# Understanding Women's Remote Collaborative Programming Experiences: The Relationship between Dialogue Features and Reported Perceptions

KIMBERLY MICHELLE YING, University of Florida, USA
FERNANDO J. RODRÍGUEZ, University of Florida, USA
ALEXANDRA LAUREN DIBBLE, University of Florida, USA
KRISTY ELIZABETH BOYER, University of Florida, USA

In recent years, remote collaboration has become increasingly common both in the workplace and in the classroom. It is imperative that we understand and support remote collaborative problem solving, particularly understanding the experiences of people from historically marginalized groups whose intellectual contributions are essential for addressing the pressing needs society faces. This paper reports on a study in which 58 introductory computer science students constructed code remotely with a partner following either predefined structured roles (*driver* and *navigator* in pair programming) or without predefined structured roles. Between the structured-role and unstructured-role conditions, participants' normalized learning gain, Intrinsic Motivation Inventory scores, and system usability scores were not significantly different. However, regardless of the collaboration condition, women reported significantly higher levels of stress, lower levels of perceived competence, and less perceived choice compared to men. Because computer science is a context in which women have been historically marginalized, we next examined the relationship between student gender and collaborative dialogues by extracting lexical and sentiment features from the textual messages partners exchanged. Results reveal that dialogue features, such as number of utterances, utterance length, and partner sentiment, significantly correlated with women's reports of stress, perceived competence, or perceived choice. These findings provide insight on women's experiences in remote programming, suggest that dialogue features can predict their collaborative experiences, and hold implications for designing systems that help provide collaborative experiences in which everyone can thrive.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; • **Social and professional topics → Women**.

Additional Key Words and Phrases: Collaborative Problem Solving; Remote Collaboration; Collaborative Programming; Gender Differences

**ACM Reference Format:**
Kimberly Michelle Ying, Fernando J. Rodríguez, Alexandra Lauren Dibble, and Kristy Elizabeth Boyer. 2020. Understanding Women's Remote Collaborative Programming Experiences: The Relationship between Dialogue Features and Reported Perceptions. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 253 (December 2020), 29 pages. https://doi.org/10.1145/3432952

Authors' addresses: Kimberly Michelle Ying, kimying@ufl.edu, University of Florida, 432 Newell Dr, Gainesville, Florida, 32611, USA; Fernando J. Rodríguez, fjrodriguez@ufl.edu, University of Florida, 432 Newell Dr, Gainesville, Florida, 32611, USA; Alexandra Lauren Dibble, a.dibble@ufl.edu, University of Florida, 432 Newell Dr, Gainesville, Florida, 32611, USA; Kristy Elizabeth Boyer, keboyer@ufl.edu, University of Florida, 432 Newell Dr, Gainesville, Florida, 32611, USA.


2573-0142/2020/12-ART253 $15.00
https://doi.org/10.1145/3432952

## 1 INTRODUCTION

In today's global society, remote collaboration has become ever present in occupational and educational interactions. Collaborative problem solving has long been identified as an essential 21st century skill [13] and has become widespread in many educational contexts to prepare learners for their future careers. Many education standards now emphasize collaboration and problem-solving skills [2]. Furthermore, a rich body of literature has established that collaborative learning, in both co-located and remote contexts, results in greater productivity and higher achievement, more supportive and committed relationships, greater self-esteem and social competence, and better mental health [32, 43].

Collaborative paradigms in a remote context are inherently different from face-to-face collaborations. CSCW researchers are increasingly investigating remote collaborative work [43, 60] and remote collaborative problem solving [52]. It is imperative that we come to better understand people's experiences and perceptions during remote collaboration as we move toward creating socio-technical systems that foster positive experiences and relationships during collaborative problem solving. This imperative may be particularly crucial in fields such as computer science, where certain groups of people (e.g., women and people of color) have been historically marginalized [47, 56]. For example, women continue to make up only a small fraction of post-secondary computer science students and earned only 20.9% of computer science bachelor's degrees in 2018 at doctoral institutions in the U.S. and Canada [70]. The intellectual contributions of people from diverse backgrounds are essential to address the substantial needs that society faces today. Thus, the differences in remote collaborative experiences with respect to gender are important to consider [31].

In computer science workplaces and classes, *pair programming* is a widely used paradigm in which two collaborators synchronously work on a shared programming task. At any given time, the *driver* writes code while the *navigator* provides feedback on the driver's actions and ideally helps with broader strategy and correctness. Within this structure, collaborators switch roles at certain time intervals or after subtasks are complete. On the other hand, collaborative programming can proceed without structured roles, with collaborators freely designating responsibilities. The structured roles of pair programming have proven beneficial for professionals as well as students [34, 42, 63]; however, research on pair programming has generally been conducted with co-located collaborators [8, 9, 34, 42]. There is a research gap concerning the impact of structured roles on collaborative programming in remote contexts. In a step toward understanding this issue, the present study compares structured and unstructured roles in remote collaborative programming with university students.

Despite the many established benefits of collaborative problem solving, previous work holds evidence of tension due to factors such as differences in prior knowledge [61], personality types [29], or degree of mutual understanding [20]. While some researchers have focused on finding optimal team formations (e.g., [17], [26]), our work embraces gender diversity in computer science team composition and examines the ways in which dialogue fosters positive outcomes for those collaborators. Our work focuses on computer science education, a context in which women in the U.S. and many other countries have been historically marginalized. Collaborative programming experiences can significantly impact these students' performance and perceptions by fostering productive communication between partners [8, 68].

This article addresses a pressing open issue within the CSCW community: the need to understand people's experiences during remote collaborative problem solving and to facilitate successful interactions. Specifically, we report on a study of remote collaborative programming through the Floobits plugin for IntelliJ, a synchronized coding interface with textual chat (see Figure 1). This

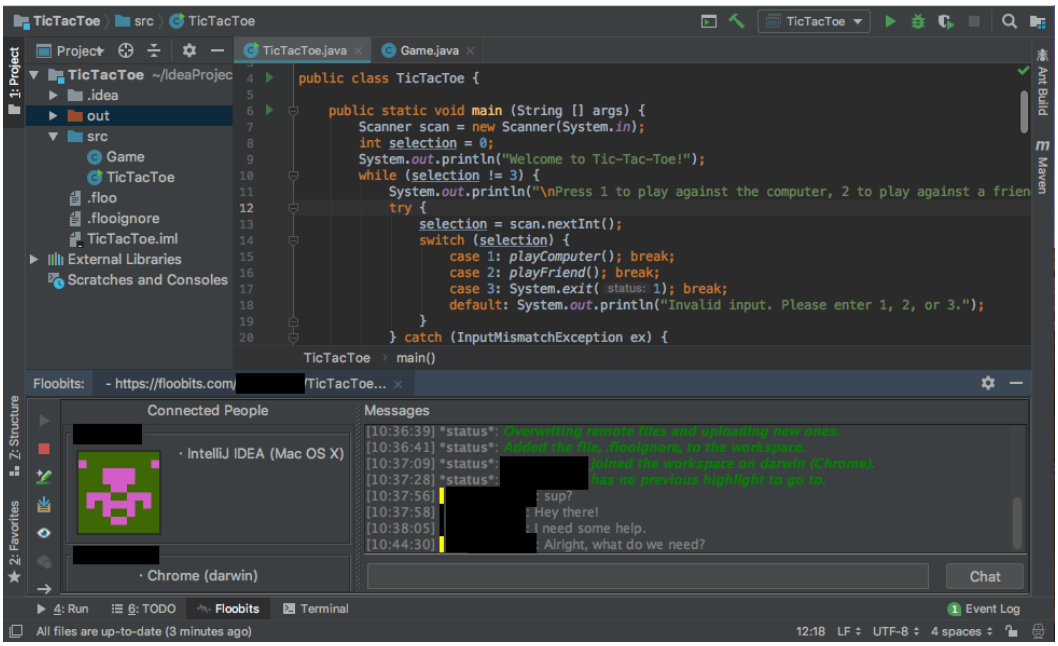


Fig. 1. Screenshot of IntelliJ IDEA and Floobits plugin interface.

work investigates the following research questions: (1) *How does the presence of structured roles in a remote programming context influence people's experiences and perceptions of the collaboration?*; (2) *In what ways do people's experiences and perceptions of the collaboration differ by gender?*; and (3) *What lexical and sentiment features of the dialogue are associated with these outcomes?* Participants from this study were recruited from an introductory computer science course at the University of Florida in the U.S., and were mainly White and between the ages of 18 and 21. To answer the first research question, we assigned participants to work with a partner with either predefined structured roles of driver/navigator (16 pairs) or without predefined structured roles (13 pairs) and analyzed their responses to an intrinsic motivation survey measuring interest/enjoyment, perceived competence, effort/importance, pressure/tension, perceived choice, value/usefulness, and relatedness [27, 40]. The results showed no significant differences between the unstructured-role versus structured-role conditions for outcomes of intrinsic motivation.

To answer the second research question, we focused our analysis on whether men's and women's experiences and perceptions of these collaborations differed. Regardless of their collaboration condition, women reported significantly lower perceived competence and perceived choice on the activity and higher levels of stress compared to men. To answer the third research question and gain insight into the collaborative processes associated with these differences, we explored lexical features, such as number of messages and message length, as well as sentiment (the positive or negative feelings conveyed through utterances), with the goal of finding correlations with women's self-reported stress, perceived competence, and perceived choice. Ordinal regression models of these survey item responses revealed, for example, that women tended to feel more relaxed if their partner sent longer messages on average or used more positive language. We examined excerpts from the dialogues to gain further insight into the collaborative processes and outcomes. Our key findings are as follows: (1) women reported more stress, less perceived competence in their

computing abilities, and less perceived choice compared to men during a remote collaborative programming activity, and (2) dialogue features can provide insight into women's experiences during remote collaborative programming. To the best of our knowledge (and as further detailed in the next section), this work is the first to investigate the impact of structured roles for remote collaborative programming, particularly with a focus on the dialogue features associated with women's outcomes.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Theoretical Frameworks on Collaboration

A typical computer-supported collaboration consists of two or more human collaborators, one or more computers with a given collaboration environment, and any other tools that may be useful to the collaborators, such as reference sheets and external devices. *Social constructivist theory* states that knowledge is built not solely by individuals, but by the interactions between individuals and their environment [23]. Through their interactions and conversations, collaborators form models of how different components of their environment behave and, as a result, anticipate how the interactions will continue to unfold. In a purely textual chat interface, dialogue is the main form of interaction between collaborators. This paper, therefore, investigates how characteristics of the dialogue might provide insight into individual's perceptions of the collaboration.

It is also important to consider the individual with respect to their group and the larger community. Seering et al. described the Social Identity Perspective [50], which encompasses the social identity theory [53] and the subtheories built on that framework, and encouraged CSCW researchers to utilize this ideology as a potential lens to discuss and analyze their data. This theory states that individuals, in reaction to their involvement in a social group, create specific self-identities that are tailored to the distinct social circumstances experienced. Based on the significance of the group to the individual, this shift may manifest as altering behaviors, shifting motivations, or generating a sense of identity specific to the social environment encountered. In our study, we investigate the hypothesis that some participants may feel differently about their interactions within the community based on their individual characteristics. Specifically in computer science communities, it important to understand and better support those that have been historically marginalized in this field, such as women.

