTRODUCTION

The whole is bigger than the sum of parts. In many real-world tasks, such as workplace collaborations and student course projects, team effort not only is required, but also plays a pivotal role in their success. Many organizations thus consider teamwork—one's ability to collaborate with others and effectively work in a team—one of the most important and desirable skills of their employees. Numerous research efforts show that team performance can be influenced by a number of factors, such as individual team members' characteristics and team personality compositions [2, 5, 18, 25, 31, 41]. It is a daunting task to understand the interactions of all these factors and their effects on team performance, let alone using these factors to guide team formation and behavior. Although tools exist in aiding teaming efforts (e.g., team formation [10]), they lack the abilities in understanding individual team members, team compositions, and their relationships with team performance [1].

Given the importance of teamwork, there has been an increasing focus on engaging students in team projects for them to practice and thereby improve their teamwork skills. Moreover, research on cooperative learning has shown that learning can be enhanced in group environments in which students can learn more actively (e.g., [23]). Educational environments also serve as a good testbed for studying

IUI '19, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302264

teaming, since there are often many teams with varying characteristics, and team performance can be objectively measured and compared by the same or similar set of assignments and projects.

To study and potentially guide student teaming efforts in an educational setting, we have developed a conversational agent called INDIGO (Individual Differences for Group Optimization) that can interact with a user in a one-on-one text-based chat. INDIGO can also automatically infer a user's personality traits based on his/her chat behavior *without* directly asking any personality test questions, which is known to be problematic [54].

INDIGO aims at achieving three goals. First, it replaces traditional online surveys to gather initial information from students, such as their team preferences and expectations. Interacting with a chatbot like INDIGO is a new experience, which may help combat survey fatigue and collect more in-depth information.

Second, INDIGO replaces a traditional personality test to automatically gauge students' personality traits objectively, preventing potential faking in such a test [54]. In an educational setting, students might provide less truthful answers in a traditional personality test to make themselves more desirable to potential teammates. To prevent faking, we want to automatically infer a student's personality traits without asking any direct personality test questions (e.g., asking for a self-reported rating on "I get angry easily"). The inferred personality traits can then be used to study team personality compositions and their effect on team performance [5, 18, 20]. While personality inference tools exist (e.g., IBM Watson Personality Insights [1]), they often require a certain amount of data (e.g., a minimal 1000 words from one's social media account) that many student might not have. Naturally, a chatbot, as described in [27], would help us address this challenge. In short, INDIGO can kill two birds with one stone: eliciting student input regarding teaming (e.g., a student's team preferences and team experience) and using the same input to automatically infer student personality traits without asking any additional personality test questions.

Third, INDIGO is intended to serve as a long-term team companion that can follow a team and interact with team members continuously during their team efforts. If shown feasible, INDIGO may therefore serve as a useful tool to help conduct longitudinal team studies, during which it will detect changing team dynamics and potentially guide team behavior based on such changes.

To achieve these goals, we deployed INDIGO in a real-world setting, where it was used to interact with 201 students who enrolled in a large engineering class at a university, formed 40 teams, and engaged in semester-long team projects. INDIGO interacted with each student at the beginning and end of the class to elicit their views and opinions about teaming, including team preferences and reflections. It also automatically inferred the students' personality traits based on their chat behavior [27]. We recorded students' interaction with INDIGO including their perceptions of INDIGO. We also tracked each team's performance throughout the semester. For comparison purpose, we also used traditional surveys to gather students' input at the beginning and end of the semester to understand their team preferences and perceptions. Figure 1 shows an example chat between INDIGO and a user.

[1]https://www.ibm.com/watson/services/personality-insights/

Based on the collected data, we performed a series of analyses to answer three key research questions.

- **RQ1**: How well do students interact with INDIGO?
- **RQ2**: How effectively can INDIGO gather information from students?
- **RQ3**: How well can the personality traits computed by INDIGO provide insights into the relations between team compositions and team performance?

Our analyses reveal that students interacted with INDIGO for an extensive period of time (e.g., 60-minute chat in their pre-course interview) and offered open and honest input. Moreover, INDIGO elicited rich information from the students that allows instructors to better understand student team preferences and perceptions. Third, specific team personality compositions based on the inferred personality traits *significantly* predict team perception and team performance.

As a result, our work offers three unique contributions. First, it suggests a new approach to team formation. Through text-based conversations, INDIGO can interview students to gather their team preferences and expectations, and measure their personality traits. It can then use such results to recommend team compositions for optimal teaming. Second, our use of INDIGO through a real-world teaming task demonstrates its practical value beyond team formation. Given that the students were willing to spend an extensive period of time with INDIGO, INDIGO could be potentially used as a team companion to accompany teams longitudinally, track team changes, and guide team behavior in real time. Third, our work presents a novel, effective method for researchers to investigate teaming in the real world and better understand how team composition impacts team performance.

2 RELATED WORK

Our work is related to several areas of work on understanding factors that impact teaming as well as the use and evaluation of conversational agents for various tasks.

2.1 Effect of Team Personality Composition

A rich body of research shows that team personality composition influences team performance and team members' perception of their team. For example, Lykourentzou et al. show that teams with a balanced personality composition outperform teams with a surplus of leader-type personality [31]. Bell uses a meta-analysis to show how psychological variables, such as personality traits, values, and abilities, predict team performance [5]. Halfhill et. al find that the average level of *Agreeableness* and *Conscientiousness* (two of Big 5 personality factors) in a group correlate with group performance in Military service teams [18]. Whelan and colleagues indicate that different team personality compositions influence virtual team performance as well as team satisfaction [48]. Humphrey et al. suggest two types of team configurations, complementary vs. supplementary fit, which maximizes or minimizes variances on different personality traits to optimize team performance [20]. In educational settings, Karn and Cowling observed although homogeneous teams may experience less team conflicts, they may fall into the no debate trap which leads lower team performance [25]. Rutherfoord used the Keirsey Temperament Sorter to form teams

in a software engineering class and found heterogeneous groups perform better in problem solving [41].

Although existing studies show a great potential of using personality as a criterion for team formation, it is non-trivial to implement such a solution in the real-world due to several challenges. In our case, one challenge is how to obtain students' authentic personality traits. Another challenge is what team personality compositions should be used to suggest student teaming in an educational setting. Our work presented here is precisely set out to explore the feasibility of addressing these challenges. As a result, not only do we present here an effective and practical approach to studying the effect of personality compositions on teaming, but our results also reveal new findings that none of prior studies has discovered.

