No Risk, No Reward: Towards An Automated Measure of Psychological Safety from Online Communication

Sharon Ferguson
sharon.ferguson@mail.utoronto.ca
Mechanical and Industrial
Engineering, University of Toronto
Toronto, ON, Canada

Georgia Van de Zande
gdvdz@mit.edu
Mechanical Engineering,
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Alison Olechowski
olechowski@mie.utoronto.ca
Mechanical and Industrial
Engineering, University of Toronto
Toronto, ON, Canada

ABSTRACT

The data created from virtual communication platforms presents the opportunity to explore automated measures for monitoring team performance. In this work, we explore one important characteristic of successful teams — Psychological Safety — or the belief that a team is safe for interpersonal risk-taking. To move towards an automated measure of this phenomenon, we derive virtual communication characteristics and message keywords related to elements of Psychological Safety from the literature. Using a mixed methods approach, we investigate whether these characteristics are present in the Slack messages from two design teams— one high in Psychological Safety, and one low. We find that some usage characteristics, such as replies, reactions, and user mentions, might be promising metrics to indicate higher levels of Psychological Safety, while simple keyword searches may not be nuanced enough. We present the first step towards the automated detection of this important, yet complex, team characteristic.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; *Computer supported cooperative work*.

KEYWORDS

Psychological Safety, Teams, Enterprise Communication Platforms

1 INTRODUCTION

With the increasing popularity of remote and hybrid working styles, teams are communicating and collaborating via computer-mediated tools more than ever. These teams are often dispersed, which makes supervision more challenging, as the team's dynamics and predictors of success are harder for managers to monitor. One important requirement for successful teams is Psychological Safety (PS) [9, 13, 22]. The concept of PS was popularized in 1999 as a team-level construct defined as the "shared belief that a space is safe for interpersonal risk-taking" [8]. Since its inception, PS has been shown to promote the ability to learn from mistakes [8, 9], share knowledge [9, 13], and achieve better overall performance [8, 10, 22]. Despite the widespread focus on PS across numerous research fields, it is almost exclusively measured via the seven-item survey measure developed in 1999 [8, 22]. Since the survey only captures perceptions at a single point in time, team leaders looking to monitor their team's dynamics over time must redistribute the survey regularly. The survey can also fail to measure the team-level construct if each team member has different perceptions.

Thus, we argue that an automated approach to measuring this important team construct can be an early warning signal for teams and their leaders when harmful dynamics emerge and require intervention. As teams rely increasingly on computer-mediated communication, we propose that the data created on these platforms is an ideal starting point for automating the measurement of PS. We consider text-based messaging platforms, termed Enterprise Communication Platforms (ECPs) [12], as the source of data in this work, as this data reflects the team's work, language, and relationships. Predicting team performance from collaboration data has proven successful in past Human-Computer Interaction (HCI) research [5, 16, 33]; for example, Cao et al. [5]'s prediction of team viability via chat messages.

Despite the success of past work, PS is a complex team construct, with various ways in which it can be displayed. Thus, its automated identification is a challenging undertaking. In this paper, we work towards this ultimate goal by addressing the following research question: **What characteristics of ECP communication might relate to high or low levels of Psychological Safety?** In this exploratory study, we analyze the ECP (Slack) messages from two hybrid design teams: one comparatively high in PS, and one comparatively low. We quantitatively characterize their use of Slack, including patterns of replies and emoji reactions, and qualitatively analyze conversations that may signal PS.

We find that Slack use characteristics, such as the number of replies, number and type of emoji reactions, and user mentions may be a starting point to consider in the automated prediction of PS, as they differed across the two teams studied here, whereas other Slack behaviours did not. Further, we found that the counts of keywords and phrases, which represent key concepts within PS (voice, supportive, learning, and familiarity behaviours), do not

differentiate the teams in our sample. Qualitative analysis revealed that this was due to false positives and nuanced differences in the way that these terms are used by the high team compared to the low team.

We contribute the first step towards holistically understanding how PS may be detected and automatically measured on online platforms, extending past work [1] which has only used a single linguistic category as a proxy for the phenomenon. We provide an initial understanding of the communication characteristics which may differentiate comparatively high and low PS teams. We bring this exploratory study to the HCI community to spur discussion on next steps and other phenomena that could be measured using collaboration data.