### 2.2 Collaborative Problem-Solving Paradigms in the Classroom

Collaborative problem solving has been identified as an essential 21st century skill [19] and is commonly employed in classrooms, especially for subjects with a heavy focus on open-ended problem solving, such as computer science. Recent studies have investigated how students' performance and learning differs between students that worked individually and students that engaged in collaborative problem solving within an online learning environment [12]. Students who collaborated exhibited higher learning gains (based on pre- and post-tests) and completed their assigned tasks in fewer steps compared to students that worked individually. This finding suggests that students who engaged in collaborative problem solving achieved a better understanding of the content and the interrelated parts of the problem, providing evidence for the benefits of collaborative problem solving to student learning.

Features of collaborative dialogue have the potential to reveal detailed information regarding the dynamics of a collaboration, including the social relationship between communicators [49]. Other work has focused on examining collaborative dialogue and the role of dialogue-related features in collaborative interactions. The effects of individual contributions on the overall quality of collaboration have been examined in the context of both spoken and text-based collaborations.

Spoken communication is inherently different than text-based communication, as in-person communication may improve the ability to bond with friends [51], while text-based communication may better facilitate the expression of affection between strangers [1] and increase their likelihood of disclosing intimate information [21, 58]. In the context of in-person communication, dialogue contributions with more positive sentiment have correlated with improved code performance and increased group satisfaction in terms of overall programming experience and developed code [22]. Additionally, another study investigated in-person collaborative dialogue through peer tutoring interactions. This study found that increased rapport with tutors significantly contributed to a supportive learning environment, as learners more willingly explained their thought processes during the collaboration [37].

Similar results have been revealed in the context of text-based collaborative dialogue. In particular, one study demonstrated that messages with more positive sentiment, higher engagement in the discussion, equal dialogue contributions between partners, and more contributions of relevant information during a collaboration was significantly correlated with higher group satisfaction and improved learning metrics [55]. Another text-based study in the context of computer science education revealed that programming project grades were positively correlated with higher numbers of posts in a class discussion board, higher numbers of replies to other students' posts, more positive dialogue contributions, and posting early with respect to the project deadline [66].

Similar to the work reported on in this article, Stewart et al. [52] conducted a study investigating the remote collaborative processes of student programmers. Specifically, they examined audio recordings of student triads collaboratively completing programming activities to understand (and automatically detect) three facets instrumental to collaborative problem solving: constructing shared knowledge, negotiating/coordinating solutions, and maintaining team function. This study yielded viable models that automated the detection of collaborative problem-solving techniques found in verbal communication. While Stewart et al. aimed to automate the analysis of collaborative dialogue, our study focuses on understanding the relationship between text-based collaborative dialogue and women's reported perceptions. The verbal communication studied by Stewart et al. holds different implications than the text-based communication that is examined in this study. Verbal communication may require less effort and better facilitate emotional expression though vocal inflections. Contrarily, text-based communication may be preferable for strangers [1, 21, 58], but might require more effort for emotional expression through the use of punctuation and response speed [15], or emoticons.

The particular collaborative problem-solving paradigm of pair programming has been widely studied within the computer science education community. Two collaborators, assigned roles of either driver or navigator and taking turns in these roles, collaborate on a programming activity [63]. Many benefits of this collaborative paradigm have been empirically demonstrated in the context of in-person pair programming: students achieved improved learning outcomes [63], better code quality [42], and increased retention rates in computer science courses [42]. Recent work has investigated modifications to this paradigm (e.g., how providing time to plan individually before collaboration benefits students [8]).

Most research studies on pair programming involve students who are co-located [9, 42] and working on the same physical computer [8, 34]. However, it is just as important to consider collaborations that happen remotely, in which each student works from a separate computer and are physically distant from each other. A study comparing both forms of collaboration found that, despite the different forms of interaction, code products created by remote collaborators were of a comparable quality to those created by co-located collaborators [3]. Other studies on remote collaboration look into how interfaces can better support collaborators, such as giving each collaborator their own cursor to select code and call attention to it [16]. In one of the few

empirical studies of remote pair programming dialogues, researchers investigated the relationship between expressing and addressing uncertainty, finding the importance of resolving uncertainty before moving on to the next subtask [45]. Understanding how students interact in a collaborative environment is the first step toward improving their experiences. Analyzing their dialogue can help provide insight into the dynamics of these collaborations. Additionally, it is important to investigate the impact of personal characteristics on collaboration. The studies mentioned in this paragraph, for example, did not include analyses with respect to gender, although men and women may experience pair programming differently [62, 65]. In this article, we compare the students' remote pair programming experiences by gender to further investigate the nuances of remote collaboration.

### 2.3 Gender and Collaborative Problem Solving

Recent years have brought a deeper understanding of gender identities including female (woman), male (man), agender, genderfluid, and non-binary [25, 69]. It is crucial to understand the collaboration experiences of people of all gender identities; however, all participants in our study self-identified as either male or female, thus limiting our study results to these two gender identities. This subsection focuses on related work on gender differences between women and men in collaboration.

To gain a deeper understanding of people's collaborative problem-solving experiences, researchers have begun investigating people's perceptions of these collaborations with respect to individual characteristics, such as gender identity. Previous work has found that women and men commonly experience and perceive remote collaborative problem solving differently. Women tend to report lower self-efficacy, or confidence, regarding both the technological medium [28, 39] and the task being performed [28]. Additionally, women collaborating in female-majority teams have been found to make more frequent verbal contributions to collaborative dialogues than male-majority teams or teams with a balanced gender distribution [67], as well as adapt their verbal contributions to have higher specificity in the absence of visual feedback [28]. Prior research has also focused on differences between men's and women's perceptions of pair programming [31, 65]. For example, in a study that aimed to compare women's and men's experiences in remote pair programming, Kuttal et al. [31] observed that men preferred working remotely due to the ease of switching between driver/navigator roles and their reported higher comfort with remote communication. Contrarily, the women in this study would have preferred to be co-located, as they felt disconnected from their partners and reported that in-person collaborations would have better facilitated communication.

It is particularly important to understand women's experiences during remote collaborative programming, because computing and related fields constitute a stable, lucrative career path that should be accessible to all people. Nevertheless, women continue to be marginalized in these fields. This disparity can be exemplified by the fact that women earned only 24.4% of computer science bachelor's degrees at U.S. non-doctoral institutions in 2019 [71]. Even in countries where educational enrollments are more equal, the situation remains troubling. For example, in India, women constitute 45% of university enrollments in computer science fields, but discriminatory workplace practices mean that women are remaining in entry-level positions rather than being promoted (80% of entry-level positions are held by women) [56]. Understanding women's experiences during remote collaborative programming is one step among many to relieve such social, educational, and occupational barriers that inhibit women's participation in computer science and related fields.

## 3 COLLABORATIVE PROGRAMMING STUDY

Participants for this study were recruited in Spring 2019 from an introductory computer science course at the University of Florida, located in the southeastern United States. The study was

Table 1. Demographic breakdown of individual participants and gender composition of pairs by condition.

| | Structured-Role Condition | Unstructured-Role Condition | Total |
|---|---|---|---|
| **Participants/Individuals** | 32 | 26 | 58 |
| **Gender** | | | |
| Woman/Female | 14 | 10 | 24 |
| Man/Male | 18 | 16 | 34 |
| **Race/Ethnicity** | | | |
| White/Caucasian | 19 | 12 | 31 |
| Hispanic/Latino | 7 | 3 | 10 |
| Asian/Pacific Islander | 3 | 7 | 10 |
| Black/African-American | 3 | 1 | 4 |
| Multiracial | 0 | 3 | 3 |
| **Age** | | | |
| 18-19 | 24 | 20 | 44 |
| 20-21 | 6 | 2 | 8 |
| 22-23 | 1 | 2 | 3 |
| 24-25 | 0 | 1 | 1 |
| 26-27 | 0 | 1 | 1 |
| 28 or older | 1 | 0 | 1 |
| **Pairs** | 16 | 13 | 29 |
| Woman-Woman Pair | 4 | 3 | 7 |
| Woman-Man Pair | 6 | 4 | 10 |
| Man-Man Pair | 6 | 6 | 12 |

conducted outside of course hours and occurred at the end of the semester after all class meetings ended and before final exam week. This study was one of three options students could choose from to earn up to two percentage points of extra credit towards their final course grade. Students expressed interest in participating by completing a form with their scheduling availability. There were 58 participants (24 women, 34 men) in the study, resulting in 29 pairs who are the target of this analysis. The vast majority (90%) of the participants were between the ages of 18 and 21. See Table 1 for additional demographic information. The demographic distribution of the participants is representative of the demographics from the introductory computer science course at this university, with approximately half of the participants identifying as White/Caucasian.

First, participants signed a consent form that specified the purpose of the research study: to identify and support effective collaboration among computer science learners. Then, we assigned the participants to one of six scheduled meeting times for the study according to their availability. All meetings were conducted in the same conference room on the university's campus. The room was set up to have 14 workstations, arranged so that participants would not have a direct view of any other participant's screen. Workstations included a laptop and mouse provided by the research team. Additionally, workstations were staggered so that stations which would be paired for remote collaboration were in separate rows and never in adjacent spots. Participants were paired according to when they arrived to the conference room by sending them to each workstation accordingly. They were not told who they were paired with at any point during the study. Some participants, however,

Table 2. Order of activities for each study session. For the structured-role condition, the pair programming paradigm was described at the same time the IntelliJ and Floobits software were introduced.

| Duration | Task |
|---|---|
| <5 min | Consent Form |
| ~5 min | Pre-Test (5 multiple-choice questions) |
| ~5 min | Introduction to the Software Environment |
| ~60 min | Collaborative Activity (create Tic-Tac-Toe game) |
| ~5 min | Intrinsic Motivation Inventory Survey (18 items, 7-point Likert) |
| <5 min | System Usability Scale (10 items, 5-point Likert) |
| ~10 min | Open-Ended Collaboration Questions (5 questions) |
| ~5 min | Post-Test (identical to Pre-Test) |
| <5 min | Demographics Survey (10 questions) |

chose to introduce themselves in the chat within the first few messages. Three pairs exchanged their first names and one person in another pair introduced himself without reciprocation. The seven participants that disclosed their name identified as men and had traditionally male names. The participant that did not reciprocate providing their name, identified as a woman. Based on the dialogue, it appears that these students did not know each other before the study. Other than these first-name disclosures, participants were unaware of the identity of their partner.

Due to late arrivals and absences, there were a minimum of four and a maximum of six pairs for each meeting time. Table 2 provides a detailed list of the order of activities. Participants first completed a pre-test, then collaborated remotely for approximately one hour on a coding task by following one of two collaborative paradigms, and finally completed a post-test and set of post-surveys.