2.2 Effect of Team Perception

Everyone wants to work in a supportive and trustworthy team. Numerous studies investigate what might affect team members' perceptions of their own team and how team perception impacts team performance [4, 9, 22, 43]. For example, Jehn et al. find that team members are more satisfied if the team is more gender balanced [22]. Burress shows that leader behavior in a team also affects team perception. A balanced personality composition in a team could also improve team member's perception [9]. Most importantly, the team perception is tightly linked to the team performance [24, 39, 44]. A coherent working relationship among team members can also help the team achieve better performance [24].

While we leverage existing findings, we must deal with special challenges in our educational setting. In our case, teams usually are formed for only one semester and team members have little time to bond with each other. If initial team configurations could project a positive team perception at the start, students in a team could dive into the teamwork right away. Therefore, our study aims at discovering how the individual differences of team members might influence team perception especially at the team formation stage.

2.3 Applications of Conversational Agents

Conversational agents have been used for various tasks [45]. In general, there are two main types of conversation agents. One type is task-oriented conversational agents that help users accomplish concrete tasks, such as information inquiry [53] and event scheduling [55]. The second type is to socialize with users without specific goals (e.g., [29, 49]). Recently, researchers have started building conversational agents that can interleave between tasks and social chitchat and handle usersâĂŹ emotional and information requests. [36, 50, 51]. Compared to these works, we are experimenting with conversational agents that can support both task-oriented and social dialogue. However, our unique focus is its use in facilitating teaming efforts.

To evaluate the quality of conversational agents, researchers have proposed many evaluation criteria based on an agent's purpose. For example, task-oriented agents were often evaluated by examining if a user can complete a task with an agent's help [47]. On the other hand, the quality of conversation is often used to evaluate social agents [19]. Since INDIGO is designed to facilitate teaming efforts, we adopt the evaluation criteria that help us evaluate the effectiveness of INDIGO in aiding teaming.

2.4 The Use of Virtual Interviewer to Assess Personality

To effectively infer one's personality traits from text, a sizable amount of text is often required (e.g., a minimal of 1000 words in [27] and at least 3000 words in [17]). To obtain one's communication text [2], recently researchers have built a chatbot that can chat with a user through a text-based conversation and automatically infer the user's personality traits based on the conversation [27]. Studies also show that users are willing to confide in an AI agent [27], including the disclosure of sensitive information [30].

Given the benefits of an AI agent, INDIGO is built to interact with students and assess the students' personality traits automatically. We can then examine how the inferred traits may predict team results.

2.5 Team Formation Methods in a Classroom

Three common methods are used to form student teams in a large class: self-selection, random assigned, and criteria-based approach [21]. In the self-selection approach, students form their own teams based on their previous experience or preference, which is often preferred by students who know their classmates and have potential teammates in mind. However, this approach often leads to homogeneous teams, which may not lead to optimal team performance due to group thinking or a lack of required complementary skills [3, 22, 25]. In addition, students may not always know each other in the same class, let alone having potential teammates in mind. Therefore, instructors often need to help students form teams. While random assignment is an alternative solution to ensure that every student be on a team, this approach may not help team dynamics, let alone team success.

To mitigate the limitations of self-selection and random approaches, instructors often use criterion-based approaches to divide students into teams. Existing studies show that teams formed based on proper criteria have a better team dynamics and team performance than self-selected and random-assigned teams [8, 48]. However, selecting the right criteria is critical to the success of this approach and is often challenging, since not every instructor is an organizational psychologist and familiar with teaming criteria and their effects on teaming and performance. Thus the goal of our work is to help simplify teaming criteria by examining whether personality composition alone helps teaming.

2.6 Existing Team Formation Tools

To support criterion-based team formation, a number of tools are developed. A notable one is CATME (Comprehensive Assessment for Team Members). It allows instructors to assign teams with 27 criteria, including skills, working styles, and demographics [10, 21]. A recent study shows that the students appreciated the use of rational criteria and thought the tool gave them a fair chance to be assigned to a good team [21]. The instructors also reported that the tool reduced their burden and stress in grouping students into teams [21]. In addition to CATME, another tool emphasizes schedule compatibility [26]. Yet another automated team formation

[2] Our experiments found that 90% of people produce fewer than 200 words on Facebook

tool by Del Val et al. focuses on collective intelligence and coalition structure generation, and has also received positive feedback from the students [11]. Compared to these tools, which mostly use traditional surveys to understand team members and derive teaming criteria, INDIGO helps teaming, including team formation, by deeply understanding and potentially guiding team members through intelligent conversations.

3 INDIGO SYSTEM OVERVIEW

To better understand and aid teaming efforts, we designed a conversational agent, called INDIGO (Individual Differences for Group Optimization). As the first step, INDIGO is designed to engage each student in a one-to-one text-based chat, eliciting valuable information from the student and automatically inferring his/her characteristics.

We built INDIGO on top of the Juji chatbot platform (juji.io), where researchers can create and deploy their own customized chatbots for various tasks. For example, one application of their chatbots is conducting job interviews [27]. We chose to use the Juji platform for two reasons. First, the customization and deployment of a Juji chatbot is very similar to creating a survey in an existing tool like CATME [10], which makes the replication of our study feasible for others. Specifically, a course instructor or his/her teaching assistant uses a GUI to input a set of interview questions and the order of the questions to be asked. A chatbot is then automatically created with a set of default capabilities, such as handling basic social dialogs [7]. Table 1 lists the chatbot's key conversational skills. The chatbot can then be distributed via a generated URL, similar to the distribution of a survey URL.

Second, Juji chatbots offer several functions that meet the practical goals of INDIGO. Specifically, we wanted a chatbot that can function robustly in the real world and support long conversations. Juji has demonstrated that their chatbots can conduct hourly-long job interviews in the real world [27]. In addition, Juji's chatbot can automatically infer users' Big 5 personality traits based on their chat behavior. This offers us a method to measure students' personality insights without asking them to take additional personalty test questions. We can also use the personality insights to investigate the relationship between team personality compositions and team performance.

4 INDIGO FIELD DEPLOYMENT

To evaluate INDIGO, we conducted a field study in a Computer Science course offered at a large public university. This course required students to team up and collaborate on a semester-long project to create a web/mobile application, which determined 40 % of their final course grade.

4.1 Participants

A total of 201 students enrolled in the class and 184 of them consented to participate in the study (N=184). While the majority of the students majored in Computer Science, there were students from non-engineering majors, such as Industrial Design, Information Science, and Psychology.

Response Feedback
Acknowledge a user's responses and provide positive feedback to encourage information disclosure
User: "I am a student majoring in Computer Science."
Juji: "Wow, you must love science. I hope you like AI and me."
Response Prompting
Ask for more information
Juji: "I'd love to hear more, go on..."
Response Probing
Probe answers to elicit deeper information
User: "I love to read books and cook."
Juji: "Why do you like to read?"
Side Talking
Handle social dialogues and unexpected situations.
User: "Why do you ask me this?"
Juji: "I guess my creators told me so. Now I've answered your question, could you answer mine?"