2 BACKGROUND

Psychological Safety has been identified as a characteristic required for teams to innovate [9, 13, 22], learn from failure [6, 22], share knowledge [9, 10, 13, 22], and be satisfied [13]. PS is similar to trust, although it extends to describe "a team climate characterized by interpersonal trust and mutual respect in which people are comfortable being themselves" [p. 354][8]. While the construct is primarily measured via the seven-item scale proposed in Edmondson [8], scholars have divided PS into four categories to observe it: voice (or silence), supportive (or unsupportive), learning, and familiarity behaviours [23]. While PS was initially used to study successful healthcare teams [8], the concept captured the attention of managers when it was found to be a key differentiator of successful teams in Google's Project Aristotle [7].

PS has also been studied in virtual teams [15, 18, 19], and is believed to be a key component for resilient virtual teams [18]. For example, Lechner and Tobias Mortlock [19] found that virtual teams are less likely to build PS than their in-person counterparts are, so they must invest in cultivating it. Hao et al. [15] found that the level of virtuality in a team strengthened the relationship between PS, trust and knowledge-sharing. Edmondson and Bransby [9] acknowledge the importance of studying how remote work will affect team PS, and they call for methodological innovations to do so.

PS is one important characteristic of successful teams, and much past HCI work looks to predict team success overall. For example, Jung [16] predicts team performance from conflict instances, Borge et al. [2] from team interaction patterns, and Zhou et al. [34] from team feedback. Researchers have recently also turned to virtual collaboration platforms as a data source when predicting team success. Wang et al. [33] focused on how successful teams organized Slack chats. Cao et al. [5] used features of chat conversations to predict team viability, and they describe a set of their predictors as relating to PS, though they are not derived from the definition of the construct.

We explore the possibility of predicting one component of team success, PS, using virtual communication data.

3 METHODS

As this work is exploratory, we cast a wide net, collecting a broad set of virtual communication characteristics that may represent PS. We use both quantitative text analysis and qualitative thematic analysis to validate these characteristics. We detail common categories of PS from past work. As the foundational survey in [8] is the most widely accepted and validated measure of PS, we began by identifying the behavioural characteristics from these items. The survey refers to making mistakes, bringing up tough issues, taking risks, accepting others, asking for help, being supportive, and valuing skills. As this is the most common measure of PS [22], most other studies also build from this definition — such as interview and observational studies [14, 23]. Ultimately, we found that the categories within the observation scheme by O'Donovan et al. [23] act as a comprehensive summary of the key parts of the PS definition: **voice**, **supportive (or unsupportive)**, **learning**, and **familiarity behaviours**. We augmented the definition of each of these categories (defining the sub-categories ourselves) from other published measures of PS [8, 10, 14]. For example, we placed the elements of the definition of PS within each of these buckets: discussing mistakes and issues is part of **voice** and accepting, supporting and valuing others are part of **supportive**. Further, Edmondson [8] identifies how PS leads to **learning** behaviour — such as asking for feedback, or providing suggestions to improve a piece of work. Lastly, teams which are high in PS often display their comfort with each other using familiarity behaviours [23, 26], such as jokes [25]. Emojis have also been shown to only be used once teams have developed a level of familiarity [29]. These elements, and their definitions, are summarized in Table 1.

We created a list of representative keywords and phrases, that may be found in Slack communication, for each sub-category (Table 1). **Voice behaviours** represent team members feeling safe to share opinions, ideas, critiques, mistakes or misunderstandings with their team [9, 10, 22]. We identified question words ("who", "what", etc.), critique words ("incorrect", "disagree", etc.), requests for help ("help", "unsure", etc.), and mistakes ("sorry", "mistake", etc.) in this category. We consider Slack behaviours such as attachments or message edits (which could signal hesitancy to share), and the number of and time to reply (as reactions to voice behaviour are important). **Supportive behaviours** are also important reactions to **voice behaviours** [22]. They are defined by agreement and appreciation [23]. We characterized these with agreement terms ("yes", "yeah", etc.) and appreciation ("congrat.*", "wow", etc.). **Supportive behaviours** may also be correlated with replies or emoji reactions to show support for ideas that someone shared — thus we look at the number and type of emoji reactions, and the equality of contributions on Slack. We describe **unsupportive behaviours** as the opposite of **supportive behaviours**, characterized by unappreciative terms ("not needed", "stop", etc.). **Learning behaviours**, such as looking to improve or asking for feedback, also characterize teams high in PS [8]. We have defined these with suggestion terms ("improv.*", "better", etc.) and requests for input ("ideas", "thoughts", etc.). We also measure user or channel mentions, as these requests for feedback may be targeted. Lastly, teams high in PS exhibit **familiarity behaviour** [23], characterized by emojis (in text or reactions), or jokes ("hah.*", "lol.*").

We used regular expression keyword searches for these words and phrases within each team's Slack messages. We tabulate the number of messages within each team that contain these keywords.

