Using Dialogue Analysis to Predict Women's Stress During Remote Collaborative Learning in Computer Science

Kimberly Michelle Ying
kimying@ufl.edu
University of Florida
Gainesville, Florida, USA

Gloria Ashiya Katuka
gkatuka@ufl.edu
University of Florida
Gainesville, Florida, USA

Kristy Elizabeth Boyer
keboyer@ufl.edu
University of Florida
Gainesville, Florida, USA

ABSTRACT

The computer science education community strives to improve equity and representation within the field, yet the proportion of women earning CS bachelor's degrees in countries such as the US remains low. In addition to recruitment and retention initiatives that support women, we need to better understand women's experiences within CS. This paper makes a novel contribution toward this effort by examining women's self-reported stress during remote collaborative programming with a peer. Women reported significantly more stress than men, so we analyzed the women's collaborative dialogues and identified the most common dialogue acts and sequences of dialogue acts. We used these dialogue acts to predict women's stress and found six significant patterns of dialogue. Women reported less stress with higher frequencies of offering suggestions, having their partner provide explanations, and having their own rapport-building messages reciprocated by their partner. In contrast, women reported more stress with higher frequencies of their own explanations, having their partner answer their questions, and having their partner send a rapport-building message that they reciprocated. Understanding the nuances of these experiences allows us to make better predictions of when women might be feeling stressed and what we might be able to do to relieve these feelings. Improving women's CS experiences holds the potential to, in turn, improve gender equity within CS.

CCS CONCEPTS

• **Social and professional topics** → **CS1**; **Women**; • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

KEYWORDS

Gender; Women; CSCL; CSCW; CS1; Collaboration

ACM Reference Format:
Kimberly Michelle Ying, Gloria Ashiya Katuka, and Kristy Elizabeth Boyer. 2021. Using Dialogue Analysis to Predict Women's Stress During Remote Collaborative Learning in Computer Science. In *26th ACM Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2021), June 26-July 1, 2021, Virtual Event, Germany.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3430665.3456369

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ITiCSE 2021, June 26-July 1, 2021, Virtual Event, Germany
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8214-4/21/06…$15.00
https://doi.org/10.1145/3430665.3456369

1 INTRODUCTION

Diversity and inclusion initiatives have become important aspects of computer science education research to tackle the disparities in the field [12, 24]. Despite ongoing efforts to reduce the gender gap in computing, women continue to be marginalized in many countries around the world. In the United States specifically, just 20% of CS bachelor's degrees were earned by women during the 2017-2018 academic year [22]. In addition to targeted recruitment and retention efforts, we need to strive to understand women's learning experiences. In this paper, we investigate women's experiences working remotely with a partner on a Java programming assignment.

This area of research is particularly relevant now due to the current COVID-19 pandemic for several reasons. First, the pandemic has resulted in many women leaving the workforce, exacerbating the gender gap and reversing some of the prior progress [29, 34]. Second, adjusting to the pandemic has meant that many professionals and students are now working and learning remotely. Lastly, this pandemic may have permanently changed the landscape and culture of remote learning and remote work, especially in computing-related fields [9].

This article describes women's experiences during collaborative learning in CS and reports a model of their dialogues that reveal which dialogue patterns are associated with women's reported stress. In a study of remote collaborative programming with CS1 students, we found that women self-reported significantly more stress than men [37]. This finding led to the research question, **what patterns of dialogue are associated with women's reported stress during remote collaborative coding?** By examining women's experiences during the process of collaborative CS learning, we can identify factors that are beneficial for women as well as those that may be harmful. These efforts, in turn, can inform appropriate interventions and the design of more gender-equitable collaborative learning activities for computer science.

2 RELATED WORK

Collaborative learning has become widely implemented in CS education, supporting long-standing educational goals, such as active learning and developing communication skills [33]. Pair programming is one such collaborative learning paradigm, which resulted in better learning gains and problem-solving skills for undergraduate students [10, 19, 20]. Pair programming allows learners to take on structured roles of *driver* or *navigator*, in which the *driver* controls the mouse and keyboard, and the *navigator* observes and provides

suggestions and feedback to the driver while they jointly produce a coding artifact [32]. In addition to improving learning gains, pair programming results in students producing higher quality code and acquiring new programming skills [21].

Pair programming has traditionally involved co-located learners working together on the same computer, but recently remote pair programming has become a popular alternative in which students reap many of the same benefits [35, 38]. While pair programming is especially beneficial for women, with the potential to help reduce the gender gap [31], remote pair programming has only recently been investigated through a gender lens [16]. Through qualitative analysis of six pairs of students, Kuttal et al. [16] found several gender differences in students' collaborative behaviors. One notable finding was the preference for pair programming roles, with women preferring to be the navigator when they *knew* how to solve the task and men preferring to be the navigator when they *did not know* what to do. In that study, students remotely pair programmed and collaborated with their partner through video conferencing, whereas in the study reported in this paper students collaborated textually through a chat messaging client within the programming interface.