Table 1: The chatbot's key Conversational features.

4.2 INDIGO Interviews

We deployed INDIGO at the beginning and the end of the semester to interview all students enrolled in the class.

Each pre-course interview included five sections. The first section was a warm-up conversation, during which INDIGO and a student introduced to each other and chatted about their hobbies and favorite movie. The second section included a 20-item Impression Management (IM) Scale questionnaire to measure how students consciously favor themselves to impress others [37]. Since the IM scale is found highly related to self-control and social adaptation in workplace [46], we hypothesized that it might affect team performance. In the third section, INDIGO asked the student a set of open-ended questions regarding his/her team experience and team preferences. For example, it asked a student "what is your preferred role in a team" and *"what kind of individuals do you want to have on your project team?"*. In the fourth part of the interview, INDIGO discussed with the student about his/her own characteristics, such as his/her strengths and weaknesses. The purpose of this discussion was to gauge how much a student is willing to trust INDIGO and share personal information including sensitive information such as one's weaknesses. The last section was to solicit the student's feedback about INDIGO. The student was asked to express his/her impression of INDIGO and rate INDIGO on a number of scales, such as its *helpfulness* and *likeability*.

At the end of the semester, INDIGO interviewed each student again to elicit their views and opinions on their overall team experience. The post-course interview included questions, such as *"What is your overall team working experience"* and *"What kind of suggestions do you have to improve the teaming experience"*.

4.3 Team Formation

To investigate whether INDIGO could offer useful insights for team formation, we wanted to learn the effect of existing team formation tools. Thus, the course instructor used CATME, an existing team

formation tool [10], to help team assignments. In this tool, the instructor first selected a set of criteria, such as team skills, working style, and demographics. By these criteria, the students were asked to take a series of surveys to obtain their self-reported measures. Based on the completed surveys, the instructor then set the weight for each criterion and ran the algorithm to obtain team assignments.

In our study, the instructor weighted the following criteria the most: languages skills, skill set (e.g., programming, UI design, and teamwork), programming capability, leadership preference (e.g., single leader vs. shared leadership), leadership role (follower or leader), and thinking style (e.g., big picture vs. detail-oriented). The algorithm was also configured to make the teams as diverse as possible by the weighted criteria.

Although all students in the course were told the teams were selected by an algorithm, only half of the teams (N=19) were assigned by the algorithm and another half (N=21) were assigned randomly. A total of 40 teams were formed with 4-5 students in each team.

4.4 Team Personality Composition

Existing studies show that team personality composition influences team dynamics and team performance [5, 18, 20, 31, 48]. To collect a team's personality composition, we used the students' Big 5 personality traits inferred by INDIDO during their interview.

4.5 Team Perception Survey

In addition to team personality composition, team dynamics is known to affect team performance [4, 9, 22, 43]. At the end of the course, we used a survey provided by the CATME tool to gather students' perception of their own team. As we will describe below, this survey allowed us to characterize a team using a scale with seven dimensions, each of which consists of a set of statements on a 1-5 Likert scale (1 totally disagree to 5 totally agree):

- *Psychological Safety.* This is a 7-item survey to measure how safe a person feels to take interpersonal risks in a team. It is known to affect team performance [13].
- *Interpersonal Cohesiveness.* This 3-item survey measures the interpersonal relationship between team members. In particular, it measures interpersonal attraction perceived by team members toward each other [52].
- *Task Commitment.* This dimension uses 3 items to measure the level of team member's commitment toward the group goal and how much effort was made by each teammate.
- *Task Attraction.* This dimension also includes 3 items to gauge the overall working atmosphere in a group. For example, it assesses how much the members enjoyed the group activities or the group's work as a whole.
- *Relationship Conflict.* This 3-item survey measures the level of tension perceived in the work group and the frequency of negative emotion generated during the working process.
- *Group Task Conflict.* This dimension uses 3 items to assess the conflicts among team members in terms of the group task, specifically, the frequency of conflicting opinions in the working process.
- *Process Conflict.* This 2-item dimension measures the conflicts during work distribution and resource allocation, e.g., how much conflict was there on task responsibilities.

5 RESULTS

To evaluate INDIGO and answer our research questions, we have examined multiple sources of data from INDIGO's field deployment, including the chat transcripts between each student and INDIGO and INDIGO-inferred personality traits of the students. Here we report the findings to answer our three research questions, respectively: (a) users' interaction with INDIGO, (b) the effectiveness of INDIGO in eliciting information from the students, and (c) the effect of INDIGO-derived personality traits on team performance and team perception.

5.1 RQ1: Students' Interaction with INDIGO

We first examined students' engagement with INDIGO. Our results showed that on average each student spent about 60 minutes (SD = 26 minutes) with INDIGO in their pre-course interview and 26 minutes (SD = 7 minutes) in the post-course interview. In addition to *engagement duration*, we also computed the *response length*, defined by the number of words in each student's responses, to gauge the amount of information that the students were willing to provide during their interaction with INDIGO. On average, each student provided 620 words (SD = 291 words) in their pre-course interview and 289 words (SD = 184 words) in their post-course interview. These measurements demonstrated that the students were willing to spend a considerable amount of time with INDIGO and offer information during their interaction with INDIGO.

5.1.1 Perceived Characteristics of INDIGO. We examined students' impression of INDIGO by examining their description and ratings of INDIGO.

The students were asked to describe their impression of INDIGO in three key words. The top-5 most mentioned keywords were *friendly*, *robotic*, *kind*, *nice*, and *polite*. From these words, it seemed that the students perceived their interaction with INDIGO positively. In fact, 80% of students provided *all* positive expressions when describing their impression of INDIGO, such as *"agreeable, friendly, perceptive"*; *"nice , charming , funny"*; and *"sweet, smart, and well built"*

The students were also asked to rate INDIGO on three dimensions, *likeable*, *helpful*, *enjoyable*, on a scale of 1 to 5, 1 being not at all and 5 being very much. The average rating for each dimension was: *likeable* 3.14 (SD = 1.29), *helpfulness* 3.12 (SD = 1.26), *enjoyable* 2.53 (SD = 1.27). Overall, the students seemed ambivalent about their experience with INDIGO. In the hope of finding explanations, we examined the students' chat transcripts with INDIGO. The transcripts helped explain the ratings from a couple of angles. First, the chat transcripts revealed that INDIGO was limited at understanding a student's complex input, which certainly made the chat less enjoyable. For example, one student commented on:

> *"[INDIGO] doesn't understand context when giving responses"*

Another student also stated:

> *"[INDIGO is] not a huge conversationalist (not too much enthusiasm or talking outside of the script)"*

This was consistent with the students' description of INDIGO (e.g., "robotic"). Although we leveraged the best conversational

agent that is available to us, INDIGO still has much to improve especially its ability to interpret a user's complex and diverse input.