Table 1: Categories of Psychological Safety and associated keywords and Slack behaviours. Categories from [23]. .* represents regex match for any characters.

Category	Sub-Category	Keywords and Phrases in Communication Messages	Slack Behaviours
Voice Behaviours	Mistakes	sorry, mistake, apolog.*	less time to reply, less time between messages, file share (incl. URLs), many replies, edits
	Critiques	incorrect, disagree, wrong, impossible, unlikely, "don't think"	
	Asking for Help	"don't know", unsure, help	
	Questions	who, what, where, why, how, ?	
Supportive Behaviours	Agreement	yes, yeah, ya, yea, agree.*	Many replies, many emoji reactions, equality of messages sent
	Appreciative	congrat.*, amazing, amaze, wonderful, wow, thank.*	
Unsupportive Behaviours	Unappreciative	"not needed", stop, waste	Few replies, few emoji reactions, non-equality of messages sent
Learning Behaviours	Suggestions	improv.*, better, instead, actual.*, "what if"	@mentions, @channel
	Asking for Input	feedback, share, thoughts, idea.*	
Familiarity Behaviours	Emojis	;[^/^\\^]*?:	emoji reactions
	Jokes	hah.*, aha.*, lol.*, lmao.*, jok.*	

We also qualitatively analyze the messages containing these keywords to further validate these choices and elaborate on the quantitative results. We used a qualitative thematic analysis [3] process to identify similarities and differences in how these keywords were used between the teams. The first author applied open codes to each individual message in the high team and the low team. Then, the codes were consolidated (aggregated and organized hierarchically) in each of the four categories **voice, supportive/unsupportive, learning,** and **familiarity.** Then, these codes were compared between the high team and the low team, identifying which codes were common across both teams, and which represented differences.

Further, we tabulate the counts and averages of the Slack behaviours between the two teams.

To validate these identified characteristics, we study the entire public channel (i.e., not direct messages) Slack communications from two design teams. This data contains the content of the messages as well as metadata, such as author, timestamp, attachments, replies, and emoji reactions (emoji "stickers" that can be used to react to a message). We study two of the six student teams that participated in a Mechanical Engineering capstone course at a major US Institution in 2023. The teams contained 14–16 students who collaborated with over 15 staff members to develop a functional physical product within three months. The course represents a realistic, though condensed, design process where teams manage their schedules and budgets, procure materials, and design and manufacture the product. This uniform context, where team members come from the same student population and are given the same resources, allows us to control for a number of external factors that may influence PS, such as organization culture [10], or leadership style [13, 17] so that the communication characteristics can be studied more clearly. Further, these teams are close to industrial teams as many products go on to be patented or commercialized. These teams are hybrid — they meet and work together in person but also communicate virtually regularly. Each team is designated a Slack workspace. More details about these teams can be found in [11, 12].

We surveyed these teams thrice throughout the semester to capture their perceptions of PS [8]. The surveys were administered directly on Slack. The survey response rates were moderate, anywhere from 10% to 65% (per-team, per-period). While PS is a team-level construct, in cases where the team construct fails to be measured (either from incomplete data, or misaligned perceptions) researchers measure *perceptions of Psychological Safety* instead [20, 27] − which is the average of the available PS questionnaire responses. We measure *perceptions of Psychological Safety* but refer to it as Psychological Safety (PS) in the remainder of this paper.

We chose our two teams based on consistently high or consistently low PS scores. Figure 1 shows all teams' perceptions of PS across time periods. We identified that Team 6 (in red) consistently had the lowest PS. While Team 5 had the highest average across time periods, they experienced significant variation throughout the semester. Since we were studying a semester-wide process, we chose the team with a consistently high score, or the lowest standard deviation, Team 2 (in blue). Thus, we compare Team 2 and Team 6 in this exploratory paper. Individuals provided informed consent for the surveys and to share their Slack data. Anyone who did not consent to sharing their Slack data had all of their Slack interactions removed from the dataset (three individuals from Team 6 and one from Team 2).