Communication is an important part of how collaborators work together during pair programming [4]. When two students communicate in a collaborative learning context, such as remote programming, they actively engage in a dialogue with their partner, working toward a solution and gradually building rapport. Rapport-building and its effects on learning have been studied in the context of group work [1], peer tutoring [23, 28], and pedagogical agents [15]. Previous work suggests that rapport can positively impact learning [15, 28]. In this paper, we identify instances of rapport-building between dyads of CS1 students during the collaborative activity. In addition to its potential impacts on learning, we were also interested in identifying rapport within the dialogues because previous research suggests that collaborative programming is particularly beneficial to women due to its social aspects [31, 36].

The dialogue exchanges between students during collaborative programming provide rich information about the intentions of each speaker. An utterance is a unit of communication that expresses a speaker's intent [3]. To capture a higher-level representation of an utterance, dialogue act labels are used to express the nature of the communicative behavior between the speakers [5]. Previous studies have analyzed dialogue patterns using dialogue act classification techniques to examine the relationship between dialogue acts and learning outcomes during pair programming [27]. Recent studies have employed machine learning techniques to classify dialogue acts, creativity stages, and current roles of the students during remote pair programming for future use in automating a facilitator agent [26]. While that study involved modeling 18 participants evenly distributed to include mixed and same-gender pairs, this paper specifically models the collaborations from the perspective of women. We labeled and analyzed collaborative dialogues of women engaged in remote collaborative programming to provide further insight on how dialogue patterns are associated with women's experiences of stress. Stress is an important affective measure to consider because it often indicates "susceptibility to failure" and can cause attrition in STEM careers [39].

3 COLLABORATIVE CODING STUDY

3.1 Participants

This study was conducted in Spring 2019, and participants were recruited from a CS1 course at a large public university in the southeastern United States. Participation occurred toward the end of the semester, after the last class meeting, but before students' final exam. The study was conducted outside of class hours and students earned two percentage points of extra credit toward their final grade as compensation for their participation.[1] There were 58 total participants in the study, 24 women and 34 men,[2] although the analysis reported here focuses on the women's perspectives. The racial/ethnic breakdown of the 58 participants were 31 (53%) White/Caucasian, 10 (17%) Hispanic/Latino, 10 (17%) Asian/Pacific Islander, four (7%) Black/African-American, and three (5%) multiracial. Of the women specifically, there were 16 White/Caucasian, two Hispanic/Latino, four Asian/Pacific Islander, and two Black/African-American. Most participants (90%) were between the ages of 18 and 21, with five participants indicating they were between 22 and 27, and one participant indicating they were '28 or older.' The women in particular were all between 18 and 21 years old.

3.2 Study Context

Participants were assigned to one of six study sessions according to their availability. For each session, the study room was set up with 14 workstations and arranged so participants would not have a direct view of other participants' screens. Participants were paired (unknowingly) in the order they arrived at the study room by sending them to each workstation accordingly. Paired participants were always seated in separate rows and never in adjacent spots. As part of a larger study, half the study sessions required participants to pair program, while the other half did not require participants to follow any specific collaboration paradigm. To simulate a remote working condition and limit collaboration to the remote work space, participants were not told who they were paired with at any point during the study. Due to latecomers and absentees, there were between four and six pairs for each study session, with 29 pairs in total (12 man-man pairs, 10 woman-man pairs, and seven woman-woman pairs). Every participant was only involved in one of the aforementioned pairs.

Participants first completed a pre-test, then had approximately one hour to collaborate remotely on a coding task, and finally completed a post-test and post-survey on their collaboration experience. The post-survey included items from the validated Intrinsic Motivation Inventory, which measures interest/enjoyment, perceived competence, effort/importance, pressure/tension, perceived choice, value/usefulness, and relatedness [14, 18]. The items are on a 7-point-Likert scale from *1 (not at all true)* to *7 (very true)*. The pre- and post-test were identical and covered the *try-catch* coding construct, which was not taught during their CS1 class. The coding task was to create a Java program to allow two people to play tic-tac-toe and to implement a try-catch construct within the program to handle erroneous user input. Participants received a one-page reference

[1] This study was one of three options students could choose from to earn extra credit.
[2] Participants were asked demographic information at the very end of the study. All participants identified as either 'female' or 'male;' there were also options of 'prefer to self describe' with a text field to enter their gender identity, or 'prefer not to say.'