Additionally, the interviews with INDIGO especially the pre-course interview was long (e.g., 60 minutes), which might have made the experience less enjoyable. However, considering people's tolerance with traditional surveys [15], INDIGO's engagement duration with the students is quite remarkable.

Despite their ambivalence about INDIGO, it is encouraging to observe that the students were still willing to interact with INDIGO and offer rich information.

5.1.2 Perceived Role of INDIGO. One of our goals was to investigate whether a conversational agent like INDIGO could serve as a team companion. To find out in which role INDIGO could serve a team, we asked students' perceived role of INDIGO by rating INDIGO on two roles: *like a friend* and *like a counselor* on a scale 1-5. Students perceived INDIGO more like a counselor (M = 3.24, SD = 3.26) than a friend (M = 2.52, SD = 1.28). This might be another reason why students felt the conversation was less enjoyable as it was not like chatting with a friend.

Moreover, when the students were asked to rate how much they trusted INDIGO on a 5-point likert scale, they indicated that they somewhat trusted INDIGO: M=3.49, SD=0.99. To understand the students' trust in INDIGO, we further examined the chat transcripts and found that the students were indeed quite open and honest at offering their opinions to INDIGO. For example, when asked about their weaknesses, one student stated:

> *"I suck at comprehending things... I feel like I'm pretty slow. It takes me a while to grasps concepts and that along with my slight laziness doesn't make for the best combo."*

Similarly, another answered:

> *"I need someone to guide me , in other words , it's hard for me to start one thing without any guidance."*

Considering that these conversations occurred *before* they had found their teammates for their class project, many students seemed having provided honest answers and did not try to hide their weaknesses.

Similarly, when asked what kind of role they want to play in a team, the students were honest to state the role they preferred to play. For example, one student stated:

> *"I would prefer to not be a leader , esp . not this semester because I have a lot of other things going on."*

Likewise, another student mentioned:

> *"I don't want to be a leader because I have too many assignments to work on."*

Similar to other findings [27], students' perception and their behavior with INDIGO showed that they somewhat trusted an agent like INDIGO and were willing to disclose personal information during the interaction. Although in our study, we did not intentionally frame INDIGO as a counselor, many students considered it as one, which might have also encouraged them to open up and offer truthful information. Understanding the students' perception of the role of INDIGO is important especially if we wish to use INDIGO as a team companion, which must be effective at eliciting authentic team information to understand the true team dynamics.

Impression of INDIGO	*SimilarPersonality*
Enjoyable	0.56, p<0.05*
Likeable	0.49, p<0.05*
Helpful	0.50, p<0.05*
Trust	0.36, p<0.05*

Table 2: The correlation between students' perceived similarity about INDIGO's personality traits to their own personality and students' impression of INDIGO.

5.1.3 Perceived Relation with INDIGO. Previous research shows that users enjoyed their interaction more with an agent if they perceive the agent has a personality similar to theirs [33]. The students were asked to rate how similar they were to the personality of INDIGO (*SimilarPersonality*). We then examined the correlation between the *SimilarPersonality* rating (5-point Likert Scale) and all other user ratings (5-point Likert Scales), such as *likeable* and *enjoyable*. The analysis revealed a moderate correlation with all these ratings (Table 2), which suggests the potential to adapt INDIGO's personality to that of a user to improve user experience with INDIGO.

5.2 RQ2: Effectiveness of INDIGO in Information Gathering

One of the main purposes of using INDIGO is to provide a more engaging way to gather information from the students, compared to the traditional, static online surveys. We thus evaluated INDIGO on its effectiveness of gathering information from the perspective of the instructor. To form teams and understand team dynamics, instructors often use traditional online surveys to learn about the students and their team experience. Compared to these surveys, the conversational interview conducted by INDIGO used more open-ended questions. While open-ended questions help elicit richer information and provide rationale behind quantitative ratings, research shows that collecting responses to open-ended questions is often difficult [35]. We thus conducted a series of analysis to examine whether the open-ended questions posed by INDIGO helped elicit useful information that can benefit the course instructor.

We first compared student's responses to the question, *"What is your preferred role in a team"* which was asked in both the CATME survey and INDIGO's interview before the teams were formed. The CATME survey used a choice-based question with five options from *strongly prefer to be a follower* to *strongly prefer to be a leader*. In contrast, INDIGO posed it as an open-ended question. To capture the gist of student responses to this question, we used an enhanced Latent Dirichlet Allocation (LDA) model [6] to analyze the 184 responses and automatically derive a set of semantic themes covered by the responses. We also used LexRank [14] to find representative sentences within each theme.

The themes produced by the LDA model not only covered *all* options presented in the choice-based question but also gave additional information, such as how they wish to play a role and the rationale why they wanted to serve a particular role, for example, one student mentioned

> *"I prefer to take turns leading and following."*

It would be difficult to put this student's answer into a category yet the information is valuable. Similarly, another student stated his preferred role on a team:

> *"I prefer to be the coordinator in the team. I would like to collect different ideas from team members, do some conclusion, share ideas among different groups and ask for advices from professor and TAs."*

Not only was additional information collected, but the instructor could also better understand the *"why"* behind the students' input. For example, at the end of the course, the students were asked to rate their overall team experience on a scale of 1 to 5, where 1 being poor and 5 being excellent. INDIGO also conducted a post-course interview that asked each student how they felt about their team experience. Coupling the students' responses to INDIGO with their ratings, the instructor got a more comprehensive picture on how the teams worked together. For example, a student who gave a high (5) rating wrote to INDIGO,

> *"... all of us were very supportive of each other and we split up the work evenly"*

In contrast, a student who gave a low (1) rating mentioned,

> *"It was a little tiring when others wanted to leave off the work until the last second"*

From the above examples, INDIGO was able to elicit useful information from the students. One might argue that the open-ended questions could be inserted into a regular survey to collect the needed information. However, extensive survey statistics shows that people are willing to spend only a few seconds per question on a survey that lasts more than 5 minutes [15]. Our use of INDIGO indicated that the students were willing to spend time interacting with it and offer useful information, which suggests an effective way to collect information from students.

5.3 RQ3: Effect of INDIGO's Personality Insights on Team Results

To examine whether and how INDIGO could help team formation, we looked into its inferred student personality traits and investigated the effect of team personality composition on team results. Specifically, we wanted to answer two questions:

- **RQ3a:** How does team personality composition impact student team performance?
- **RQ3b:** How does team personality composition impact team members' perceptions of their own team?