4 RESULTS

4.1 Quantitative Tabulations

We tabulated the Slack behaviours for the high and low PS teams, as seen in Table 2. We comment only on the characteristics which differ between teams. The low PS team sent more public channel messages than the high team. Within **voice behaviours**, the high team sent more replies, and a higher proportion of their messages had at least one reply. Within **supportive behaviour**, the high team used more emoji reactions than the low team (also a factor of **familiarity behaviour**). Slack contribution was also more equal in the high team, with a standard deviation of 3.4% of total messages, compared to 6.4% in the low team. In **learning behaviours**, the high team used @user mentions slightly more often than the low team. This
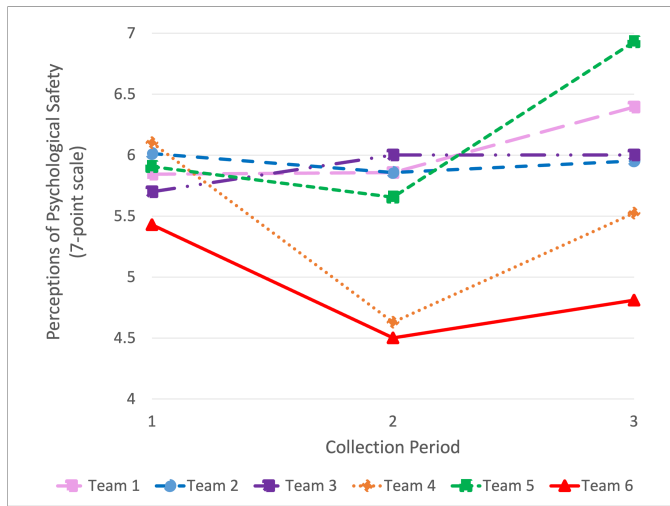
Figure 1: Average perceptions of Psychological Safety values per each team over the semester.

suggests that number of replies may be a promising avenue in signalling **voice** and **supportive behaviour**, number of emoji reactions and equality of online communication may contribute to **supportive behaviour**, and user mentions may signal teams' **learning behaviour**.

As the high PS team used more emoji reactions than the low team did, we were interested in the type of emojis that were used. Table 3 lists the ten most commonly used emoji reactions per team. The teams studied are assigned colours, and we see team members use the heart emojis in their team's colour often, perhaps to build team spirit. A custom emoji of the photo of one team member was the second most commonly used emoji reaction on the high team, perhaps representing team bonding. Five of the top ten emoji reactions were custom in the high team, while only three were custom in the low team. This could represent more **familiarity behaviour** in the high PS team, such as inside jokes.

We identified the percentage of messages within each team that fell into each sub-category from Table 1, shown in Table 4. We find that the high and low teams do not seem to differ in frequency in any of the sub-categories — all values are within a few percent of each other. Keyword searches can hide nuance in how words are used [28]; thus, we conducted a follow-up qualitative analysis to identify if there are patterns between how these terms are used across teams.

4.2 Qualitative Observations

4.2.1 Voice Behaviours. Within **voice behaviours**, we comment on mistakes, critique, asking for help, and questions.

The word "sorry" was commonly used across teams when someone was late: *"i might be about 30 min late..so sorry"*[high]. This may not be a high-stakes enough mistake to signal an interpersonal risk and correspondingly high PS. There was more variety in reasons why the individuals in the low team apologized, such as for asking for clarification, *"We talked about so many things I got confused so sorry"* [low]; or for not communicating clearly, *"sorry..I said that in*

the wrong channel oops" [low]. Both teams also discussed mistakes regarding the product or design process, which might be a better signal of PS, as they have a larger impact on the team's success. For example, *"I just came into lab and found that the soldering iron was left on...No need to claim this mistake.."*[high]; *"I'd like to apologize because I assumed our [product] was running off 120VAC."*[low]. In both these cases, a tool or product part was referenced — perhaps future iterations of this algorithm need to filter for a combination of mistake keywords with task-specific terms to improve the detection of PS. Further, a wider variety of apology types (apologies combined with communication words or questions) may signal low PS, as it could represent a team that does not feel like they can ask clarifying questions without apologizing.

We found elements of critiques of the team's process and product in Slack messages. The low team more commonly critiqued the process, *"..There is no path forward for [this product]; we don't have a sketch...have done zero testing. It will be almost impossible..."*[low]. On the other hand, product critiques were more common in the high team, *"actually [I don't know] if we have enough pins to run a 3 digit 7-bit display...?"*[high]. While it undoubtedly can also be scary to critique the team's process, this could be perceived by team members more personally, eroding PS in the team over time. Thus, critique words paired with task-focused terms, might better signal high PS.

Within the category of asking for help, we caught instances when someone said they did not know something or needed help with it. We found that in the high team, this often happened at the end of a message, after the team member had shared some research or work they had done, or an idea they had. The author would share information, and then bring up a lack of knowledge as a roadblock, where they needed their teammates' help to move forward: *"@user <link> this site talks about dipoles over ground...I don't know nearly enough about antennas to try and solve the issue theoretically"*[high]. On the other hand, individuals on the low team tended to admit they didn't know something in response to a question asked or when feedback was sought: *"I think that would be good for buoyancy but...I don't know how good it'll be at cushioning"*[low]. Thus, high PS might be signalled when the uncertainty or request follows information sharing, which may be characterized by attachments, URLs, or long descriptions. This could be the team members showing a "good faith" effort to try to solve the problem themselves first.