Based on which collaborative condition participants were assigned to, participants completed the study following one of two collaborative paradigms. In the *structured-role* condition, participants followed the pair programming paradigm, with each person in the pair performing the role of the driver or navigator throughout the collaboration, swapping roles periodically. In the *unstructured-role* condition, participants were not given specific roles to follow, allowing the participants to decide how the collaboration would unfold. Participants were assigned to either the structured-role or unstructured-role condition based on the study meeting time they were assigned to, since the structured-role condition required additional scaffolding and instructions by researchers. We aimed to have an equal number of participants for the conditions, but due to absences and late participants, the structured-role condition had three more pairs than the unstructured-role condition. Sixteen pairs (14 women and 18 men) were assigned to the structured-role condition and were reminded by researchers to change roles every 15 minutes. The remaining participants (13 pairs; 10 women; 16 men) were paired and assigned to the unstructured-role condition. Table 1 shows the breakdown in gender composition for the pairs in each condition. The structured-role condition had four woman-woman pairs, six woman-man pairs, and six man-man pairs, while the unstructured-role condition had three woman-woman pairs, four woman-man pairs, and six man-man pairs. The participants in the structured-role condition received printed instructions explaining the pair programming paradigm and the responsibilities of each role (see Figure 2). The facilitator also announced the instructions verbally at the start of the meeting. No role-structure announcements or related materials were provided to participants in the unstructured-role condition.

Participants had approximately one hour to complete a programming task with their remote partner using the Floobits plugin on the IntelliJ integrated development environment (see Figure 1).

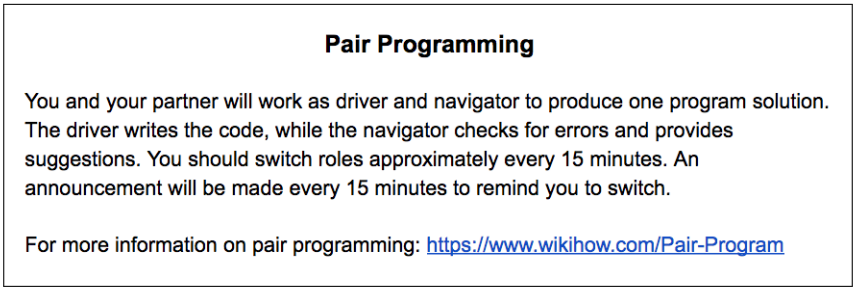


Fig. 2. Pair programming description provided to students in the structured-role collaboration condition.

They programmed using the Java programming language, which was the language used in their programming course. We did not directly ask the students about their familiarity with the IntelliJ environment or the Floobits plugin, but the researcher who conducted the study observed that the participants were not familiar with Floobits prior to this study. The Floobits plugin is open-source software that supports real-time collaborative coding. It allows users to chat textually via a built-in messenger, write code in the same project or file simultaneously, and has various features such as summoning collaborators to an active cursor position. For our research purposes, we modified the plugin to add event logging to a local text file, which captured participants' chat messages and clicks within the interface. Participants practiced incorporating *try-catch* blocks into code as they created a program to allow two people to play tic-tac-toe. We provided all participants with three reference materials in printed form: a quick-reference on the try-catch construct, a one-page guide to using the Floobits collaboration platform, and a copy of the requirements for their tic-tac-toe program.

We opted for textual chat communication between the programming partners for some of the same reasons that collaborators in classes and workplaces often choose textual collaboration [60]. For example, when compared to video/audio communication, textual chat requires less online bandwidth and is more robust to changes in online signal strength. Additionally, textual chat does not require additional hardware, such as a microphone or a camera, which can vary in quality. Finally, textual chat is less invasive and provides users with a chat history, allowing them to refer back to what has been discussed during the conversation. This feature in particular is useful in a problem-solving context such as programming.

To measure conceptual understanding and learning, we crafted a pre-test consisting of five multiple-choice questions to measure prior knowledge on the try-catch coding construct, which the students' introductory course had not yet covered. We purposely introduced the participants to this new coding construct, which they were required to use in their coding solution, so that we could measure whether there were any differences in normalized learning gain between the collaboration conditions. The post-test was identical to the pre-test.

After the collaborative activity, participants completed a set of post-surveys. Participants were first asked about affective/motivational outcomes, for which we turned to the widely used and validated Intrinsic Motivation Inventory (IMI) instrument, originally created in 1989 [40]. The IMI includes seven subscales measuring Interest/Enjoyment, Perceived Competence, Effort/Importance, Pressure/Tension, Perceived Choice, Value/Usefulness, and Relatedness. The seventh subscale, Relatedness, focuses on interpersonal interactions between collaborators; it was added to the IMI and validated in 2016 [27]. Survey items are on a Likert scale from 1 (not at all true) to 7 (very true). Next, participants completed the System Usability Scale (SUS) survey [6], which gives insight into

Table 3. Racial breakdown, comfort with computers, and prior programming experience by gender. Prior programming experience refers to any programming experience prior to taking the introductory computer science course.

| | Women (*n*=24) | | Men (*n*=34) | | Total (*n*=58) | |
|---|---|---|---|---|---|---|
| **Race/Ethnicity** | | | | | | |
| White/Caucasian | 16 | (67%) | 15 | (44%) | 31 | (53%) |
| Hispanic/Latino | 2 | (8%) | 8 | (24%) | 10 | (17%) |
| Asian/Pacific Islander | 4 | (17%) | 6 | (18%) | 10 | (17%) |
| Black/African-American | 2 | (8%) | 2 | (6%) | 4 | (7%) |
| Multiracial | 0 | (0%) | 3 | (9%) | 3 | (5%) |
| **Comfort with Computers** | | | | | | |
| I am very comfortable and have used computers extensively. | 11 | (46%) | 18 | (53%) | 29 | (50%) |
| I am comfortable but have not used them extensively. | 4 | (17%) | 13 | (38%) | 17 | (29%) |
| I am moderately comfortable with computers. | 7 | (29%) | 3 | (9%) | 10 | (17%) |
| I am a little uncomfortable using computers. | 2 | (8%) | 0 | (0%) | 2 | (4%) |
| I am very uncomfortable using computers. | 0 | (0%) | 0 | (0%) | 0 | (0%) |
| **Prior Programming Experience** | | | | | | |
| no | 17 | (71%) | 18 | (53%) | 35 | (60%) |
| yes | 7 | (29%) | 16 | (47%) | 23 | (40%) |
| **Prior Java Programming Experience** | | | | | | |
| none | 15 | (63%) | 15 | (44%) | 30 | (52%) |
| a little | 6 | (25%) | 11 | (32%) | 17 | (29%) |
| some | 3 | (13%) | 7 | (21%) | 10 | (17%) |
| a lot | 0 | (0%) | 1 | (3%) | 1 | (2%) |

how easily students adapted to the new interface. We solicited this usability information to account for any variability in how the interface features supported each collaborative paradigm.

The post-survey next included open-ended questions about the collaborative experience. These questions asked participants about their collaboration methods and what they liked or disliked about the software they had used (see Table 4). In the unstructured-role condition, we asked participants what their collaboration strategy was and whether they felt it was a good approach. We present a sample of these responses in section 7, but the majority of this paper focuses on analyzing the 18 intrinsic motivation outcomes (see Table 5). At the very end (after the post-test), participants were asked demographic information such as age, gender identity, race, student classification, and prior programming experience. All participants identified their gender as either "male" or "female"; two additional options (an "other" option with a field to self-describe, and a "prefer not to say" option) were also available. See Table 3 for the racial breakdown and prior experience of the participants stratified by gender. Prior experience indicated here is that of experience prior to taking the introductory computer science course.

Every participant individually completed a pre- and post-test on try-catch blocks to measure normalized learning gain. Normalized learning gain (Equation 1) is calculated as the difference in pre- and post-test scores divided by the difference in maximum score and the pre-test score all

Table 4. Open-ended questions and sample responses from post-survey. Grammar from original student responses is preserved.

| **Question (Unstructured-role Condition)** | **Example Response** |
|---|---|
| What was your strategy for working together on this program? (i.e. Did you divide-and-conquer? Did you focus on the same piece of code at a time?) Please describe. | We mostly worked on different things at the same time while consulting with each other about our approaches to solving these problems. |
| Do you think this was a good strategy? Please explain why or why not. | I think it was a good strategy because we were able to get more done. |
| **Question (Structured-role Condition)** | **Example Response** |
| Did you enjoy pair programming? Please explain why or why not. | not really; it felt like whenever it wasn't my turn to be the driver i was just itching to do it myself - it can be hard to explain ideas to another person through text |
| Do you think pair programming was a good strategy for collaborating on this program? Please explain why or why not. | sure; i don't really know any other method to collaborate on a program that doesn't result in one person doing all the work |
| **Question (Both Conditions)** | **Example Response** |
| What software features were crucial for collaborating and why? | The chat box and the fact that we were able to see where our partner's cursor was on the screen was very helpful. We were able to see what the other person was working on so that no one started coding the same thing at once. |
| How could this software be improved to support your collaboration? Please describe any additional features that would help you. | Adding a live call chat could help with the collaboration since we wouldn't have to keep looking down to see if someone typed something. A notification system could also help so that we are notified if someone is typing or if someone has typed something. |
| Did you feel limited by the software? If so, please explain why. | No, it reminded me of Google Docs. |

multiplied by 100. This metric captures the percent of improvement over the pre-test compared to maximum potential improvement.

$$normalized\ learning\ gain = \frac{post\text{-}test\ score - pre\text{-}test\ score}{100 - pre\text{-}test\ score} * 100 \quad (1)$$

## 4 DIALOGUE FEATURE EXTRACTION

In collaborative problem-solving research, dialogue analysis is often used to investigate outcomes such as satisfaction [38] and perception [5, 14] through techniques including sentiment analysis

Table 5. Intrinsic Motivation Inventory (IMI) items used in post survey.

| Scale | Item |
|---|---|
| Interest/Enjoyment | This activity was fun to do.<br>This activity did not hold my attention at all.<br>I thought this activity was quite enjoyable. |
| Perceived Competence | I think I did pretty well at this activity, compared to other students.<br>I am satisfied with my performance at this task.<br>This was an activity that I couldn't do very well. |
| Effort/Importance | I put a lot of effort into this.<br>I didn't try very hard to do well at this activity.<br>It was important to me to do well at this task. |
| Pressure/Tension | I felt very tense while doing this activity.<br>I was very relaxed in doing these. |
| Perceived Choice | I didn't really have a choice about doing this task.<br>I did this activity because I wanted to. |
| Value/Usefulness | I believe doing this activity could be beneficial to me.<br>I think this is an important activity. |
| Relatedness | I felt really distant to this person.<br>I'd like a chance to interact with this person more often.<br>It is likely that this person and I could become friends if we interacted a lot. |

[14, 30, 33, 52] and utterance intent classification [52]. In learning contexts, the amount of talk is also an important feature within collaborations, particularly with regard to balanced contributions from collaborators [34].

### 4.1 Extracting Lexical Features

In our study, distribution of talk can be measured using the number of messages each participant sent as well as the number of words they used. Study participants communicated through textual chat, and we logged these chat messages as part of the study and used them to extract six lexical features: 1) the total number of messages sent by the participant (i.e., number of messages), 2) the total number of words typed by the participant (i.e., number of words), 3) the average words per message of the participant (i.e., average words per message), 4) the total number of messages sent by the participant's partner (i.e., number of partner messages), 5) the total number of words sent by the partner (i.e., number of partner words), and 6) the average words per message of the partner (i.e., average partner words per message).