We used the data from three sources: (a) 184 students' 35 Big 5 personality traits inferred by INDIGO, (b) 184 students' self-reported team perception by seven dimensions, and (c) 40 teams' project performance. Because of the number of data dimensions involved, we first performed factor analyses to examine the relationships among the relevant data dimensions.

5.3.1 Factor Analysis of Inferred Personality Traits. We first examined the factorability of the inferred 35 personality traits. The results showed that 33 of 35 measures correlated with at least one other measure ($R^2 \geq 0.3$). Moreover, the Bartlett's test of sphericity was significant ($\chi^2(595) = 35091.39$, $p < .05$). Thus factor analysis was suitable for all 35 traits. A Principal Components Analysis (PCA) indicated a seven-factor solution, which explained 53% of the variance. A screen plot also showed the sharp leveling off of Eigenvalues after the seven factors. Table 3 lists the seven factors. For each of the seven factors, we then measured a composite trait score, a regression-weighted mean of items with primary loadings greater than 0.5 in the factor. Internal consistency for each score was also examined using Cronbach's alpha. The alphas were: 0.78 for *Emotional* (4 items), 0.63 for *Collaborative* (3 items), 0.53 for *Social* (3 items), 0.55 for *Open-minded* (3 items), 0.43 for *Responsible* (3 items), 0.73 for *Sensitive* (3 items), and 0.39 for *Self-Disciplined* (2 items). Overall, our analyses indicated that seven distinct factors were underlying the students' personality measures with reasonable internal consistency. Only one item had a cross-loading above 0.5 (*Vulnerability*), however, this item had a strong primary loading of 0.64.

Factors	Personality Traits
Emotional	**Neuroticism**, Depression, Impulsiveness, Vulnerability
Collaborative	**Agreeableness**, Cooperation, Sympathy
Social	**Extroversion**, Friendliness, Gregariousness
Open-minded	**Openness**, Imagination, Intellectual Curiosity
Responsible	**Conscientiousness**, Dutifulness, Cautiousness
Sensitive	Feelings
Self-Disciplined	Self-Discipline, Anxiety, Vulnerability

Table 3: Personality traits loaded onto 7 separate factors. The bold trait indicates the trait is one of the Big five personality traits

	Team Relationship	Team Conflict
Psychology Safety	0.55	
Interpersonal Cohesiveness	0.91	
Task Commitment	0.58	
Task Attraction	0.85	
Relationship Conflict		0.73
Task Conflict		0.56
Process Conflict		0.76

Table 4: Factor Loadings for 2 Factors from the Student's Team Reflection

5.3.2 Factor Analysis on Team Perception Measures . We also performed factor analysis on the seven dimensions that measured students' preception of their own team. The Bartlett's test of sphericity was significant ($\chi^2(21) = 119.74$, $p < .05$). A two-factor solution was derived from PCA, which explained 59% of the variances. Accordingly, we created two composite scores for both factors, respectively. Table 4 showed these two factors. Each composite score was regression-weighted on the relevant items with their primary loadings greater than 0.5. Cronbach's alphas were also computed: 0.84 for *Team Cohesion* (4 items) and 0.69 for *Team Conflict* (3 items). And no cross loading was found.

5.3.3 Analysis Variables. To answer our questions above, we computed a set of measures as independent and dependent variables, respectively.

5.3.4 Independent Variables.

- *Team Personality Composition.* Numerous studies show that both mean and variance of team personality composition scores influence team performance [18, 20, 31]. For each team, we measured its personality composition by the mean and variance of the individual team members' personality scores. Each individual's personality scores were computed by the seven extracted personality factors (Table 3).
- *Team Formation Method.* A binary variable indicates whether a team was determined by an algorithm (value = 1) or randomly assigned (value = 0).
- *Impression Management Score (IM Score).* The IM score reflects how students consciously favor themselves to impress others. For each team, the IM Score is a mean of its team member's individual IM scores.

5.3.5 Dependent Variables.

- *Team Performance.* For each team, its performance was based on the team's final project score, which included the scores of all project milestones throughout the course.
- *Team Perception.* For each team, we computed the mean of each of the two factors, Team Relationship and Team Conflict, extracted based on the ratings reported by each team member in their survey.

5.3.6 Analysis Methods.

Before examining the effect of the above independent variables on each of the dependent variables, we first analyzed the relationship between the dependent variables. A correlation test showed no significant correlation between *Team Relationship* and *Team Performance*: r(38) = -0.05, p = 0.76 and neither between *Team Conflict* and *Team Performance*: r(38) = 0.29, p = 0.07. The results indicates our dependent variables measures different aspects of the teams.

5.3.7 ***RQ3a: Effect of Team Personality Composition on Team Performance.***

Using *Team Personality Composition* as an independent variable and *Team Performance* as a dependent variable, we built a regression model with Impression Management (IM) scores and *Team Formation Method* as control variables.

The analysis results showed that the *Emotional* variance in team personality composition (β= .43, t(30) = 2.30, p < .05) and the level of Impression Management (β=.37, t(30)=2.22, p<0.05) significantly predicted team performance (Table 5). The *Emotional* variance in team personality composition alone explained a significant proportion of variance in team performance, R2 = .34, p < .05. The level of IM alone explained a significant of the variance as well, R2=.36, p<.05. No other effect was significant.

In other words, the more diverse a team was in their Emotional makeup, the better their project performance was. Although no prior study on teaming reports such a finding, this result is consistent with prior findings in personality research. In particular, the *Emotional* factor consisted of 4 items on the Neuroticism dimension (Table 3). According to Oertig et al. [34], people high in Neuroticism tend to perform better in short-term goals with deadlines, which is very similar to our class project setting. On the other hand, people low in Neuroticism handle stressors (e.g. upcoming deadlines) better, which in turn helps overall team effort [12]. In summary, a team with members at varied levels of Neuroticism could benefit from both sides to perform better.

Predictor	*B*	β	*sr^2*	*R2*
(Intercept)	91.05**			
Emotional	.07*	.43	.12	.34*
Collaborative	-.03	-.12	-.01	-.05
Social	-.04	-.16	.02	.01
Open-minded	.01	.04	.00	.03
Responsible	-.05	-.22	.05	-.13
Sensitive	-.01	-.02	.00	.20
Self-Disciplined	-.06	-.23	.05	-.10
Impression Management	.56*	.37	.11	.36*
Team-formation Method	-.05	-.01	.00	.10

Table 5: Regression results using Team Performance as the outcome variable and the variance of the individual team members personality scores as the independent variables. *Note. * indicates p < .05. ** indicates p < .01.*

Moreover, not only is our unique finding derived from a field study, but it also helps verify several previous teaming theories. For example, Lykourentzou shows a balanced personality composition benefit team performance [31] and how emotional traits affect teamwork [12, 31].