Questions were common among both teams, and we found no difference in how they were asked.

4.2.2 Supportive Behaviours. Under **supportive behaviours**, we discuss agreement and appreciation. Since the project requires the team to make many decisions, both teams frequently referenced agreement. We found that mentions of agreement from the high team were mostly agreeing to a specific decision or proposition: *"I agree that we should buy all 4.."*[high], and while the low team also had these instances of agreement, they commonly also questioned agreement: *"I just wanted to be clear about this..are we all agreed that the ideal user is..."*[low]. Thus, agreement combined with uncertain language might signal low PS teams where the members may want to disagree, but feel they can't.

Table 2: Count of Slack behaviours. Percentages represent the percentage of total messages sent.

Slack Behaviour	High Value	Low Value	Slack Behaviour	High Value	Low Value
Total Number of Messages	5494	7496	**Supportive Behaviour**		
Voice Behaviours			Number of Emoji Reactions	4857	2279
Number of Replies	3897	3340	% of Messages With at Least One Emoji Reaction	23%	18%
% of Messages With at Least One Reply	18.27%	12.91%	Standard Deviation of % Contribution to Total Slack Messages	0.034	0.064
Average Time to Reply	3:35:59	3:43:10	**Learning Behaviours**		
Average Time Between Messages in a Channel	18:33	17:18	Number of @Channel Mentions	41 (1%)	160 (2%)
Number of File Shares	990 (18%)	1162 (16%)	Number of @User Mentions	539 (10%)	504 (7%)
Number of Edits	243 (4%)	357 (5%)			

Table 3: Top ten most used emoji reactions for each team. † represents a custom-made emoji. Percentages are based on the total amount of emoji reactions used.

Rank	High Team	Low Team	Rank	High Team	Low Team
1	Team colour heart (1657, 34%)	Team colour heart (749, 33%)	6	Team spinner† (146, 3%)	Course emoji† (84, 4%)
2	Team member photo† (422, 9%)	Thumbs up (541, 24%)	7	Heart (143, 3%)	Laughing (83, 4%)
3	Thumbs up (356, 7%)	Heart (156, 7%)	8	Team colour pet† (138, 3%)	Eyes (58, 3%)
4	White and green check mark (206, 4%)	Cat "party"† (126, 6%)	9	Team colour fire† (126, 3%)	Exclamation mark (56, 2%)
5	Team "party"† (185, 4%)	Raised hands (105, 5%)	10	Laughing (116, 2%)	Team mascot† (33, 1%)

Table 4: Number and percentage of total messages containing keywords within each sub-category of Psychological Safety.

Concept	Sub-concept	High Team	Low Team
Voice Behaviours	Mistakes	82 (1%)	76 (1%)
	Critiques	32 (<1%)	51 (<1%)
	Asking for Help	232 (4%)	229 (3%)
	Questions	1876 (34%)	2339 (31%)
Supportive Behaviours	Agreement	241 (4%)	391 (5%)
	Appreciative	244 (4%)	244 (3%)
Unsupportive Behaviours	Unappreciative	22 (<1%)	40 (<1%)
Learning Behaviours	Suggestions	134 (2%)	197 (3%)
	Asking for Input	248 (5%)	303 (4%)
Familiarity Behaviours	Emojis	295 (5%)	317 (4%)
	Jokes	85 (1%)	101 (1%)

Both teams showed appreciation for their peers. It was more common in the high PS team to call out a specific person for a specific action, such as, *"big thanks to [names] for helping clean up our space yesterday"*[high]. On the other hand, the low team shared more general compliments, *"Slideshow looks amazing! Great job everyone!"*[low].

4.2.3 Unsupportive Behaviours. The keywords used for **unsupportive behaviours** were not effective at capturing instances of people underappreciating someone's effort. They resulted primarily in false positives such as saying someone "stopped by", discussing waste produced by the product, or stopping a function the product should do. These phrases only captured a few instances where teams were reiterating perhaps unsupportive feedback they had received from outside the team. Thus, these keywords are not effective signals without further refinement.

4.2.4 Learning Behaviours. Under **learning behaviours** we discuss instances where teams shared suggestions or asked for input

from others. In general, both teams shared suggestions about the product or the process, and we noticed no distinct patterns. There were also false positives, such as someone correcting themselves, telling a teammate to feel better, or asking about meeting logistics: *"I'm currently in [location] but can move up if that's better?"* [high]. Perhaps combining these keywords with opinion marking terms such as "I think" could help to prevent these false positives.