### 4.2 Sentiment Analysis

Sentiment analysis is an active area of natural language processing research that focuses on identifying the sentiment (for example, positive/negative or favorable/unfavorable) expressed within an utterance or piece of text. Most sentiment analysis research has focused on online content such as blog posts [14], but recent work in computer-supported collaborative learning has begun

to examine sentiment during collaborative problem solving in computer science education [52]. Because of its link to affective or emotional factors as expressed within dialogue, we examined sentiment of collaborators' utterances by automatically assigning sentiment intensity using the VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner) sentiment analysis package [18]. Each word within a message is assigned a sentiment intensity based on a predefined table of words with corresponding sentiment intensity values: the numeric values represent the intensity of the sentiment for a given word, and the signs represent the polarity of the word's sentiment (positive or negative), with neutral sentiment represented as a zero. VADER has been used widely for sentiment analysis on social media posts [11, 24] and as such takes into consideration whether the text is written in all uppercase letters, whether there are repeated punctuation marks, and whether the text includes emoticons or common online acronyms, among other things. Those textual communication phenomena are common within our corpus of collaborative dialogue, making VADER suitable for our analysis.

The sentiment intensity of each message is calculated as a compound score by adding the sentiment intensity of each word in the message and standardizing the result to a value between -1 and 1. We took these scores and calculated the average compound sentiment of each participant by averaging the compound sentiment score of each message from that given participant. This process resulted in two features: 1) average compound sentiment of the participant and 2) average compound sentiment of the partner.

## 5 DATA ANALYSIS

### 5.1 Comparison of Structured- and Unstructured-Role Conditions

*5.1.1 Outcome Metrics.* Table 6 summarizes the comparisons between the two collaboration conditions. Since our data could not be assumed to follow a normal distribution, we use nonparametric tests for statistical comparisons. There were no significant differences detected in normalized learning gain, SUS scores, or any items from the Intrinsic Motivation Inventory (IMI) between students in the structured-role (pair programming) and unstructured-role collaboration conditions, according to Wilcoxon rank sum tests. Within each condition, participants exhibited a significant normalized learning gain (see Table 7), according to one-sample Wilcoxon signed rank tests.

SUS scores for those in the unstructured-role collaboration condition averaged 66.54, and SUS scores for those in the structured-role (pair programming) condition averaged slightly higher (but not significantly) at 67.81. Both of these results fall just below the average SUS score of 68. This average indicates an acceptable degree of usability, as it functions as the center of the Sauro-Lewis curved grading scale [35], which was specifically designed to interpret SUS scores [46]. Since both groups of participants averaged SUS scores lower than this threshold, they likely struggled to adapt to this new interface while they were collaborating and may have felt overwhelmed. Future studies should include longer collaboration sessions so that participants have time to familiarize themselves with the collaboration software.

*5.1.2 Dialogue Features.* An overview of the dialogue features in our dataset can be found in Table 8. We removed two pairs (four participants) from the unstructured-role condition because at least one of the partners did not send any messages during the study, leaving 22 participants from the unstructured-role condition. We found significant differences in the number of messages and number of words between the collaboration conditions, according to Wilcoxon rank sum tests and after applying the Benjamini-Hochberg procedure as a correction for multiple comparisons over an alpha of 0.05 (corrected $p$-value<0.0125) [57]. Students in the structured-role condition had higher number of messages, number of words, number of partner messages, and number of partner

Table 6. Summary of normalized learning gain, SUS scores, and IMI survey item responses by condition and by gender. Shows averages and standard deviations, numbers in bold are significantly different ($p$<0.0125).

| Category | Structured-role ($n$=32) | Unstructured-role ($n$=26) | Women ($n$=24) | Men ($n$=34) |
|---|---|---|---|---|
| **Normalized Learning Gain** (out of 100) | 17.3 (22.9) | 15.4 (24.4) | 16.2 (23.2) | 16.7 (23.8) |
| **System Usability Scale Scores** (out of 100) | 67.8 (16.8) | 66.5 (17.5) | 63.3 (17.6) | 70 (16.1) |
| **Intrinsic Motivation Inventory** (1 = not at all true, to 7 = very true) | | | | |
| **Interest/Enjoyment** | | | | |
| This activity was fun to do. | 5.56 (2.46) | 4.96 (1.61) | 4.96 (1.73) | 5.53 (1.38) |
| This activity did not hold my attention at all. | 1.97 (1.12) | 2.31 (1.35) | 2.17 (1.34) | 2.09 (1.16) |
| I thought this activity was quite enjoyable. | 5.25 (1.37) | 4.92 (1.47) | 4.67 (1.63) | 5.41 (1.16) |
| **Perceived Competence** | | | | |
| I think I did pretty well at this activity, compared to other students. | 3.72 (1.75) | 3.42 (1.63) | **2.88 (1.7)** | **4.09 (1.5)** |
| I am satisfied with my performance at this task. | 4.34 (1.7) | 4.12 (1.61) | 3.71 (1.76) | 4.62 (1.48) |
| This was an activity that I couldn't do very well. | 3.97 (1.93) | 4.12 (1.97) | 4.58 (2.0) | 3.65 (1.81) |
| **Effort/Importance** | | | | |
| I put a lot of effort into this. | 5 (1.02) | 4.92 (1.23) | 4.79 (1.1) | 5.09 (1.11) |
| I didn't try very hard to do well at this activity. | 1.81 (1) | 2.23 (1.11) | 2.04 (0.95) | 1.97 (1.14) |
| It was important to me to do well at this task. | 4.66 (1.12) | 4.77 (1.24) | 4.71 (0.81) | 4.71 (1.38) |
| **Pressure/Tension** | | | | |
| I felt very tense while doing this activity. | 3.19 (1.69) | 2.92 (1.7) | **3.96 (1.52)** | **2.44 (1.52)** |
| I was very relaxed in doing these. | 4.47 (1.5) | 4.35 (1.5) | **3.67 (1.49)** | **4.94 (1.25)** |
| **Perceived Choice** | | | | |
| I didn't really have a choice about doing this task. | 2.19 (1.47) | 2.54 (1.65) | 2.75 (1.62) | 2.06 (1.46) |
| I did this activity because I wanted to. | 5.22 (1.41) | 4.92 (1.67) | **4.5 (1.44)** | **5.5 (1.46)** |
| **Value/Usefulness** | | | | |
| I believe doing this activity could be beneficial to me. | 5.16 (1.48) | 4.88 (1.48) | 4.75 (1.22) | 5.24 (1.62) |
| I think this is an important activity. | 5.53 (1.14) | 5.73 (0.92) | 5.33 (0.92) | 5.82 (1.09) |
| **Relatedness** | | | | |
| I felt really distant to this person. | 2.94 (1.72) | 3.65 (1.77) | 3.17 (1.58) | 3.32 (1.9) |
| I'd like a chance to interact with this person more often. | 4.56 (1.58) | 4.42 (1.47) | 4.42 (1.35) | 4.56 (1.65) |
| It is likely that this person and I could become friends if we interacted a lot. | 4.84 (1.42) | 4.46 (1.17) | 4.5 (1.18) | 4.79 (1.41) |

Table 7. Average pre-test and post-test scores (shows average and standard deviation) and one-sample Wilcoxon signed rank tests for normalized learning gain by condition and by gender.

| Group | Pre-Test (out of 5) | Post-Test (out of 5) | Normalized Learning Gain | $p$-value |
|---|---|---|---|---|
| Unstructured-role ($n$=26) | 1.15 (0.92) | 1.88 (0.77) | 15.38 (24.39) | 0.0012 |
| Structured-role ($n$=32) | 0.93 (0.8) | 1.72 (0.77) | 17.34 (22.85) | < 0.0001 |
| Women ($n$=24) | 1.04 (0.91) | 1.79 (0.78) | 16.18 (23.24) | 0.0012 |
| Men ($n$=34) | 1.03 (0.83) | 1.79 (0.77) | 16.67 (23.79) | 0.0001 |

Table 8. Overview of dialogue features by condition and by gender. Shows averages and standard deviations, numbers in bold are significantly different ($p$<0.0125). Because partners belonged to the same collaboration condition, the values for participant features and partner features in the role condition columns are the same.

| Dialogue Feature | Structured-role ($n$=32) | Unstructured-role ($n$=22) | Women ($n$=24) | Men ($n$=30) |
|---|---|---|---|---|
| No. Messages | **49.03 (27.31)** | **29.68 (27.2)** | 37.42 (28.0) | 44.13 (29.3) |
| No. Words | **417.63 (225.61)** | **212.82 (147.42)** | 324.33 (223.37) | 342.07 (222.01) |
| Avg. Words | 8.88 (2.82) | 8.01 (3.41) | 9.21 (3.58) | 7.97 (2.53) |
| Avg. Sentiment | 0.1 (0.06) | 0.11 (0.09) | 0.11 (0.08) | 0.1 (0.07) |
| Partner No. Messages | **49.03 (27.31)** | **29.68 (27.2)** | 39.42 (26.26) | 42.53 (30.82) |
| Partner No. Words | **417.63 (225.61)** | **212.82 (147.42)** | 326.04 (200.32) | 340.7 (238.91) |
| Partner Avg. Words | 8.88 (2.82) | 8.01 (3.41) | 9.13 (3.33) | 8.04 (2.81) |
| Partner Avg. Sentiment | 0.1 (0.06) | 0.11 (0.09) | 0.1 (0.08) | 0.11 (0.07) |

words than students in the unstructured-role condition. However, according to Cohen's $d$, these four items show small to no effect size (0.03, 0, 0.03, and 0, respectively) [10].

### 5.2 Comparison by Gender

Because there were no significant differences in outcomes based on the structured- and unstructured-role conditions, and the sample size does not support a disaggregated analysis by both condition and gender, we pooled the data from both conditions and proceeded with a by-gender analysis of outcomes and dialogue features. We do not compare pairs based on their gender composition (woman-woman, woman-man, man-man) due to the limited sample size.

*5.2.1 Outcome Metrics.* We examined the differences in women's and men's self-reported experiences from their survey responses, and the results revealed several significant differences by gender on items from the IMI according to Wilcoxon rank sum tests (shown in Table 6). After applying the Benjamini-Hochberg procedure as a correction for multiple comparisons, four items remained significant ($p$<0.0125): two from the Pressure/Tension subscale, one from the Perceived Competence subscale, and one from the Perceived Choice subscale. According to Cohen's $d$, the Pressure/Tension items show a large effect size (0.66 and 0.68, respectively), while the Perceived Competence and Perceived Choice items show a medium effect size (0.48 and 0.47. respectively).

Both women and men had significant normalized learning gains from the activity, and normalized learning gain was not significantly different between them (see Table 7).