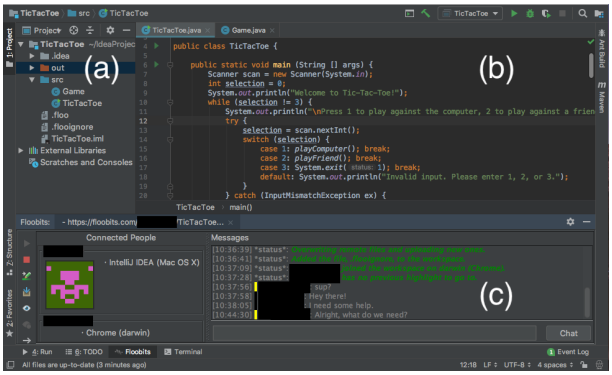

Figure 1: IntelliJ with Floobits plugin: (a) file explorer, (b) synchronous collaborative coding editor, (c) chat messenger

sheet with a brief description and example of the try-catch coding construct for use only during the collaborative task. This reference sheet was not available to them during either the pre- or post-test. Participants programmed using the IntelliJ environment with a plugin (Floobits) that allowed them to collaboratively code in real-time and provided them a built-in messenger for communicating textually (see Figure 1). The Floobits plugin is open-source and we modified it to log participants' chat messages.

4 DATA ANALYSIS

The analysis reported in this paper is part of a larger study. A previous analysis on this dataset has been published [37] and those findings motivate the additional dialogue analysis described in this paper.[3] We summarize those findings in this paragraph. Overall, students (n=58) had significant normalized learning gains as calculated from their pre- and post-tests using one-sample Wilcoxon signed rank tests (p<0.01). Splitting by gender, both the women and men had significant learning gains (p<0.01), and there were no significant differences between these populations on learning gain (p»0.05). Despite both women and men having favorable learning gains, on the post-survey, women reported significantly lower levels of perceived competence, lower levels of perceived choice, and higher levels of pressure/tension. For the remainder of the paper, to model women's stress, we focus specifically on the survey item "*I was very relaxed in doing these*," which is part of the pressure/tension subscale. On this item, women had an average Likert response of 3.67 (σ=1.49), while men had an average Likert response of 4.94 (σ=1.25). This difference was statistically significant according to a Wilcoxon rank sum test, with a p-value of 0.0024 and a large effect size (d=0.68) according to Cohen's d [7]. For the remainder of the paper, we refer to this survey item as the stress item and reverse score the survey responses for modeling.

The previously published analysis did not investigate dialogue acts during collaborative remote computer science learning. Both the dialogue act classification reported here, and the model using those dialogue acts (and their n-grams) to predict women's stress are novel contributions of this paper. To the best of the authors' knowledge, this paper is the first to use dialogue act analysis to predict women's stress during remote collaborative coding.

[3]There were no significant differences found between the two conditions (pair programming versus freestyle collaboration) on learning gain or any survey items (p»0.05).

Table 1: Dialogue Act Classification Scheme

Tag	Dialogue Act (DA)	Examples from the Corpus
SU	Suggestion or Directive	*maybe use a for loop to initialize* | *lets test what we have now*
SA	Statement of Action	*I am reading the javadocs now* | *im trying to set up if a game is won*
FP	Positive Feedback	*nice nice nice* | *omg it works now* | *ok cool we're getting somewhere*
FN	Non-Positive Feedback	*Yeah, I don't think that's right lol* | *oh yikes!* | *hmmm that didnt work*
Q	Question	*where would they be added though?*
A	Answer to a Question	*yeah I think the try-catch block is done* | *dont see why not*
U	Uncertainty	*um not sure yet* | *but idk how*
ACK	Acknowledgement	*thanks* | *ok* | *gotcha* | *yea* | *no problem*
E	Explanation	*the red means it wont work*
MNV	Meta or Non-Verbal Comment	*uhmmmmmmmmmm* | :3 | *pftsh* | *haha* | *wait*
RB	Rapport-Building	*im in the same boat honestly*

4.1 Dialogue Classification

We extracted all the chat logs from the collaborative coding sessions that involved at least one female collaborator and manually tagged them using a dialogue act classification scheme, as shown in Table 1. There were seven woman-woman pairs and 10 woman-man pairs, resulting in 17 dialogues to be tagged. The tagging scheme was created by adapting an existing tagging scheme by Rodríguez et al. [27], which they used to annotate textual dialogues in the context of remote collaborative block-based programming among introductory CS students.