When teams asked each other for input, they often either requested feedback, thoughts, edits, votes, or ideas. The high team displayed more vulnerability in their asks, as they explicitly welcomed feedback or edits *"..but def[initely] feel free to make edits"*[high], whereas the low team was more likely to make general calls for thoughts or ideas: *"Any thoughts on a section name or mascot? Been saying [mascot name]...but more than happy to tweak if there's something you think would be better"*[low]. Perhaps directly asking for a teammate to critique or edit your work represents a stronger belief that the team is a safe space for risk-taking, and we should search for these vulnerability signalling terms.

4.2.5 Familiarity Behaviour. **Familiarity behaviour** is characterized by jokes and emojis. While there were many instances of jokes in both teams' communication, particularly making fun of themselves, the product, or referencing inside jokes, we found no difference in the way jokes were made between the high and low teams.

Emojis were also used *within* messages (note that these are the same set of emojis as emoji reactions, but a different use case). Teams commonly used their team colour emoji, and crying, smiling or laughing faces. Both teams used them in a self-deprecating way *"you're telling me I left RIGHT when this started :sob:"*[high], and a supportive way *"FANTASTIC job everyone with the posters and pitches!! :baby_chick::hatching_chick:"* [low]. However, emojis were

mostly used in the low team to direct their team's attention towards important information: *":alert: URGENT UPDATES...:alert:"*[low]. This practice was less common in the high team, perhaps suggesting that the low team had poor communication practices (which can both result from [22], or cause [30], low PS) and had to use other methods to capture attention.

5 DISCUSSION

In this exploratory study, we aimed to identify characteristics of virtual communication that may differentiate between higher and lower levels of PS in teams, suggesting promising directions for future investigation. Our quantitative analysis revealed that the number of replies, number and type of emoji reactions, and number of user mentions should be further investigated. We also found that simple keyword lists, alone, are not likely to be effective signals of PS. Qualitative analysis identified that this was due to nuance in how teams use these terms, such as the low team using emojis primarily to capture attention, whereas the high team used them in supportive messages.

Our results also highlight the complexity of measuring team-wide social phenomenon via text. Voice behaviour is an important sign of high PS [23]. While we can identify the presence of voice behaviour, text-based measures miss when someone chooses to remain silent, which signals lower levels of PS. Further, while all of the PS sub-categories are described in literature [8–10, 22, 23], we did not detect differences in the amount of corresponding keyword-textual evidence in the teams studied here. First, the small 1-2% differences we saw could increase if more teams were studied. Another possible interpretation is that there are different levels of PS — as suggested by its scale measure [8]— and that these categories are enacted differently at different levels of PS. This is best illustrated by our finding that the high team asks for feedback and edits while the low team asks for thoughts and ideas. Even the low team rated themselves above the midpoint on the PS scale, which may explain why we see evidence of some learning behaviours, but not all. In future work, we will define not the presence or absence of a behaviour, but the level to which it is displayed.

This exploratory study relied on only two teams (over 40 individuals and 13,000 messages). Only the messages sent by consenting team members were included in the analysis; thus, conversations reviewed manually may be missing messages. We also measure perceptions of PS based on the survey responses from a portion of the team, meaning that we may not have a complete picture of PS within the team. Despite these limitations, this exploratory work identified promising directions for the next iteration of an automated approach to detecting PS from virtual communications.

To further advance the automation of this team-wide measurement, we plan to test the communication patterns identified on a larger set of teams. We will continue iterating upon these keywords lists using the insights derived in this work (such as combining agreement terms with uncertainty measures). We will manually disentangle messages displaying PS from false positives, refine our search, and create a large, labelled dataset to validate our algorithm. We will experiment with more sophisticated detection methods, such as language models and network measures.

As we continue working towards an automated measure of PS, there are a number of validity concerns to keep in mind. While here we begin by searching for text characteristics corresponding with the foundational definition of PS, future work is needed to ensure construct validity — that the behaviours captured in the algorithm actually represent PS. To move towards this, we will broaden the text characteristics studied to look for characteristics that correlate with survey measures of PS across a larger number of teams. We will further validate this using interviews with the teams studied, verifying the intent behind, or outcome of, instances of their Slack communication. Studying a broader number of teams across contexts, and verifying our proposed measure with team members will also help to address the concern of endogeneity. We will continue to calibrate the automated measure with the well-accepted survey measure of PS [8], and triangulate our new measures with interviews or observations. Lastly, we will engage social psychologists with deep expertise in PS in comparing these different measures of PS, and validating the proposed algorithm.