The SUS scores reported between men and women were not significantly different, with men reporting an average of 70.0 and women reporting an average of 63.3. However, the higher average SUS score given by men and the lower average score by women support findings from prior remote collaborative work that men tend to prefer remote collaboration while women prefer to collaborate when co-located [31].

Additionally, we found no significant difference in prior programming experience between women and men according to a likelihood ratio Chi-square test ($p$=0.1666). After collapsing students' responses about their prior Java programming experience into two responses (no experience vs. some experience), we found no significant difference in prior Java programming experience between women and men according to a likelihood ratio Chi square test ($p$=0.1660). The vast majority of students also reported comfort with using computers (92% of women and 100% of men).

*5.2.2 Dialogue Features.* An overview of these features in our dataset can be found in Table 8. We removed four men from this analysis because they belonged to pairs in which at least one person did not send any messages during the study, leaving 30 male participants. There were no significant differences detected in these dialogue features between men and women, according to Wilcoxon rank sum tests.

## 6 EXAMINING COLLABORATIVE DIALOGUE: MODELING WOMEN'S EXPERIENCES

Given the significant differences in responses to the Pressure/Tension, Perceived Competence, and Perceived Choice items between men and women, we next took a deeper look at the collaborative process itself, namely the dialogues exchanged between collaborators. For the remainder of this paper, we focus on understanding how dialogue features from the collaboration might indicate women's stress, perceived competence, or perceived choice through the use of regression modeling.

### 6.1 Feature Selection

The extraction of lexical and sentiment features from the dialogues resulted in eight dimensions of dialogue features. Our goal was to determine which of these features were significantly correlated with women's outcomes. To identify whether any groups of features were highly correlated, we performed Principal Component Analysis [64]. The output of this analysis is an 8x8 correlation matrix $R$, where each value in the matrix represents the correlation $r$ between each pair of factors (see Table 9). Based on standard practice, we deemed feature pairs with $r > 0.7$ as strongly correlated features [7]: for our set of eight features, the number of words was strongly correlated with the number of messages, and the same was true for the partner words and messages. The decision of whether to eliminate word-based or message-based features was made based on the auxiliary correlations: collaborators' number of messages were correlated with each other at 0.6266, while number of words were correlated at 0.6565. Due to this slightly lower correlation between message-based features than word-based features across collaborators, we opted to remove the number-of-words-based features and build the regression models using the following six features: 1) number of messages, 2) average words per message, 3) average compound sentiment, 4) number of partner messages, 5) average partner words per message, and 6) average partner compound sentiment.

### 6.2 Dialogue Features that Correlate with Women's Reported Experiences

Our goal was to identify significant relationships between dialogue features and the womens' self-reported outcomes. Given that the outcome responses take the form of a 7-point Likert scale, we used *ordinal logistic regression* [41] to model these responses. Ordinal regression is used to

Table 9. Correlation Matrix for Principal Components Analysis of Dialogue Features. Strong correlations (r > 0.7) are bolded.

| | No. Messages | No. Words | Avg. Words | Avg. Sentiment | Partner No. Messages | Partner No. Words | Partner Avg. Words | Partner Avg. Sentiment |
|---|---|---|---|---|---|---|---|---|
| No. Messages | - | **0.8857** | -0.2116 | 0.0356 | 0.6266 | 0.5613 | -0.3025 | 0.1449 |
| No. Words | **0.8857** | - | 0.1607 | 0.0656 | 0.5715 | 0.6565 | -0.1598 | 0.0501 |
| Avg. Words | -0.2116 | 0.1607 | - | 0.1553 | -0.1820 | -0.0400 | 0.0683 | -0.1989 |
| Avg. Sentiment | 0.0356 | 0.0656 | 0.1553 | - | 0.0776 | -0.0108 | -0.3110 | -0.0957 |
| Partner No. Messages | 0.6266 | 0.5715 | -0.1820 | 0.0776 | - | **0.8095** | -0.4024 | 0.2820 |
| Partner No. Words | 0.5613 | 0.6565 | -0.0400 | -0.0108 | **0.8095** | - | 0.1153 | 0.1795 |
| Partner Avg. Words | -0.3025 | -0.1598 | 0.0683 | -0.3110 | -0.4024 | 0.1153 | - | -0.2834 |
| Partner Avg. Sentiment | 0.1449 | 0.0501 | -0.1989 | -0.0957 | 0.2820 | 0.1795 | -0.2834 | - |

model outcomes that follow an ordinal format, such as a Likert scale. The model defines a logistic probability function for the thresholds between two items on an ordinal scale (e.g., for a 7-point Likert scale, six functions are defined). Each function in the model takes the form of Equation 2 [36]. The logit of the probability that an outcome $Y$ will fall at or below a given threshold $j$ is the intercept of the corresponding threshold $\alpha_j$ minus the sum of products of each feature $x_i$ and its corresponding coefficient $\beta_i$ over all $n$ features. Since the equation subtracts the sum of products of features and coefficients from the intercepts, a negative coefficient actually indicates a positive correlation of the corresponding feature with the outcome, and vice versa.

$$logit(P(Y \leq j)) = \alpha_j - \sum_{i=1}^{n} \beta_i x_i \tag{2}$$

In our case, responses to all four survey items ranged between six values (one of the seven values on the scale did not occur for each of the items). We built separate regression models for each of the survey responses that had emerged as significantly different for women than men. We decided to treat each item individually because none of the items comprised a full subscale of the validated IMI survey.

*6.2.1 Pressure/Tension.* We first built ordinal regression models for women's responses to the Pressure/Tension items from our study, one for each survey item: the *tension* survey item ("I felt very tense while doing this activity") and the *relaxation* survey item ("I was very relaxed in doing

Table 10. Ordinal regression model of women's responses to the relaxation survey item.

| Feature/Parameter | Estimate | Standard Error | $p$-value |
|---|---|---|---|
| **number of messages** | **-0.0419** | **0.0201** | **0.0369*** |
| avg. words per message | -0.1921 | 0.1208 | 0.1118 |
| avg. compound sentiment | -6.144 | 5.2273 | 0.2398 |
| number of partner messages | 0.0372 | 0.0221 | 0.0921 |
| **avg. partner words per message** | **-0.3578** | **0.1574** | **0.0231*** |
| **avg. partner compound sentiment** | **-14.5901** | **6.2633** | **0.0198*** |
| intercept 1\|2 | 4.6078 | 2.5745 | 0.0735 |
| intercept 2\|3 | 5.0367 | 2.5871 | 0.0516 |
| intercept 3\|4 | 6.7871 | 2.7609 | 0.0140* |
| intercept 4\|5 | 9.1959 | 3.0776 | 0.0028* |
| intercept 5\|6 | 9.5628 | 3.1072 | 0.0021* |

these"). Both models used the six dialogue features described previously. Based on likelihood ratio Chi-square tests, the model for the tension survey item did not fit the data significantly ($p = 0.2058$). However, the model for the relaxation survey item did result in a significant fit ($p = 0.0497$). Table 10 shows the resulting parameter estimates and their corresponding $p$-values.

Responses ranged between 1 and 6, resulting in five intercept estimates representing the thresholds between each pair of responses on the Likert scale. For this model, the number of messages, the average partner words per message and the average partner compound sentiment were significantly positively correlated with women's responses to the relaxation survey item ($p < 0.05$).

*6.2.2 Perceived Competence.* Next, we built an ordinal regression model for women's responses to the Perceived Competence survey item from our study, which we refer to as the *competence* survey item ("I think I did pretty well at this activity, compared to other students"). The model used the same six dialogue features described previously. Based on a likelihood ratio Chi-square test, the model for the competence survey item resulted in a significant fit to the data ($p = 0.0011$). Table 11 shows the resulting parameter estimates and their corresponding $p$-values indicating their correlation to the response. Women's reports of lower perceived competence support prior research that has found women tend to report lower self-efficacy in computer-supported collaboration [28, 39].

Responses featured scores of 1, 2, 3, 4, 5, and 7, resulting in five intercept estimates representing the thresholds between each pair of responses on the Likert scale. For this model, the number of messages, the average words per message, and the average partner compound sentiment were significantly positively correlated with women's responses to the competence survey item ($p < 0.05$). Additionally, the number of partner messages was significantly negatively correlated with women's responses to the competence survey item ($p < 0.05$).

*6.2.3 Perceived Choice.* Finally, we built an ordinal regression model for women's responses to the Perceived Choice survey item from our study, which we refer to as the *choice* survey item ("I did this activity because I wanted to"). The model used the same six features described in the previous section. Based on a likelihood ratio Chi-square test, the model for the choice survey item resulted in a significant fit to the data ($p = 0.0022$). Table 12 shows the resulting parameter estimates and their corresponding $p$-values indicating their correlation to the response.

Table 11. Ordinal regression model of women's responses to the competence survey item.

| Feature/Parameter | Estimate | Standard Error | $p$-value |
|---|---|---|---|
| **number of messages** | **-0.0741** | **0.0259** | **0.0042*** |
| **avg. words per message** | **-0.3486** | **0.1397** | **0.0126*** |
| avg. compound sentiment | -1.6204 | 5.3791 | 0.7632 |
| **number of partner messages** | **0.1026** | **0.0371** | **0.0057*** |
| avg. partner words per message | -0.2761 | 0.1541 | 0.0731* |
| **avg. partner compound sentiment** | **-19.0354** | **6.7264** | **0.0047*** |
| intercept 1\|2 | 4.1740 | 2.6268 | 0.1121 |
| intercept 2\|3 | 6.6701 | 2.9334 | 0.0230* |
| intercept 3\|4 | 8.4127 | 3.0896 | 0.0065* |
| intercept 4\|5 | 10.1856 | 3.2507 | 0.0017* |
| intercept 5\|6 | 10.7508 | 3.2989 | 0.0011* |

Table 12. Ordinal regression model of women's responses to the choice survey item.

| Feature/Parameter | Estimate | Standard Error | $p$-value |
|---|---|---|---|
| number of messages | -0.0376 | 0.0212 | 0.0754 |
| **avg. words per message** | **-0.4746** | **0.156** | **0.0024*** |
| avg. compound sentiment | -2.6347 | 5.4034 | 0.6258 |
| **number of partner messages** | **0.0670** | **0.0256** | **0.0088*** |
| avg. partner words per message | -0.1964 | 0.1498 | 0.1899 |
| **avg. partner compound sentiment** | **-19.1850** | **6.7892** | **0.0047*** |
| intercept 1\|3 | 2.0568 | 2.5721 | 0.4239 |
| intercept 3\|4 | 4.7530 | 2.7145 | 0.0800 |
| intercept 4\|5 | 7.9017 | 3.1286 | 0.0115* |
| intercept 5\|6 | 8.8283 | 3.1926 | 0.0057* |
| intercept 6\|7 | 11.0607 | 3.4043 | 0.0012* |

Responses featured scores of 1, 3, 4, 5, 6, and 7, resulting in five intercept estimates representing the thresholds between each pair of responses on the Likert scale. For this model, the average words per message and the average partner compound sentiment were significantly positively correlated with women's responses to the choice survey item ($p < 0.05$). Additionally, the number of partner messages was significantly negatively correlated with women's responses to the choice survey item ($p < 0.05$).