We made some modifications to Rodríguez's scheme by combining some more granular tags into higher-level tags. Specifically, *directives* and *suggestions* were combined into one dialogue act, and we tagged questions and answers without specifying whether they were closed (yes/no) or open-ended. We modified the *meta* dialogue act to also include non-verbal comments like emoticons. Our *explanation* dialogue act is synonymous with the *statement* dialogue act in Rodríguez's scheme. The *rapport-building* dialogue act has some overlap with Rodríguez's *off-task* dialogue act; however, the rapport-building tag can indicate statements of encouragement, opinions about the task or themselves, as well as as general (off-task) socializing. Lastly, we included the *statement of action* dialogue act to distinguish from general explanations and indicate when the student was describing what they were doing, what they just did, or what they were about to do.

Two researchers independently tagged the corpus, and the dialogue scheme was iteratively refined until there was substantial agreement between them. The researchers labeled each sentence with one dialogue act tag. Before tagging, one researcher manually split messages that included more than one sentence so that

appropriate dialogue act tags could be assigned to each part of the message. The tagging process was as follows: First, two researchers collaboratively tagged one conversation for training and initial refinement of the tagging scheme. Then, for each iteration the two researchers independently tagged 20% of the data, kappa (a measurement of agreement) was calculated, discrepancies were resolved, and the dialogue scheme and tag prioritization rules were refined. After two iterations on unique sets of data, the kappa reached 0.633, which indicates substantial agreement [17]. One of the two researchers tagged the remainder of the corpus.

4.2 Dialogue Acts and *n*-grams

From the 17 collaborative dialogues, there were 1292 total chat messages sent. After splitting messages containing multiple sentences, there were 1363 total utterances which were tagged with their corresponding dialogue act. The shortest conversation had 13 messages and 14 tags, while the longest conversation had 166 messages and 171 tags. On average, the conversations had 76 messages and 80 tags. Shorter conversations were generally indicative of a pair following mostly a divide-and-conquer approach, while longer conversations were common for those following a pair programming strategy.

After dialogue act tagging, *n*-grams were compiled from the corpus, following standard protocol from previous dialogue research by Forbes-Riley and Litman [8]. In this analysis, *n*-grams are sequences of dialogue act tags, and we extracted all unigrams (n=1), bigrams (n=2), and trigrams (n=3) from the corpus using a sliding window approach.[4] To ensure our analysis was student-centered, we compiled these *n*-grams from the perspective of each student, and we also included subscripts on the tags to indicate whether the message came from that student or their partner. Because we are concerned with understanding women's experiences in particular, for our modeling we only included *n*-grams that were from the perspectives of women, meaning any tag with the *student* subscript is a tag representation of a message that a woman sent during her collaboration. Tags with the *partner* subscript, however, could be from either a woman or a man, depending on who the woman was collaborating with. For example, the bigrams extracted from the dialogue excerpt in Table 4 from S3's perspective were {$\mathbf{E}_{stu}$, $\mathbf{E}_{stu}$}, {$\mathbf{E}_{stu}$, $\mathbf{RB}_{stu}$}, {$\mathbf{RB}_{stu}$, $\mathbf{SA}_{par}$}, and {$\mathbf{SA}_{par}$, $\mathbf{ACK}_{stu}$}.

The total number of unique *n*-grams extracted from only the women's perspectives was 1731, consisting of 22 unigrams, 378 bigrams, and 1331 trigrams. Before modeling, we reduced the number of *n*-grams provided as features to the model by excluding all *n*-grams that occurred in less than half of the conversations from the women's perspectives. In other words, if an *n*-gram was not present in at least 12 of the 24 women's conversations, it was removed. After this reduction, there were 30 total unique *n*-grams that would be considered by the model, 19 unigrams and 11 bigrams.

4.3 Modeling

We modeled women's experiences using the best subset method for generalized regression with the JMP Pro statistical software [13]. The 30 *n*-grams identified previously were used as features in the model and the women's responses to the stress item were used as the outcome metric for prediction. The best subset method uses an exhaustive algorithm that assesses all possible models with the given features and chooses the one with the best fit based on the selected goodness-of-fit measure; we used the Akaike information criterion (AIC) in this case. Using AIC and R^2, the best subset method chooses the model that explains the most variance with the fewest possible number of predictors.[5]

Table 2: Generalized regression model for women's stress (R^2=0.697) with parameter estimates (Est.) and standard error (Std Err) for centered and scaled predictors.