To better illustrate our finding, we chose and compared the characteristics of one team that had a diverse Emotional makeup and achieved a high team performance (Team A) with another team that had a more homogeneous Emotional makeup and achieved a low team performance (Team B) (see Table 6). The compositions of these two teams were very similar in terms of the skills, leadership preference, leadership role preference, and thinking style. Those criteria were the most weighted criteria set by the instructor during the team formation, which are also commonly used by other project-based classes [21]. While these characteristics are very similar, the variances of team *Emotional* composition were very different (Var_{TeamA} = 713, Var_{TeamB} = 219.7).

To understand how the team's *Emotional* makeup impacted team performance, we interviewed the teaching assistant (TA) who mentored both teams. The TA mentioned that "Team A was often concerned with their grade and kept sending me emails to ask what's the requirement and when the next assignment will be due." It is also interesting to observe that although instructors often intuitively believe that general academic performance (e.g., GPA) of students is most predictive of team performance, we can see from Table 6 that members of Team A (M = 3.28) actually had a *lower* mean GPA than that of those in Team B (M = 3.70). Also, one student on team A who had a relatively high emotional score also had a low GPA. Based on the TA's comments, one explanation why Team A performed better was that the members with a high *Emotional* score tended to remind the team about upcoming deadlines and nudge the team to finish the work on time. On the other hand, emotionally more calm members helped hold the team together without being overwhelmed by the deadlines. It is possible that this was the reason that Team A achieved a much higher overall project score than Team B (98.7% vs 90.4%). As an example, the analyses afforded by

	Team A ($Var_{Emotional}$ = 713)				Team B ($Var_{Emotional}$ = 219.7)				
Team Member	P1	P2	P3	P4	P1	P2	P3	P4	P5
Emotional	232	222	214	274	258	225	236	247	223
GPA	3.92	3.4	3.5	2.30	3.9	3.96	3.93	3.0	3.72
Skill Set	T,W,P	T,P	P	T,U,P	T,W,U,P	P	T,W	T,P	T,W,P
Leadership Preference	O	O	O	S	O	O	O	S	S
Leadership Role	N	F	F	N	F	F	F	N	N
Thinking Style	I	B	B	I	B	B	B	I	I

Note. T indicates Team-work. P indicates Programming. W indicates Writing Skill. D indicates Design Skill. O indicates One Leader with Input. S indicates Shared Leadership. F indicates Follower. B indicates No Preference between Follower and Leader. I indicates Idea Oriented. B indicates Balanced between Idea Oriented and Detail Oriented.

Table 6: Team Characteristics Comparison Between Team A and Team B

Predictor	*B*	*β*	*sr^2*	*R2*
(Intercept)	3.30			
Emotional	.01	.23	.03	-.09
Collaborative	.01	.25	.08	-.36*
Social	-.00	-.07	.01	-.12
Open-minded	.01	.28	.03	.26
Responsible	.00	-.02	.00	-.01
Sensitive	-.02*	-.42	.10	.23
Self-Disciplined	-.01	-.14	.02	-.13
Impression Management	.02	.08	.01	-.01
Team-formation Method	-.27*	-.33	.10	.35*

Table 7: Regression results using Team Conflict as the outcome variable and the mean of the individual team members' personality scores as the independent variables. *Note.* * *indicates p < .05.*

INDIGO led us to focus on important aspects of the teams, which provided important insights on how team composition could plays a pivotal role in influencing team performance.

Our finding above suggests that emotional makeup of a team impact team performance. To create teams with high performance especially for accomplishing short-term goals, INDIGO could be used to first understand individuals' personality traits and then suggest teams that are made up of members with varied emotional characteristics.

As we hypothesized, our analysis indicated that the Impression Management Scale (IM) also influenced team performance. In particular, the higher the average IM score was in a team, the better the team performed. Our finding seems consistent with previous research on relating IM scale with self-control and social adaptation [46]. In other words, a team would perform better, if all team members have a high level of self-control and can adapt well socially in a team setting. In addition, we further examined whether the inferred personality traits could predict the IM scores. We found that among the seven factors, *Sensitive* (β=0.25, p=0.01**), *Emotional* (marginal, β = -0.19,p = 0.05), and *Social* (marginal, β = -0.15, p = 0.07). This suggests that personality traits may be used to infer IM scores *automatically*, which can then be used to suggest team formations (e.g., trying to form teams with a higher average IM scores). The significant relationship between team personality composition, impression management score, and team performance indicates the potential of an agent like INDIGO: it could be used to understand individuals by inferring their personality traits and then use the inferred traits to recommend high-performance teams.

5.3.8 ***RQ3b: Effect of Team Personality Composition on Team Perception.*** To answer our second question above, we built a regression model that used team personality composition as independent variables and team perception as dependent variables. Regression results showed that the mean of a team's *Sensitive* score significantly predicted *Team Conflict*, β = -.42, t(30) = -2.10, p<.05 7. In particular, the less sensitive a team was, the fewer conflicts a team experienced. Moreover, *Team Formation Method* played a role. Teams assigned by the tool reported fewer team conflicts, β = -0.33, t(30) = 2.24, p< .05. No other effect was significant,see Table 7.

Our results suggest that teams that were lower on the *Sensitive* measure experienced fewer conflicts during the teaming process. The *Sensitive* measure is highly loaded on *Feelings*, *Anxiety* and *Vulnerability*, which suggest that people high on these dimensions are more sensitive and vulnerable to conflicts and negative feelings [16]. When facing upcoming deadlines, those team members might have expressed more negatives emotions. Furthermore, their similarly vulnerable teammates could not cope with the negative feelings and might cast their own negative emotions. The intensified negative feelings would then create more team conflicts and affect their team relationship.

In addition, we found teams that were assigned by the team formation tool were more satisfied with their team. This finding was aligned with previous research that team formation algorithm could help improve team dynamics [21]. We further tested whether the team formation method interacted with the team personality composition. The result showed no significant effect.

Again, the relationship between students' inferred personality and team perception indicates INDIGO's potential to recommend teams based on team personality compositions that will optimize team experience.

6 DISCUSSION

The field deployment of INDIGO and its demonstrated value offered encouraging results. First, INDIGO elicited rich information from the students, which enabled the instructor to gain deeper insights

into the students as unique individuals along with their teaming preferences and experience. We believe that the novel use of a conversational agent like INDIGO contributes to the rich information harvested. Before INDIGO, traditional teaming tools require students to take multiple surveys before assigning teams. Survey fatigue may prevent students from giving truthful and in-depth information. On the other hand, the interactive nature of INDIGO may reduce survey fatigue.