As teams continue to rely on computer-mediated communication, HCI has focused on understanding and improving this experience [4, 31], and recently, using the data from these platforms to do so [5, 33]. We contribute to this effort by moving towards the automated identification of PS, a concept important for the success of virtual teams [18], yet rarely discussed in HCI (see [24] for one example). By studying characteristics related to each category in this complex phenomenon, we provide a starting point for researchers to move towards automatically measuring successful team dynamics in a number of online environments, such as open-source software teams [21] or Reddit communities [32].

The successful development of an automated measure for PS would allow system designers to build dashboards into virtual communication platforms for teams to monitor and adjust their dynamics. For example, we imagine this tool to be most useful for managers and team leaders — PS is a strong predictor of a number of team outcomes [9, 10, 13, 22]; thus, leaders having the opportunity to monitor the team's levels of PS and identify drops would allow timely interventions to get the team back on track. However, any tool which provides monitoring capabilities comes with risks. Employees may feel like they are being surveilled and like they need to filter their communication, acting in opposition to PS. Team members must be taught to treat these values as potential signals of a problem which should be investigated further, not as absolute truth, and should be careful not to make critical assumptions based on these values. To better understand how the tool will be used by teams, as well as to test the validity of the algorithm, we plan to conduct extensive user tests with industry teams.

6 CONCLUSION

Psychological Safety is a prerequisite for many positive team outcomes, yet its measurement is currently limited to perceptions measured at a single point in time. This exploratory study aims to move towards an automated measure of Psychological Safety for teams to monitor, based on data from the virtual collaboration platforms already used by most teams. Using a mixed-method study of the Slack messages from two design teams, we find that the number of replies, number and type of reactions, and the number of

user mentions are promising characteristics to further study, while simple keyword searches through messages hide many nuances. We contribute a set of communication characteristics that are a promising scope of focus for future iterations of an automated Psychological Safety measure.

REFERENCES

[1] Rohan Ahuja, Daniyal Khan, Danilo Symonette, Marie desJardins, Simon Stacey, and Don Engel. 2019. A Digital Dashboard for Supporting Online Student Teamwork. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing.* 132–136.

[2] Marcela Borge, Craig H. Ganoe, Shin-I Shih, and John M. Carroll. 2012. Patterns of team processes and breakdowns in information analysis tasks. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12).* Association for Computing Machinery, New York, NY, USA, 1105–1114. https://doi.org/10.1145/2145204.2145369

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[4] Hancheng Cao, Chia-Jung Lee, Shamsi Iqbal, Mary Czerwinski, Priscilla N Y Wong, Sean Rintel, Brent Hecht, Jaime Teevan, and Longqi Yang. 2021. Large Scale Analysis of Multitasking Behavior During Remote Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 448, 13 pages. https://doi.org/10.1145/3411764.3445243

[5] Hancheng Cao, Vivian Yang, Victor Chen, Yu Jin Lee, Lydia Stone, N'godjigui Junior Diarrassouba, Mark E Whiting, and Michael S Bernstein. 2021. My team will go on: Differentiating high and low viability teams through team interaction. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.

[6] Abraham Carmeli and Jody Hoffer Gittell. 2009. High-quality relationships, psychological safety, and learning from failures in work organizations. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 30, 6 (2009), 709–729.

[7] Charles Duhigg. 2016. What Google learned from its quest to build the perfect team. *The New York Times Magazine* 26, 2016 (2016), 2016.

[8] Amy Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative science quarterly* 44, 2 (1999), 350–383.

[9] Amy C Edmondson and Derrick P Bransby. 2023. Psychological safety comes of age: Observed themes in an established literature. *Annual Review of Organizational Psychology and Organizational Behavior* 10 (2023), 55–78.

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

[11] Sharon A Ferguson, Kathy Cheng, Lauren Adolphe, Georgia Van de Zande, David Wallace, and Alison Olechowski. 2022. Communication patterns in engineering enterprise social networks: an exploratory analysis using short text topic modelling. *Design Science* 8 (2022), e18.

[12] Sharon A Ferguson and Alison Olechowski. 2023. Are We Equal Online?: An Investigation of Gendered Language Patterns and Message Engagement on Enterprise Communication Platforms. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[13] M Lance Frazier, Stav Fainshmidt, Ryan L Klinger, Amir Pezeshkan, and Veselina Vracheva. 2017. Psychological safety: A meta-analytic review and extension. *Personnel psychology* 70, 1 (2017), 113–165.

[14] Kate Grailey, Clare Leon-Villapalos, Eleanor Murray, and Stephen Brett. 2021. Exploring the factors that promote or diminish a psychologically safe environment: a qualitative interview study with critical care staff. *BMJ open* 11, 8 (2021), e046699.