Dialogue Act *n*-gram	Est.	Std. Err.	*p*-value
Intercept	4.333	0.165	<0.0001
$\mathbf{Rapport}_{student}$, $\mathbf{Rapport}_{partner}$	-9.065	1.552	<0.0001
$\mathbf{Suggestion}_{student}$	-6.335	0.936	<0.0001
$\mathbf{Explanation}_{partner}$	-4.920	0.976	<0.0001
$\mathbf{Answer}_{partner}$	3.742	0.908	<0.0001
$\mathbf{Explanation}_{student}$	5.308	1.417	0.0002
$\mathbf{Rapport}_{partner}$, $\mathbf{Rapport}_{student}$	11.448	1.628	<0.0001

5 RESULTS

The generalized regression model using the best subset method resulted in six *n*-grams as predictors for the women's self-reported stress. Table 2 shows the resulting model with the *n*-gram predictors centered and scaled estimates, along with their standard error, and *p*-value, showing statistical significance. The model has an R^2 of 0.697, which indicates that the model explains nearly 70% of the variance in this dataset. Three *n*-grams are negatively correlated with women's stress, while three *n*-grams are positively correlated with women's stress. Specifically, higher frequencies of rapport-building messages from their partner followed by the woman reciprocating rapport-building messages, higher frequencies of the woman explaining, and higher frequencies of partner answers are predictive of the woman reporting more stress. On the other hand, higher frequencies of her own rapport-building messages being reciprocated by her partner, higher frequencies of her own suggestions, and higher frequencies of explanation messages by her partner, are predictive of the woman reporting less stress.

6 DISCUSSION

From these results, we focus our discussion on the dialogue acts initiated by the women, in other words the dialogue acts in our model that have the student subscript. All example excerpts provided are shown verbatim from the corpus, including keeping typos and capitalization as it was originally written. These excerpts are included in the discussion to provide context for interpreting the model; a formal qualitative analysis is left for future work.

6.1 Rapport-Building (RB)

There were 148 rapport-building messages in the corpus. Further annotations revealed that 63 were an opinion or of a self-deprecating

[4]We created a public GitHub repository to share the scripts we used for generating the n-grams and preparing the data for modeling: https://github.com/LearnDialogue/N_gram_Gen_Dialogue_Analysis

[5]Due to the small dataset (n=24), we capped the number of possible predictors at six to avoid overfitting the model [11].

nature ("*uh im gonna be honest i'm not the most proficient programmer*"), 58 were more general off-task socializing ("*haha 9:20 am is only good for drinking coffee*"), and 27 were of an encouraging, supportive, or motivational nature ("*i have faith in u*"). According to the model, higher frequencies of a woman sending rapport-building messages that were subsequently reciprocated by her partner was predictive of her reporting less stress. There were 39 instances of this bigram. Table 3 shows an example of a messaging exchange between a woman (S1) and her male partner (S2). This woman was one of four who reported their stress level as a two on the Likert scale out of seven, the lowest score any woman reported for stress. Although she may not have been confident during the exercise, her partner's willingness to admit that he was also struggling may have put her at ease.

Table 3: Reciprocated Rapport-Building Example

ID	Message	DA
S1	im kinda clueless ahaha	RB
S2	im in the same boat honestly	RB

In contrast, the excerpt shown in Table 4, which is from a woman-woman pair, shows a potential missed opportunity to connect and offer reassurance. Both women in this pair had high stress scores, with S4 reporting a seven (the highest possible score) and S3 reporting a six. S3 explains a mistake she thinks they made and then follows it up with a self-deprecating remark. Although her partner acknowledges the mistake and responds by saying she will fix it, she does not offer any sympathy or try to lighten the mood.

Table 4: Unreciprocated Rapport-Building Example

ID	Message	DA
S3	we also like	E
S3	forgot get/set methods which i think might help	E
S3	omg im so dumb	RB
S4	ok i can do some of those	SA
S3	kk	ACK

Tickle-Degnen and Rosenthal [30] define three aspects of rapport: (1) *mutual attentiveness* is the "feeling as one", (2) *positivity* is "mutual friendliness and caring", and (3) *coordination* is "predictability and equilibrium." In these two opposing examples, Table 3 shows mutual attentiveness and positivity between the pair, while Table 4 shows some mutual attentiveness by taking initiative on the task, but less positivity because the self-deprecating remark is dismissed in a simple acknowledgment instead of through a caring response.

6.2 Suggestions (SU)

Most of the SU utterances were suggestions, with very few directives. Specifically, out of the 223 SU dialogue acts in the corpus, 201 were suggestions and 22 were directives. There were 157 SU dialogue acts sent by women. The model shows that the more the student makes suggestions, the less likely she will report stress. Suggestions occurred most frequently in batches with the student sending an initial suggestion and then following up with more ideas. It was also relatively common for students to offer suggestions after their partner made a suggestion, as shown in Table 5. The woman in this pair (S5) had a low stress score (2). These instances of dense suggestion activity may indicate that the students feel comfortable expressing their thoughts and are able to build on each other's ideas. Table 6 shows an excerpt from a woman-woman pair with two instances of idea building. From this pair, one woman (S8) had a relatively low stress score (3), although her partner (S7) had a relatively high stress score (5).