Moreover, team personality composition that was inferred by INDIGO predicted team outcomes. Our result showed that teams with a higher variance in their *Emotional* makeup performed better and teams with lower on the Sensitive measure experienced fewer conflicts. Without INDIGO, such relationships were difficult for traditional tools to discover. Additionally, INDIGO saved the extra effort required for the students to assess their personality, not mentioning the objectivity in these results due to social desirability bias. The relationships discovered by INDIGO also provide guidance for instructors to choose the appropriate criteria to form effective teams.

6.1 Design Implications

The study findings demonstrated INDIGO's ability to proactively engage with students, gather useful teaming information, and provide insights on team outcomes through the lens of team personality composition. Such findings can benefit the design of intelligent teaming tools in general. First, from the interaction between INDIGO and the students in the class, we learned that the students trusted INDIGO and perceived INDIGO as a counselor. In the future design of an intelligent agent for teaming, we could leverage such perception by framing the agent as a team coach. Similar to a counselor in the real life, who would follow up with their clients, a team coach can follow up with individual team members throughout their teaming efforts. Such a team coach can collect students' team perception in real time and track the changes. Instructors and teaching assistants can then use the gathered information to intervene or guide group activities, such as helping reduce interpersonal tension or resolve interpersonal conflicts.

Second, our results show that students prefer INDIGO more if they perceive INDIGO having a personality similar to theirs. With its personality inference capability, an intelligent agent like INDIGO can learn a user's personality on the fly and then adapt its behavior to that of the similar personality (e.g., using similar wording).

Third, since our findings reveal that team personality composition predicts team outcomes, we can use such findings to augment INDIGO. In particular, we can extend INDIGO to automatically recommend team formation based on the inferred personality traits of potential team members. For example, it can select each team by maximizing the variance of emotional characteristics of the team members.

6.2 Limitations

Our current work has several limitations. First, the measure of team performance in our study was limited to one project score. The goal for education should go beyond a simple score. From example, the project score does not fully reflect the teamwork skills that students have learned in the process. Moreover, research suggests that a lower score sometimes may even imply better learning [38, 40, 42]. Second, our field deployment of INDIGO was situated in a Computer Science class in a large U.S. public university, which may not be representative of student teaming situations in other cultures, since culture often influences teaming and team success [28, 32]. Third, in our evaluation, we collected the self-reported team perception at the end of the semester. Such measures may not fully reflect a team's status, let alone capturing the changes of such status. Ideally, we want to collect team perceptions throughout teaming efforts and detect their changes over time to better assess team dynamics. Although the student teaming effort lasted for a full semester, it is still considered short-term teaming. It is unclear whether our findings would hold for longer teaming efforts, for example, sports teams or workplace teams that may last for many years.

7 CONCLUSIONS

We have presented the novel application of a conversational agent called INDIGO (Individual Differences for Group Optimization) in aiding teaming efforts. We evaluated INDIGO through a field deployment involving about 200 university students in 40 teams working on semester-long team projects. INDIGO interviewed each student twice to learn the student's team preferences and team experience. From the perspective of students and instructors, INDIGO demonstrates its ability to effectively collect valuable information from the students, which could help instructors form effective project teams. In addition, the use of INDIGO discovered the relationships between students' personality traits inferred from their interaction with INDIGO and their team outcomes. This demonstrates the potential of using INDIGO to recommend team formation for optimal teaming experience and outcomes. Overall, our findings bear design implications on developing intelligent agents for aiding teaming efforts at various stages, such as recommending team formation, tracking team dynamics, and guiding team behavior.

ACKNOWLEDGMENTS

This work is supported in part by the Air Force Office of Scientific Research under FA9550-15-C-0032. We would also like to thank all the reviewers for their valuable feedback.

REFERENCES

[1] Elizabeth F Barkley, K Patricia Cross, and Claire H Major. 2014. *Collaborative learning techniques: A handbook for college faculty.* John Wiley & Sons.
[2] Bruce Barry and Greg L Stewart. 1997. Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied psychology* 82, 1 (1997), 62.
[3] Daniel J Beal, Robin R Cohen, Michael J Burke, and Christy L McLendon. 2003. Cohesion and performance in groups: a meta-analytic clarification of construct relations. *Journal of applied psychology* 88, 6 (2003), 989.
[4] Julia B Bear and Anita Williams Woolley. 2011. The role of gender in team collaboration and performance. *Interdisciplinary science reviews* 36, 2 (2011), 146–153.
[5] Suzanne T Bell. 2007. Deep-level composition variables as predictors of team performance: a meta-analysis. *Journal of applied psychology* 92, 3 (2007), 595.
[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[7] Dan Bohus and Alexander I Rudnicky. 2009. The RavenClaw dialog management framework: Architecture and systems. *Computer Speech & Language* 23, 3 (2009), 332–361.
[8] Lt Col James L Brickell, Lt Col David B Porter, Lt Col Michael F Reynolds, and Capt Richard D Cosgrove. 1994. Assigning students to groups for engineering design projects: A comparison of five methods. *Journal of Engineering Education* 83, 3 (1994), 259–262.

[9] Mary Ann Burress. 1996. *The relationship between team leader behaviors and team performance and satisfaction.* Ph.D. Dissertation. University of North Texas.

[10] CATEME. 2018. CATME. (2018). http://info.catme.org/

[11] Elena del Val, Juan Miguel Alberola, Victor Sanchez-Anguix, Alberto Palomares, and Ma Dolores Teruel. 2014. A team formation tool for educational environments. In *Trends in Practical Applications of Heterogeneous Multi-agent Systems. The PAAMS Collection*. Springer, 173–181.

[12] Vanessa Urch Druskat and Steven B Wolff. 2001. Building the emotional intelligence of groups. *Harvard business review* 79, 3 (2001), 80–91.

[13] Amy C Edmondson and Zhike Lei. 2014. Psychological safety: The history, renaissance, and future of an interpersonal construct. *Annu. Rev. Organ. Psychol. Organ. Behav.* 1, 1 (2014), 23–43.

[14] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.

[15] Colette Des Georges and Eric Van Susteren. 2018. 5 ways to get the survey data you want. (2018). https://www.surveymonkey.com/curiosity/5-best-ways-to-get-survey-data/

[16] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[17] Liang Gou, Michelle X Zhou, and Huahai Yang. 2014. KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 955–964.