We also built models for men's responses to the Pressure/Tension, Perceived Competence, and Perceived Choice items from our study. However, none of the models fit the data significantly, according to likelihood ratio Chi-square tests ($p > 0.05$). Therefore, we do not directly compare men's and women's dialogue features, nor do we discuss the relationship between men's dialogue features and their responses to the survey items. Our discussion instead focuses on women's dialogue features and how they relate to women's responses to the survey items.

Table 13. Summary of dialogue feature correlations with women's relaxation, competence, and choice survey items (+ = positive correlation, − = negative correlation, ∅ = no correlation).

| Dialogue Features | Relaxation | Competence | Choice |
|---|---|---|---|
| No. Messages | + | + | ∅ |
| Avg. Words | ∅ | + | + |
| Avg. Compound Sentiment | ∅ | ∅ | ∅ |
| Partner No. Messages | ∅ | − | − |
| Partner Avg. Words | + | ∅ | ∅ |
| Partner Avg. Compound Sentiment | + | + | + |

## 7 DISCUSSION

Of the six dialogue features considered, five were significantly correlated to at least one of the survey items for women: (1) number of messages, (2) average words per message, (3) number of partner messages, (4) average partner words per message, and (5) average partner compound sentiment (see Table 13). In the following subsections, we discuss how each of these features related to their corresponding survey items. The discussion of each feature is accompanied by textual excerpts from the students' collaborative dialogue, presented as illustrative examples to the quantitative results. We remind the reader that these results are in the context of a university in the southeastern United States, and that the participants were primarily White women between the ages of 18 and 21 (see Tables 1 and 3 for more detail).

### 7.1 Number of Messages

According to the ordinal regression models, the number of messages sent by women was positively correlated with their responses to the relaxation survey item and to the competence survey item. The more messages sent by the student, the more relaxed and competent they reported feeling. In an equitable pair programming collaboration, both partners contribute about the same number of messages to the conversation [34]; participants in our study generally adhered to this equity trait, with most sessions featuring a fairly balanced distribution of messages between the two partners. Therefore, a higher number of messages from a student can imply a more active conversation between the partners. Additionally, unique to textual dialogue messages, students would sometimes communicate a thought through several consecutive chat messages. In the excerpt in Figure 3, S1 (woman, high relaxation survey item score) explains her thought process to S2, her male partner, on how to display a game board across multiple messages (typos from original excerpts are preserved).

S1 how should we display a board that doesn't change but that has the insides that change?
S2 maybe we do something with those multi arrays
S1 oh maybe
S1 i was thinking we could do either that or we could like
S1 have a printed board with every space as a variable and then they just replace the variables
S1 bc tbvh im not too hot w arrays
S1 so i'll def need help if we go the array route

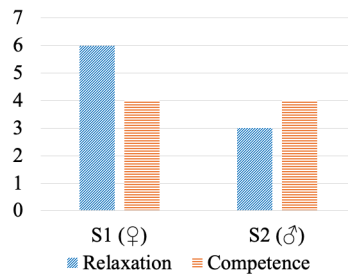


Fig. 3. Excerpt from collaborative programming session and survey item scores for S1 and S2.

It is possible that students who feel more at ease with the collaboration tend to contribute more actively to the conversation while students who feel less relaxed may tend to avoid sending as many messages, perhaps even resorting to sending longer messages. As an example, the excerpt in Figure 4 shows S3 (woman, low relaxation survey item score) communicating with her partner S4 (also a woman) using longer messages instead of splitting a contribution across multiple messages.

S3 do you think the try part is where we are supposed to list all possible wrong inputs, then we specifiy in each catch?
S4 I don't think we have to specify the wrong inputs
S4 nevermind we have to
S3 ahh, is this what you think specifying wrong inputs looks like? I'm assuming that is what it is based off the handout
S4 I think so

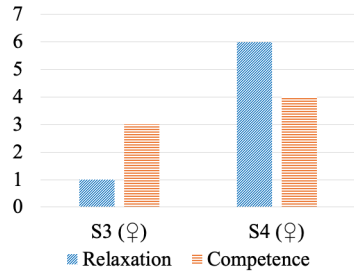


Fig. 4. Excerpt from collaborative programming session and survey item scores for S3 and S4.

### 7.2 Average Words per Message

According to the ordinal regression models, the average words per message sent by women was positively correlated with their responses to the competence survey item and to the choice survey item. The longer the messages sent by a woman collaborator, the more competent she reported feeling and the more likely she felt she had a choice in completing the given activity. Longer messages do not necessarily imply fewer messages, but they could imply fewer instances of short, consecutive messages as described in the previous subsection. With longer messages, ideas are expressed in a single, long message rather than many shorter messages. Shorter messages may not contain all of the details necessary to convey one's thought process, so longer messages could indicate that a collaborator had a concrete idea in their mind and knew exactly how much detail to provide so their partner could understand. Students who reported a high competence survey item score also tended to report a high choice survey item score, which could imply that students who feel more competent also feel more in control of the activity. Figure 5 shows an excerpt in which S5 (woman, high competence and choice survey item scores) explains to her partner S6 (also a woman) how they can use the try/catch block to check for proper user input in the tic-tac-toe program. Her explanations are lengthy, providing detail as to how she thinks they can implement their solution.

### 7.3 Number of Partner Messages

According to the ordinal regression models, the number of messages sent by women's partners was negatively correlated with women's responses to the competence survey item and to the choice survey item. The more messages sent by their partner, the less competent the women reported feeling and the less likely the women felt they had a choice in completing the given activity. This feature is the only feature negatively correlated with any of the survey item responses, and it has the opposite effect of the average words per message. It is possible that by sending a large number of messages, the woman's partner may have appeared to take control of the collaborative programming activity, leading the woman to feel less competent, corresponding with her decreased control of the outcome. In response to the open-ended post-survey question, two participants mentioned they were frustrated when they were in the navigator role due to having a less experienced partner and "itching to do it [themselves]". Similarly, Figure 6 shows an excerpt in which S8 (a male student)

S5 Im in prog 1 as well so I don't have much experience either. I think we need use a try-catch block to read in the user-input and if it's wrong then we do the catch block. Once we have the value from the user we can convert it to index locations and then chose an x or o depending on the turn
S6 would the user input be a number 1-9 for the location?
S5 yeah so i guess we would read in the number and see if it's correct but first we need to make sure it is a number
S5 and not a char or something like that
S6 Ok

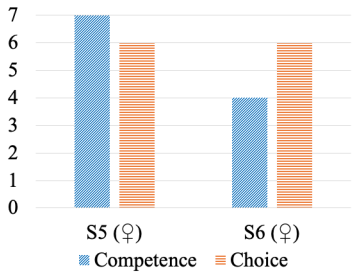


Fig. 5. Excerpt from collaborative programming session and survey item scores for S5 and S6.

as the driver takes control of the conversation on how to search the tic-tac-toe grid by row and column, even completing tasks reserved for the navigator, leaving S7 (woman, low competence survey item score) with nothing to contribute. In this pair, S8 sent almost three times as many messages as S7.

S8 ill just create the method and we can fill it in later
S8 yeah first for should search through each row
S8 second through each column
S7 i just wasnt sure if thats how you structure the code with a 2d array
S8 im looking right now
S8 yeah thats it
S8 we can run it whenever to see its output
S8 just lmk

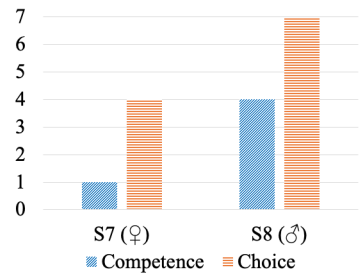


Fig. 6. Excerpt from collaborative programming session and survey item scores for S7 and S8.

### 7.4 Average Partner Words per Message

According to the ordinal regression models, the average words per message sent by women's partners was positively correlated with women's responses to the relaxation survey. The longer the messages sent by the woman's partner, the more relaxed she reported feeling. It is possible that longer partner messages may be preferred by the women, since they receive all of the information they need in a single message, rather than having to wait for their partner to send every message related to a single idea. Relatedly, receiving many messages in rapid succession may lead to feeling less relaxed [59].

It is possible for collaborators to have vastly different experiences, despite having worked together on the same collaborative activity. For example, we previously discussed in Figure 4 how S3 (woman, low relaxation survey item score) communicated with her partner S4 (woman, high relaxation survey item score) using longer messages. While S3's longer messages may be an indicator of her lower relaxation survey item score, they also suggest a higher average words per message. Despite S3's low relaxation survey item score, her higher average words per message may have contributed to S4's high relaxation survey item score, since S4 received all of the information from her partner grouped together rather than in many separate and shorter messages.

### 7.5 Average Partner Compound Sentiment

According to the ordinal regression models, the average compound sentiment of women's partners was positively correlated with all three survey items. The more positive sentiment present in the messages sent by the woman's partner, the more relaxed and competent she reported feeling, and the more likely she felt she had a choice in completing the given activity. Positive sentiment may encompass several related phenomena such as encouragement, which has been associated with positive outcomes [4]. The excerpt in Figure 7 shows S10 (woman, high relaxation survey item score) receiving words of encouragement from her partner S9 (also a woman) stating they both tried their best.

| | | |
|---|---|---|
| S9 | – | idk why the else is acting up |
| S9 | + | hahhah yes |
| S10 | 0 | i don't have a clue |
| S9 | + | oh well we tried our best |
| S10 | 0 | we really did |

*Sentiment polarity:
positive = +, negative = –, neutral = 0

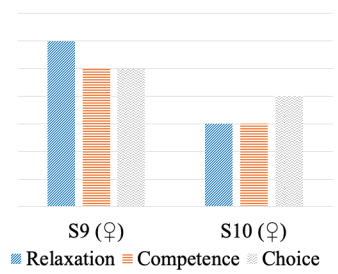


Fig. 7. Excerpt from collaborative programming session and survey item scores for S9 and S10.

Another form of positive sentiment is positive feedback, which has been associated with improved student performance during collaboration [44]. Figure 8 shows an excerpt in which S11 (woman, high competence and choice survey item score) and her partner S12 (a male student) discussed how to divide their tasks, with S12 giving S11 permission to edit his code if she found any errors. When S11 offers a suggestion, S12 expresses gratitude and acknowledges her idea, providing positive feedback by stating that S11's suggestion "sounds good."

| | | |
|---|---|---|
| S11 | + | So do you want to divide the different methods amoung ourselves? There should be 3 different ones we could work on (including main) |
| S12 | + | I tried to create the display for the board using the for loop im not sure if it is right though |
| S12 | + | Feel free to change it and make it correct |
| S11 | 0 | I think you will need to include Arrays.deepToString() so that it is able to display |
| S12 | + | awesome |
| S12 | + | thanks |
| S11 | - | No problem |
| S12 | + | yes dividing methods up sounds good |

*Sentiment polarity:
positive = +, negative = –, neutral = 0

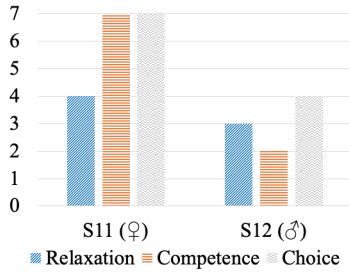


Fig. 8. Excerpt from collaborative programming session and survey item scores for S11 and S12.