Table 5: Suggestions Example

ID	Message	DA
S5	I feel like we should add more methods instead of just doing everything in this one	SU
S6	Sure, what methods do we need?	Q
S6	I think me might need Player, turn, and something that check if the game ended	SU
S5	yeah and an initializeBoard too probably	SU
S5	and then we can just run everything in our driver	SU

Table 6: Suggestions Example 2

ID	Message	DA
S7	we should prob start with an int array of two dimensions	SU
S7	we can write the array like	SU
S7	int board[] = new int[3][3];	SU
S7	i dont know how we would print the board in the begining	U
S8	I'm thinking print the values in a box with the []	SU
...	...	...
S7	okay so should we initalize the array first?	SU
S8	yea maybe with all 0's	SU
S7	hmmm that didnt work	FN
S8	maybe use a for loop to initialize	SU
S8	oh we gotta put that in its own method i think	SU

6.3 Explanations (E)

The explanation dialogue act tag was used for messages of explanation, elaboration, and information. It had a lower priority in the tagging protocol, meaning if an explanation message also fit another dialogue act, the other dialogue act was prioritized and used as the final tag. For example, if a student answered their partner's question by providing information, that message would be annotated as an *answer* rather than an *explanation*. There were 150 $Explanation_{student}$ unigrams and 171 $Explanation_{partner}$ unigrams in the corpus. Higher frequencies of explanations from the students were predictive of that student reporting more stress.

If a student does a lot of explaining and elaborating to her partner, it could be a sign that the learners are struggling to build common ground. In dialogue, *grounding* refers to the interaction between people to establish a common ground [6]. Grounding is important

because it leads to mutual understanding [2], and it is particularly vital for success in collaborative problem solving [25].

Table 7 shows an example excerpt from a woman-woman pair where S9 offers several explanations and suggestions to her partner, and her partner (S10) seems to struggle to keep up. After several tries, she requests taking over writing the code and her partner is happy to oblige. This pair of women reported opposite levels of stress, with S9 reporting the highest possible stress score (7) and S10 reporting among the lowest stress (2).

Table 7: Explanations Example

ID	Message	DA
S9	So it looks like you're the driver right now.	E
S9	This try catch is similar to if else.	E
S9	Should you do a system.out.println for the board to be printed	SU
S10	that what i was thinking but I not sure where to start	U
S9	or String stringBoard = "__| " to represent one cell of the board and a for loop?	SU
S9	for (i=0;i<row; i++) then a nested loop for(j=0;j<column;j++) return stringBoard	SU
S9	for (i=0;i<row;i++) nested loop for(j=0;j<column;j++) return stringBoard	SU
S9	sorry, this chat box won't let me type code.	E
S9	She wnats us to print a real board though	E
S10	oh your fine, I thought she said we did not have to	U
S9	Same lol!	U
S9	can I show you what I was thinking?	Q
S10	yes please	A

Looking at the model on a higher-level, when a student's messages include a lot of explanations but not a lot of suggestions, it may be an indication of inaction or indecision. This might occur when a woman is trying to make sense of things by sending messages about what she knows, but she might be struggling to determine how to contribute to the code or problem solving.

7 IMPLICATIONS

CS instructors should be aware that students, especially those from historically marginalized groups, may be having negative learning experiences even when grades or learning gains suggest otherwise. Instructors can collect affective feedback from their students on assignments through brief surveys and use this information to determine whether assignments are enjoyed equitably or whether some might be disproportionately favorable to certain groups and harmful to others. When supervising collaborative coding between students, instructors might consider intervening when (1) the student is giving a lot of explanations but not forming a lot of suggestions, and (2) when her partner is providing a lot of answers but not providing a lot of explanations. Instances of rapport-building between women and their partners are particularly interesting because the model indicates that when initialized by the woman and reciprocated by her partner, it is predictive of less stress, but when her partner instead initializes the rapport-building that she subsequently reciprocates, it is predictive of greater stress. These occurrences warrant deeper investigation. Gathering additional collaborative dialogue data to expand the corpus may allow us to use the more granular annotations of rapport-building (self-deprecation/opinion, encouraging/supportive/motivational, off-task/socializing) to better model women's stress and understand these nuances.

Limitations. This modeling was based on a small set of data: 17 total dialogues, and 24 women's perspectives. The model will be more robust with additional data and a larger population of participants. Additionally, the model results should be interpreted in light of the context, a CS1 student population from a large public research university in the southeastern United States. This model, while predictive, does not imply causality. The analysis presented here is exploratory and further research is needed to determine both generalizability to new contexts as well as establish whether any of the predictors in the model have a causal effect. Student stress levels were self-reported after the collaborative coding task. While the survey items were specifically in the context of the task, it is possible that some students (or women in particular) are more stressed in general.