[18] Terry Halfhill, Tjai M Nielsen, Eric Sundstrom, and Adam Weilbaecher. 2005. Group personality composition and performance in military service teams. *Military Psychology* 17, 1 (2005), 41.

[19] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 415.

[20] Stephen E Humphrey, John R Hollenbeck, Christopher J Meyer, and Daniel R Ilgen. 2007. Trait configurations in self-managed teams: A conceptual examination of the use of seeding for maximizing and minimizing trait variance in teams. *Journal of Applied Psychology* 92, 3 (2007), 885.

[21] Farnaz Jahanbakhsh, Wai-Tat Fu, Karrie Karahalios, Darko Marinov, and Brian Bailey. 2017. You Want Me to Work with Who?: Stakeholder Perceptions of Automated Team Formation in Project-based Courses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3201–3212.

[22] Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. 1999. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly* 44, 4 (1999), 741–763.

[23] David W Johnson, Roger T Johnson, and Karl Smith. 2007. The state of cooperative learning in postsecondary and professional settings. *Educational Psychology Review* 19, 1 (2007), 15–29.

[24] Peter J Jordan and Neal M Ashkanasy. 2006. Emotional Intelligence, Emotional Self-Awareness, and Team Effectiveness. (2006).

[25] JS Karn and Anthony J Cowling. 2005. A study of the effect of disruptions on the performance of software engineering teams. In *Empirical Software Engineering, 2005. 2005 International Symposium on*. IEEE, 9–pp.

[26] Richard A Layton, Misty L Loughry, Matthew W Ohland, and George D Ricco. 2010. Design and Validation of a Web-Based System for Assigning Members to Teams Using Instructor-Specified Criteria. *Advances in Engineering Education* 2, 1 (2010), n1.

[27] Jingyi Li, Michelle X Zhou, Huahai Yang, and Gloria Mark. 2017. Confiding in and Listening to Virtual Agents: The Effect of Personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 275–286.

[28] Xuan-Hui Lin, Ran Bian, Rui Zhu, and Hong-Sheng Che. 2008. Team personality composition and team effectiveness: The mediating effects of team process. *Acta Psychologica Sinica* 40, 4 (2008), 437–447.

[29] Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse* 8, 1 (2017), 31–65.

[30] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[31] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P Dow. 2016. Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 260–273.

[32] Erin Meyer. 2014. *The culture map: Breaking through the invisible boundaries of global business.* PublicAffairs.

[33] Clifford Nass, Youngme Moon, Brian J Fogg, Byron Reeves, and Chris Dryer. 1995. Can computer personalities be human personalities?. In *Conference companion on Human factors in computing systems*. ACM, 228–229.

[34] Daniela Oertig, Julia Schüler, Veronika Brandstätter, and Adam A Augustine. 2014. The Influence of Avoidance Temperament and Avoidance-Based Achievement Goals on Flow. *Journal of personality* 82, 3 (2014), 171–181.

[35] Marije Oudejans and Leah Melani Christian. 2010. Using interactive features to motivate and probe responses to open-ended questions. *Social and behavioral research and the Internet* (2010), 215–244.

[36] Ioannis Papaioannou, Christian Dondrup, Jekaterina Novikova, and Oliver Lemon. 2017. Hybrid chat and task dialogue for more engaging hri using reinforcement learning. In *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on*. IEEE, 593–598.

[37] Delroy L Paulhus. 1991. Measurement and control of response bias. (1991).

[38] Maria Pennock-Roman. 1992. Interpreting test performance in selective admissions for Hispanic students. (1992).

[39] Marilyn Ann Perkins. 1991. Team Member Implicit Theories of Team Development: Relationships with Team Member Behavior, Team Viability and Team Performance. (1991).

[40] Neda Ratanawongsa, Patricia A Thomas, Spyridon S Marinopoulos, Todd Dorman, Lisa M Wilson, Bimal H Ashar, Jeffrey L Magaziner, Redonda G Miller, Gregory P Prokopowicz, Rehan Qayyum, et al. 2008. The reported validity and reliability of methods for evaluating continuing medical education: a systematic review. *Academic Medicine* 83, 3 (2008), 274–283.

[41] Rebecca H Rutherfoord. 2001. Using personality inventories to help form teams for software engineering class projects. In *ACM Sigcse Bulletin*, Vol. 33. ACM, 73–76.

[42] Winny Shen, Paul R Sackett, Nathan R Kuncel, Adam S Beatty, Jana L Rigdon, and Thomas B Kiger. 2012. All validities are not created equal: Determinants of variation in SAT validity across schools. *Applied Measurement in Education* 25, 3 (2012), 197–219.

[43] Greg L Stewart. 2006. A meta-analytic review of relationships between team design features and team performance. *Journal of management* 32, 1 (2006), 29–55.

[44] Joel Anthony Thurston. 2012. *Exploring Group Perception: The Relationship Between the Perception of Entitativity and Assessments of Cohesion.* University of California, Santa Barbara.

[45] David Traum. 2017. Computational Approaches to Dialogue. *The Routledge Handbook of Language and Dialogue* (2017), 143.

[46] Liad Uziel. 2014. Impression management ("lie") scales are associated with interpersonally oriented self-control, not other-deception. *Journal of personality* 82, 3 (2014), 200–212.

[47] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. 2018. On Evaluating and Comparing Conversational Agents. *arXiv preprint arXiv:1801.03625* (2018).

[48] Thomas J Whelan, L Aiman-Smith, C Kimbrough, and Larry Taylor. 2009. Group personality composition, satisfaction and performance in virtual teams. In *24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA*.

[49] Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274* (2017).

[50] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3506–3510.

[51] Zhou Yu, Alan W. Black, and Alexander I. Rudnicky. 2017. Learning Conversational Systems That Interleave Task and Non-task Content. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. AAAI Press, 4214–4220. http://dl.acm.org/citation.cfm?id=3171837.3171875

[52] Stephen J Zaccaro and M Catherine McCoy. 1988. The effects of task and interpersonal cohesiveness on performance of a disjunctive group task. *Journal of applied social psychology* 18, 10 (1988), 837–851.

[53] Michelle X Zhou, Keith Houck, Shimei Pan, James Shaw, Vikram Aggarwal, and Zhen Wen. 2006. Enabling context-sensitive information seeking. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, 116–123.

[54] Matthias Ziegler, Carolyn MacCann, and Richard Roberts. 2011. *New perspectives on faking in personality assessment.* Oxford University Press.

[55] Victor Zue, Stephanie Seneff, Joseph Polifroni, Michael Phillips, Christine Pao, David Goodine, David Goddeau, and James Glass. 1994. PEGASUS: A spoken dialogue interface for on-line air travel planning. *Speech Communication* 15, 3-4 (1994), 331–340.