### 7.6 Implications

The implications of this study are presented within a particular demographic context. Specifically, the participants were recruited from an introductory computer science course at a large public university in the southeastern United States, and were mainly White and between the ages of 18 and 21. It is important to note that, if situated in another context, this study might produce different results, as this specific context may not apply to students who are more experienced or different participant demographic distributions.

*7.6.1 Separate and Not Equal: Women's Experiences During Remote Collaborative Programming.* The findings reported here indicate a systematic difference between women's and men's experiences during remote collaborative programming. In their responses to the individual items within the IMI subscales, women reported significantly higher levels of stress, lower levels of perceived competence, and less perceived choice compared to men in a survey regarding their collaborative experience. Women experienced more tension during the collaboration, which has the potential to impact the quality of their performance on the programming activity [54]. In a society where remote collaborations are increasingly prevalent, it is imperative for the research community to investigate ways in which historically marginalized groups are experiencing this modality of work differently in many domains, not just computing, and ways in which we can support and empower users.

*7.6.2 Dialogue Features as Early Detectors of Collaborative Experience.* As the field moves toward supporting all users within remote collaborative work, we will need ways to detect whether a collaborator is having a positive experience even as the collaboration is unfolding. The results reported here suggest that the dialogue observed during remote collaborative programming can provide a window into a woman's experience, allowing for early detection of her tension and other outcomes. In fact, real-time language feedback systems have been used to facilitate productive group interactions [55], but have not been developed to support women specifically. By detecting collaborators' outcomes early, we can offer timely support such as scaffolding collaborations to foster the dialogue features that are positively correlated with women's experiences, while potentially reducing dialogue features that are negatively correlated with women's experiences. For example, does fostering the expression of positive sentiment toward the task have a positive impact on women's experience during collaborative problem solving? The correlations found here must be investigated in the context of broader collaborative strategies and roles.

*7.6.3 Dialogue Modalities for Collaborative Programming.* It is important to consider how the system's affordances impact the way users collaborate. From the open-ended responses in the present study, we see evidence that when working synchronously on a dialogue-heavy collaborative problem-solving activity, collaborators might benefit from a non-textual means of communication (voice chat or video chat). For example, all participants were asked about how the software could be improved or if they felt limited by the software, and many (19/58) indicated a preference for communicating via voice (or video) chat, specifying that communicating via text chat was difficult and that communicating via voice chat would be more efficient. Four participants mentioned a preference for being "face-to-face" or physically next to each other. Alternatively, the existing interface with only textual chat might be better suited for cooperative work, like a divide-and-conquer strategy in which team members can communicate and create a task plan prior to using the system, and then use the textual chat for convenient check-ins or clarifications with team members while otherwise working independently. This means of communication may also support social processes during cooperative work. In a response to an open-ended question, one participant from this study wrote, "although talking by voice would be quicker, I felt because there was a message

conversation we got along better at first then if we were face-to-face since we were less nervous." Textual chat may be more conducive to remote collaborations where collaborators are unfamiliar with their partners or team members. In particular, previous work has revealed that strangers meeting for the first time via text-based communication are more likely to express affection [1] and disclose intimate information [21, 58], compared to face-to-face communication.

### 7.7 Limitations

The most notable limitations of the present study involve its sample size and the nature of the experiment revealing only correlational, not causal, relationships. First, as to sample size, the null result regarding difference in structured- versus unstructured-role conditions may be due to low statistical power to reveal those results. Larger-scale studies can shed light on nuanced differences between different collaborative paradigms in a remote context.

Next, with regard to the relationships reported here between dialogue features and women's outcomes, the analyses presented here do not imply causation of women's reported perceptions based on those features. It is possible that the woman's perceptions of stress, competence, and choice are influencing their dialogue moves or those of their partners (rather than the dialogue moves influencing the outcomes). Future research is needed to clarify this relationship. Additionally, the dialogue features used in the model are summative, characterizing the entire collaboration as a whole. Future work needs to consider the temporal aspects of collaboration to understand how collaborations unfold over time and determine at what time, if at all, a system could provide guidance or support to facilitate productive collaboration.

Finally, due to the limited participant population size (See Table 1), we did not investigate outcome differences based on gender composition of the student pairs. Further research with a larger sample size could support analysis of collaboration condition and dialogue features with respect to gender composition in collaborative programming. Additionally, some participants disclosed their name to their partner, which could have resulted in assumptions of gender or racial identity, possibly influencing the partner's behavior. Specifically, one woman had a partner that disclosed his (traditionally male) name, which may have invoked stereotype threat [48]. Future studies could mitigate the assumption of partner identity by telling participants not to disclose their name or other identifying characteristics. Alternatively, future studies could be conducted such that student partners see and/or hear each other before or during their collaborative session to determine the effects of gender composition on the collaboration and the participants' perceptions of the collaboration.

## 8 CONCLUSION

Remote collaboration has become increasingly common in industry and educational settings, with online courses and work becoming more prevalent due to growing technological capabilities. Remote collaborations have unique characteristics and nuances from those of co-located collaborations, and to better support these endeavors we must understand how people experience and perceive their remote collaborations. It is important to consider how individual characteristics may impact these experiences and perceptions, especially to understand and support historically marginalized groups. In this study, we investigated the influence of collaborative paradigms on remote collaborative problem solving in the context of computer science, a field in which women have been historically marginalized. Specifically, participants from a university computer science course coded in pairs, following either structured or unstructured collaborative roles, and we measured outcomes of normalized learning gain, intrinsic motivation, and system usability scores. There were no significant differences between the two collaboration conditions.

Our analysis with respect to gender identity uncovered differences in men's and women's experiences of the activity altogether, regardless of the collaboration condition. We found that women's perceptions of their experiences differed significantly from men's in several aspects. Women on average reported more stress, less perceived competence in their computing abilities, and less perceived choice compared to men. These negative emotions may inhibit the comfort and success of women in the computer science field, and should be researched more extensively to best promote desirable interactions between collaborators and support women during remote collaborative programming. In that vein, we inspected women's collaborative dialogues to gain more insight into their experiences, and our analyses revealed five important relationships between dialogue features and women's outcomes: (1) women's number of sent messages was positively correlated with their reported relaxation and competence; (2) women's average words per message was positively correlated with their reported competence and choice; (3) women's number of received messages was negatively correlated with their reported competence and choice; (4) the average words per message in women's received messages was positively correlated with their reported relaxation; and (5) the average compound sentiment of the women's received messages was positively correlated with their reported relaxation, competence, and choice. These findings hold important implications for the CSCW community and for future work. For example, our study did not find a significant correlation between women's own average compound sentiment and any of their outcomes. This result suggests that despite reporting lower levels of relaxation, perceived competence, and perceived choice, women may not have expressed these emotions and perceptions in their written messages. Future work should delve deeper into the pragmatic and semantic structure of the dialogues, as this investigation may provide insight into the ways in which women are expressing themselves through their dialogue and problem-solving approaches. Additionally, more research is needed to understand the structures and affordances that support all users during remote collaborative problem solving.

## ACKNOWLEDGMENTS

Thanks to the members of the LearnDialogue Group for their help, as well as the reviewers for the time and effort they spent providing suggestions for improvement. Special thanks to Julia Woodward and Kyla McMullen for their feedback on an early version of this paper. This material is based upon work supported by the National Science Foundation under grant CNS-1622438. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

[1] Marjolijn L. Antheunis, Patti M. Valkenburg, and Jochen Peter. 2012. The Quality of Online, Offline, and Mixed-Mode Friendships Among Users of a Social Networking Site. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 6, 3 (2012).

[2] National Education Association. 2012. Preparing 21st Century Students for a Global Society: An Educator's Guide to the "Four Cs". *Alexandria, VA: National Education Association* (2012).

[3] Prashant Baheti, Edward Gehringer, and David Stotts. 2002. Exploring the Efficacy of Distributed Pair Programming. In *Conference on Extreme Programming and Agile Methods*. 208–220.

[4] Kristy Elizabeth Boyer, Robert Phillips, Michael D. Wallis, Mladen A. Vouk, and James C. Lester. 2008. Balancing Cognitive and Motivational Scaffolding in Tutorial Dialogue. In *Proceedings of the International Conference on Intelligent Tutoring Systems (ITS)*. 239–249.

[5] Danielle Bragg, Kyle Rector, and Richard E. Ladner. 2015. A User-Powered American Sign Language Dictionary. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing (CSCW)*. 1837–1848.

[6] John Brooke. 1996. SUS - A quick and dirty usability scale. In *Usability Evaluation in Industry*. 4–7.

[7] Andrzej Buda and Andrzej Jarynowski. 2010. *Life Time of Correlations and Its Applications*. Wydawnictwo Nieżalezne.

[8] Mehmet Celepkolu, Joseph B. Wiggins, Kristy Elizabeth Boyer, and Kyla McMullen. 2017. Think First: Fostering Substantive Contributions in Collaborative Problem-Solving Dialogues. In *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning (CSCL)*. 295–302.

[9] Jan Chong and Rosanne Siino. 2006. Interruptions on Software Teams: A Comparison of Paired and Solo Programmers. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW)*. 29–38.

[10] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

[11] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*. 512–515.

[12] Mona Emara, Ramkumar Rajendran, Gautam Biswas, Mahmod Okasha, and Adel Alsaeid Elbanna. 2018. Do Students' Learning Behaviors Differ when they Collaborate in Open-Ended Learning Environments? *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.

[13] Stephen M. Fiore, Art Graesser, Samuel Greiff, Patrick Griffin, Brian Gong, Patrick Kyllonen, Christine Massey, Harry O'Neil, Jim Pellegrino, Robert Rothman, Helen Soulé, and Alina von Davier. 2017. Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress. (2017).

[14] Alastair J. Gill, Robert M. French, Darren Gergle, and Jon Oberlander. 2008. The Language of Emotion in Short Blog Texts. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW)*. 299–302.

[15] Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. 2007. Expressing Emotion in Text-Based Communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 929–932.

[16] Brian Hanks. 2008. Empirical Evaluation of Distributed Pair Programming. *International Journal of Human-Computer Studies* 66, 7 (2008), 530–544.

[17] Emily M. Hastings, Farnaz Jahanbakhsh, Karrie Karahalios, Darko Marinov, and Brian P. Bailey. 2018. Structure or Nurture? The Effects of Team-Building Activities and Team Composition on Team Outcomes. *Proceedings of the ACM on Human-Computer Interaction (CHI)* 2, CSCW (2018).

[18] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. 216–225.

[19] Henry Jenkins. 2009. *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*.

[20] Patrick Jermann and Marc-Antoine Nüssli. 2012. Effects of Sharing Text Selections on Gaze Cross-Recurrence and Interaction Quality in a Pair Programming Task. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*. 1125–1134.

[21] L. Crystal Jiang, Natalie N. Bazarova, and Jeffrey T. Hancock. 2011. The Disclosure–Intimacy Link in Computer-Mediated Communication: An Attributional Extension of the Hyperpersonal Model. *Human Communication Research* 37, 1 (2011), 58–77.