8 CONCLUSION

The gender gap in computer science is a pressing issue. To progress toward gender-equity within tech, we need to not only recruit and prepare women for the field, but also understand and improve their experiences during their education. This paper makes a novel contribution toward that end by investigating the relationship between dialogue acts and women's remote collaborative coding experiences. The findings suggests that a woman may feel less stressed if (1) her partner is empathetic to her self-deprecating remarks, (2) she provides more suggestions than explanations, (3) her partner provides more explanations than answers, and (4) her partner does not initialize rapport-building dialogue.

These results point to many important directions for future work. Other individual characteristics, such as prior experience, could also be investigated as potential predictors in the model. Additional data may also support modeling women's experience with respect to different gender pairings to determine whether there are differences in women's experiences when they are paired with other women versus when they are paired with someone of a different gender. Future work should determine whether interventions that influence women's collaborative conversations have any bearing on their affective state during CS learning activities. Additionally, future studies should investigate how the communication medium (e.g. audio, textual, video) and collaboration interface may afford more equitable experiences. Finally, further data collection of women's remote collaborative programming during a broad range of assignments will also refine the model and our understanding of how to create more gender-equitable learning opportunities.

ACKNOWLEDGMENTS

Thanks to the members of the LearnDialogue Group for all their help. 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] Annelie Ädel. 2011. Rapport building in student group work. *Journal of Pragmatics* 43, 12 (2011), 2932–2947.

[2] Michael Baker, Tia G. B. Hansen, Richard Joiner, and David Traum. 1999. The role of grounding in collaborative learning tasks. *Collaborative Learning: Cognitive and Computational Approaches* 31 (1999), 63.

[3] Mikhail Mikhaĭlovich Bakhtin. 2010. *Speech genres and other late essays*. University of Texas Press.

[4] Andrew Begel and Nachiappan Nagappan. 2008. Pair Programming: What's in It for Me?. In *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (Kaiserslautern, Germany) *(ESEM '08)*. 120–128. https://doi.org/10.1145/1414004.1414026

[5] Harry Bunt. 2005. A Framework for Dialogue Act Specification. *Joint ISO-ACL Workshop on the Representation and Annotation of Semantic Information* (2005).

[6] Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science* 13, 2 (1989), 259–294.

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

[8] Kate Forbes-Riley and Diane J. Litman. 2005. Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*. 87–96.

[9] Joey Hadden, Laura Casado, Tyler Sonnemaker, and Taylor Borden. 2020. 21 major companies that have announced employees can work remotely long-term. *Business Insider* (2020). https://www.businessinsider.com/companies-asking-employees-to-work-from-home-due-to-coronavirus-2020

[10] Brian Hanks. 2006. Student Attitudes toward Pair Programming. In *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education* (Bologna, Italy) *(ITICSE '06)*. 113–117. https://doi.org/10.1145/1140124.1140156

[11] Douglas M. Hawkins. 2004. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences* 44, 1 (2004), 1–12. https://doi.org/10.1021/ci0342472 PMID: 14741005.

[12] Madeline Hinckle, Arif Rachmatullah, Bradford Mott, Kristy Elizabeth Boyer, James Lester, and Eric Wiebe. 2020. The Relationship of Gender, Experiential, and Psychological Factors to Achievement in Computer Science. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. 225–231.

[13] SAS Institute Inc. 2021. *JMP® Pro 15*. https://www.jmp.com/en_us/software/predictive-analytics-software.html

[14] 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.

[15] Nicole C. Krämer, Bilge Karacora, Gale Lucas, Morteza Dehghani, Gina Rüther, and Jonathan Gratch. 2016. Closing the gender gap in STEM with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers Education* 99 (2016), 1 – 13. https://doi.org/10.1016/j.compedu.2016.04.002

[16] 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)*. IEEE, 75–85.

[17] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[18] 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.

[19] Ian McChesney. 2016. Three Years of Student Pair Programming: Action Research Insights and Outcomes. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (Memphis, Tennessee, USA) *(SIGCSE '16)*. 84–89. https://doi.org/10.1145/2839509.2844565

[20] Charlie McDowell, Brian Hanks, and Linda Werner. 2003. Experimenting with Pair Programming in the Classroom. In *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE '03)*. 60–64.

[21] 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 SIGCSE Technical Symposium on Computer Science Education* (Cincinnati, Kentucky) *(SIGCSE '02)*. 38–42. https://doi.org/10.1145/563340.563353

[22] U.S. Department of Education: National Center for Education Statistics. 2019. Higher Education General Information Survey (HEGIS): Degrees in computer and information sciences conferred by postsecondary institutions, by level of degree and sex of student: 1970-71 through 2017-18. https://nces.ed.gov/programs/digest/d19/tables/dt19_325.35.asp Accessed: 01-07-2021.