[22] Malte Jung, Jan Chong, and Larry Leifer. 2012. Group Hedonic Balance and Pair Programming Performance: Affective Interaction Dynamics as Indicators of Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 829–838.

[23] Beaumie Kim. 2001. Social Constructivism. In *Emerging Perspectives on Learning, Teaching, and Technology*, Michael Orey (Ed.). 55–61.

[24] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. 2016. Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies. *PLOS ONE* 11, 8 (2016).

[25] Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, Foad Hamidi, and Stacy M. Branham. 2019. HCI Guidelines for Gender Equity and Inclusivity. Retrieved May 13, 2020 from https://www.morgan-klaus.com/sigchi-gender-guidelines.

[26] Johannes Konert, Henrik Bellhäuser, René Röpke, Eduard Gallwas, and Ahmed Zucik. 2016. MoodlePeers: Factors Relevant in Learning Group Formation for Improved Learning Outcomes, Satisfaction and Commitment in E-learning Scenarios using GroupAL. In *Adaptive and Adaptable Learning: Proceedings of the 11th European Conference on Technology-Enhanced Learning (EC-TEL)*. 390–396.

[27] Brian Kooiman, Wenling Li, Michael Wesolek, and Heeja Kim. 2015. Validation of the Relatedness Scale of the Intrinsic Motivation Inventory through Factor Analysis. *International Journal of Multidisciplinary Research and Modern Education* 1, 2 (2015), 302–311.

[28] Theodora Koulouri, Stanislao Lauria, and Robert D. Macredie. 2017. The Influence of Visual Feedback and Gender Dynamics on Performance, Perception and Communication Strategies in CSCW. *International Journal of Human-Computer Studies* 97 (2017), 162–181.

[29] Amit Kramer, Devasheesh P. Bhave, and Tiffany D. Johnson. 2014. Personality and Group Performance: The Importance of Personality Composition and Work Tasks. *Personality and Individual Differences* 58 (2014), 132–137.

[30] Adam D. I. Kramer, Susan R. Fussell, and Leslie D. Setlock. 2004. Text Analysis as a Tool for Analyzing Conversation in Online Support Groups. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA)*. 1485–1488.

[31] Sandeep Kaur Kuttal, Kevin Gerstner, and Alexandra Bejarano. 2019. Remote Pair Programming in Online CS Education: Investigating through a Gender Lens. In *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 75–85.

[32] Marjan Laal and Seyed Mohammad Ghodsi. 2012. Benefits of Collaborative Learning. *Procedia-Social and Behavioral Sciences* 31 (2012), 486–490.

[33] Cliff Lampe, Rebecca Gray, Andrew T. Fiore, and Nicole Ellison. 2014. Help is on the Way: Patterns of Responses to Resource Requests on Facebook. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work Social Computing (CSCW)*. 3–15.

[34] Colleen M. Lewis and Niral Shah. 2015. How Equity and Inequity Can Emerge in Pair Programming. In *Proceedings of the 11th Annual Conference on International Computing Education Research (ICER)*. 41–50.

[35] James R. Lewis and Jeff Sauro. 2018. Item Benchmarks for the System Usability Scale. *Journal of Usability Studies* 13, 3 (2018), 158–167.

[36] Xing Liu and Hari Koirala. 2012. Ordinal Regression Analysis: Using Generalized Ordinal Logistic Regression Models to Estimate Educational Data. *Journal of Modern Applied Statistical Methods* 11, 1 (2012), 242–254.

[37] Michael Madaio, Kun Peng, Amy Ogan, and Justine Cassell. 2018. A Climate of Support: A Process-Oriented Analysis of the Impact of Rapport on Peer Tutoring. In *Proceedings of the 13th International Conference of the Learning Sciences (ICLS)*. 600–607.

[38] Tara Matthews, Jalal U. Mahmud, Jilin Chen, Michael Muller, Eben Haber, and Hernan Badenes. 2015. They Said What?: Exploring the Relationship Between Language Use and Member Satisfaction in Communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work  Social Computing (CSCW)*. 819–825.

[39] Hanna Maurin, Diane H. Sonnenwald, Bruce Cairns, James E. Manning, Eugene B. Freid, and Henry Fuchs. 2006. Exploring Gender Differences in Perceptions of 3D Telepresence Collaboration Technology: An Example from Emergency Medical Care. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles (NordiCHI)*. 381–384.

[40] Edward McAuley, Terry Duncan, and Vance V. Tammen. 1989. Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis. *Research Quarterly for Exercise and Sport* 60, 1 (1989), 48–58.

[41] Peter McCullagh. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society: Series B (Methodological)* 42, 2 (1980), 109–127.

[42] Charlie McDowell, Linda Werner, Heather Bullock, and Julian Fernald. 2002. The Effects of Pair-Programming on Performance in an Introductory Programming Course. In *Proceedings of the 33rd ACM Technical Symposium on Computer Science Education (SIGCSE)*. 38–42.

[43] Meike Osinski and Nikol Rummel. 2019. Towards Successful Knowledge Integration in Online Collaboration: An Experiment on the Role of Meta-Knowledge. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019).

[44] Jennifer Robison, Scott McQuiggan, and James C. Lester. 2009. Evaluating the Consequences of Affective Feedback in Intelligent Tutoring Systems. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–6.

[45] Fernando J. Rodríguez, Kimberly Michelle Price, and Kristy Elizabeth Boyer. 2017. Expressing and Addressing Uncertainty: A Study of Collaborative Problem-Solving Dialogues. In *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning (CSCL)*. 207–214.

[46] Jeff Sauro and James R. Lewis. 2016. *Quantifying the User Experience: Practical Statistics for User Research*.

[47] Linda J. Sax, Kathleen J. Lehman, Jerry A. Jacobs, M. Allison Kanny, Gloria Lim, Laura Monje-Paulson, and Hilary B. Zimmerman. 2017. Anatomy of an Enduring Gender Gap: The Evolution of Women's Participation in Computer Science. *The Journal of Higher Education* 88, 2 (2017), 258–293.

[48] Toni Schmader. 2012. *Stereotype Threat: Theory, Process, and Application*. Oxford University Press.

[49] Andrew J. Scholand, Yla R. Tausczik, and James W. Pennebaker. 2010. Social Language Network Analysis. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)*. 23–26.

[50] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of Social Identity Theory to Research and Design in Computer-Supported Cooperative Work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018).

[51] Lauren E. Sherman, Minas Michikyan, and Patricia M. Greenfield. 2013. The Effects of Text, Audio, Video, and In-Person Communication on Bonding Between Friends. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 7, 2 (2013).

[52] Angela E.B. Stewart, Hana Vrzakova, Chen Sun, Jade Yonehiro, Cathlyn Adele Stone, Nicholas D. Duran, Valerie Shute, and Sidney K. D'Mello. 2019. I Say, You Say, We Say: Using Spoken Language to Model Socio-Cognitive Processes during Computer-Supported Collaborative Problem Solving. *Proceedings of the ACM on Human-Computer Interaction (CHI)* 3, CSCW (2019), 1–19.

[53] Henri Tajfel. 1974. Social Identity and Intergroup Behaviour. *Information (International Social Science Council)* 13, 2 (1974), 65–93.
[54] Chiew Seng Sean Tan, Johannes Schöning, Kris Luyten, and Karin Coninx. 2014. Investigating the Effects of Using Biofeedback as Visual Stress Indicator during Video-Mediated Collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 71–80.
[55] Yla R. Tausczik and James W. Pennebaker. 2013. Improving Teamwork Using Real-Time Language Feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 459–468.
[56] Divy Thakkar, Nithya Sambasivan, Purva Kulkarni, Pratap Kalenahalli Sudarshan, and Kentaro Toyama. 2018. The Unexpected Entry and Exodus of Women in Computing and HCI in India. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI)*. 1–12.
[57] David Thissen, Lynne Steinberg, and Daniel Kuang. 2002. Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *Journal of Educational and Behavioral Statistics* 27, 1 (2002), 77–83.
[58] Patti M. Valkenburg and Jochen Peter. 2009. Social Consequences of the Internet for adolescents: A Decade of Research. *Current Directions in Psychological Science* 18, 1 (2009), 1–5.
[59] Jan M. van Bruggen, Paul A. Kirschner, and Wim Jochems. 2002. External Representation of Argumentation in CSCL and the Management of Cognitive Load. *Learning and Instruction* 12, 1 (2002), 121–138.
[60] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How Data Scientists Use Computational Notebooks for Real-Time Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[61] Armin Weinberger, Karsten Stegmann, and Frank Fischer. 2007. Knowledge Convergence in Collaborative Learning: Concepts and Assessment. *Learning and Instruction* 17, 4 (2007), 416–426.
[62] Linda L. Werner, Brian Hanks, and Charlie McDowell. 2004. Pair-Programming Helps Female Computer Science Students. *ACM Journal of Educational Resources in Computing* 4, 1 (2004), 1–8.
[63] Laurie Williams and Richard L. Upchurch. 2001. In Support of Student Pair-Programming. In *Proceedings of the 32nd ACM Technical Symposium on Computer Science Education (SIGCSE)*. 327–331.
[64] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 1-3 (1987), 37–52.
[65] Kimberly Michelle Ying, Lydia G. Pezzullo, Mohona Ahmed, Kassandra Crompton, Jeremiah Blanchard, and Kristy Elizabeth Boyer. 2019. In Their Own Words: Gender Differences in Student Perceptions of Pair Programming. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE)*. 1053–1059.
[66] Jaebong Yoo and Jihie Kim. 2014. Can Online Discussion Participation Predict Group Project Performance? Investigating the Roles of Linguistic Features and Participation Patterns. *International Journal of Artificial Intelligence in Education* 24, 1 (2014), 8–32.
[67] Chien Wen (Tina) Yuan, Yu-Hsuan Liu, Hao-Chuan Wang, and Yuan-Chi Tseng. 2019. Gender Effects on Collaborative Online Brainstorming Teamwork. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA)*. 1–6.
[68] Mark Zarb, Janet Hughes, and John Richards. 2013. Industry-Inspired Guidelines Improve Students' Pair Programming Communication. In *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE)*. 135–140.
[69] Kristina M. Zosuls, Cindy Faith Miller, Diane N. Ruble, Carol Lynn Martin, and Richard A. Fabes. 2011. Gender Development Research in Sex Roles: Historical Trends and Future Directions. *Sex Roles* 64, 11-12 (2011), 826–842.
[70] Stuart Zweben and Betsy Bizot. 2019. 2018 Taulbee Survey: Undergrad Enrollment Continues Upward; Doctoral Degree Production Declines but Doctoral Enrollment Rises. *Computing Research News* 31, 5 (2019), 3–74.
[71] Stuart Zweben, Jodi Tims, and Yan Timanovsky. 2019. ACM-NDC Study 2018–2019: Seventh Annual Study of Non-Doctoral-Granting Departments in Computing. *ACM Inroads* 10, 3 (2019), 40–54.

Received June 2020; accepted July 2020