[23] Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012. Rudeness and Rapport: Insults and Learning Gains in Peer Tutoring. In *Intelligent Tutoring Systems*, Stefano A. Cerri, William J. Clancey, Giorgos Papadourakis, and Kitty Panourgia (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 11–21.

[24] Vahab Pournaghshband and Paola Medel. 2020. Promoting Diversity-Inclusive Computer Science Pedagogies: A Multidimensional Perspective. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. 219–224.

[25] Sadhana Puntambekar. 2006. Analyzing collaborative interactions: divergence, shared understanding and construction of knowledge. *Computers Education* 47, 3 (2006), 332 – 351. https://doi.org/10.1016/j.compedu.2004.10.012

[26] Peter Robe, Sandeep Kaur Kuttal, Yunfeng Zhang, and Rachel Bellamy. 2020. Can Machine Learning Facilitate Remote Pair Programming? Challenges, Insights & Implications. In *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–11.

[27] Fernando J. Rodríguez, Kimberly Michelle Price, and Kristy Elizabeth Boyer. 2017. Exploring the Pair Programming Process: Characteristics of Effective Collaboration. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (Seattle, Washington, USA) *(SIGCSE '17)*. 507–512. https://doi.org/10.1145/3017680.3017748

[28] Tanmay Sinha and Justine Cassell. 2015. We Click, We Align, We Learn: Impact of Influence and Convergence Processes on Student Learning and Rapport Building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And InfLuence* (Seattle, Washington, USA) *(INTERPERSONAL '15)*. 13–20. https://doi.org/10.1145/2823513.2823516

[29] Rachel Thomas, Marianne Cooper, Gina Cardazone, Sarah Coury, Kate Urban, Ali Bohrer, Madison Long, Lareina Yee, Alexis Krivkovich, Jess Huang, Sara Prince, and Ankur Kumar. 2020. Women in the Workplace 2020. (2020). https://wiw-report.s3.amazonaws.com/Women_in_the_Workplace_2020.pdf Accessed: 01-12-2021.

[30] Linda Tickle-Degnen and Robert Rosenthal. 1990. The Nature of Rapport and Its Nonverbal Correlates. *Psychological inquiry* 1, 4 (1990), 285–293.

[31] Linda L. Werner, Brian Hanks, and Charlie McDowell. 2004. Pair-Programming Helps Female Computer Science Students. *Journal on Educational Resources in Computing* 4, 1 (2004), 4–es. https://doi.org/10.1145/1060071.1060075

[32] Laurie Williams, Robert R. Kessler, Ward Cunningham, and Ron Jeffries. 2000. Strengthening the case for pair programming. *IEEE software* 17, 4 (2000), 19–25.

[33] Ursula Wolz, Jacob Palme, Penny Anderson, Zhi Chen, James Dunne, Göran Karlsson, Atika Laribi, Sirkku Männikkö, Robert Spielvogel, and Henry Walker. 1997. Computer-Mediated Communication in Collaborative Educational Settings (Report of the ITiCSE '97 Working Group on CMC in Collaborative Educational Settings). In *The Supplemental Proceedings of the Conference on Integrating Technology into Computer Science Education: Working Group Reports and Supplemental Proceedings* (Uppsala, Sweden) *(ITiCSE-WGR '97)*. 51–69.

[34] UN Women. 2020. COVID-19 and its economic toll on women: The story behind the numbers. *UN Women* (2020). https://www.unwomen.org/en/news/stories/2020/9/feature-covid-19-economic-impacts-on-women

[35] Stelios Xinogalos, Maya Satratzemi, Alexander Chatzigeorgiou, and Despina Tsompanoudi. 2017. Student Perceptions on the Benefits and Shortcomings of Distributed Pair Programming Assignments. In *2017 IEEE Global Engineering Education Conference (EDUCON)*. 1513–1521.

[36] 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 '19)*. 1053–1059. https://doi.org/10.1145/3287324.3287380

[37] 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 (2020), Article 253.

[38] Nick Z. Zacharis. 2011. Measuring the Effects of Virtual Pair Programming in an Introductory Programming Java Course. *IEEE Transactions on Education* 54, 1 (2011), 168–170.

[39] Amy L. Zeldin and Frank Pajares. 2000. Against the Odds: Self-Efficacy Beliefs of Women in Mathematical, Scientific, and Technological Careers. *Am. Educ. Res. J.* 37, 1 (2000), 215–246. https://doi.org/10.3102/00028312037001215