[15] Qi Hao, Bin Zhang, Yijun Shi, and Qizhong Yang. 2022. How trust in coworkers fosters knowledge sharing in virtual teams? A multilevel moderated mediation model of psychological safety, team virtuality, and self-efficacy. *Frontiers in Psychology* 13 (2022), 899142.

[16] Malte F. Jung. 2016. Coupling Interactions and Performance: Predicting Team Performance from Thin Slices of Conflict. *ACM Trans. Comput.-Hum. Interact.* 23, 3, Article 18 (jun 2016), 32 pages. https://doi.org/10.1145/2753767

[17] William A Kahn. 1990. Psychological conditions of personal engagement and disengagement at work. *Academy of management journal* 33, 4 (1990), 692–724.

[18] Bradley L Kirkman and Adam C Stoverink. 2021. Building resilient virtual teams. *Organizational Dynamics* 50, 1 (2021), 100825.

[19] Alexandra Lechner and JM Tobias Mortlock. 2022. How to create psychological safety in virtual teams. *Organizational dynamics* 51, 2 (2022).

[20] Cynthia Letting, Nicole Calpin, Nicolás F Soria Zurita, and Jessica Menold. 2023. From Obstacles to Opportunities: How Perceptions of Psychological Safety Affect Founders' Abilities to Identify Resources. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference,* Vol. 87349. American Society of Mechanical Engineers, V006T06A030.

[21] Nora McDonald and Sean Goggins. 2013. Performance and participation in open source software on GitHub. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13).* Association for Computing Machinery, New York, NY, USA, 139–144. https://doi.org/10.1145/2468356.2468382

[22] Alexander Newman, Ross Donohue, and Nathan Eva. 2017. Psychological safety: A systematic review of the literature. *Human resource management review* 27, 3 (2017), 521–535.

[23] Róisín O'Donovan, Desirée Van Dun, and Eilish McAuliffe. 2020. Measuring psychological safety in healthcare teams: developing an observational measure to complement survey methods. *BMC medical research methodology* 20 (2020), 1–17.

[24] So Yeon Park, Nicole K Santero, Blair Kaneshiro, and Jin Ha Lee. 2021. Armed in ARMY: A case study of how BTS fans successfully collaborated to# MatchAMillion for black lives matter. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–14.

[25] Wisanupong Potipiroon and Michael T Ford. 2021. Does leader humor influence employee voice? The mediating role of psychological safety and the moderating role of team humor. *Journal of Leadership & Organizational Studies* 28, 4 (2021), 415–428.

[26] Michael A Roberto. 2002. Lessons from Everest: The interaction of cognitive bias, psychological safety, and system complexity. *California Management Review* 45, 1 (2002), 136–158.

[27] Mathis Schulte, N Andrew Cohen, and Katherine J Klein. 2012. The coevolution of network ties and perceptions of team psychological safety. *Organization Science* 23, 2 (2012), 564–581.

[28] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8, 9 (2013), e73791.

[29] Esha Shandilya, Mingming Fan, and Garreth W Tigwell. 2022. "I need to be professional until my new team uses emoji, GIFs, or memes first": New Collaborators' Perspectives on Using Non-Textual Communication in Virtual Workspaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–13.

[30] Enno Siemsen, Aleda V Roth, Sridhar Balasubramanian, and Gopesh Anand. 2009. The influence of psychological safety and confidence in knowledge on employee knowledge sharing. *Manufacturing & Service Operations Management* 11, 3 (2009), 429–447.

[31] Qiao Tang, Xinmu Hu, Ziwen Zeng, and Yufeng Zhao. 2022. Co-Orb: Fostering Remote Workplace Gratitude with IoT Technology. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22).* Association for Computing Machinery, New York, NY, USA, Article 488, 6 pages. https://doi.org/10.1145/3491101.3514488

[32] Franz Waltenberger, Simon Höferlin, and Michael Froehlich. 2023. Reddit Insights: Improving Online Discussion Culture by Contextualizing User Profiles. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23).* Association for Computing Machinery, New York, NY, USA, Article 245, 6 pages. https://doi.org/10.1145/3544549.3585671

[33] Dakuo Wang, Haoyu Wang, Mo Yu, Zahra Ashktorab, and Ming Tan. 2022. Group Chat Ecology in Enterprise Instant Messaging: How Employees Collaborate Through Multi-User Chat Channels on Slack. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 94 (apr 2022), 14 pages. https://doi.org/10.1145/3512941

[34] Sharon Zhou, Melissa Valentine, and Michael S. Bernstein. 2018. In Search of the Dream Team: Temporally Constrained Multi-Armed Bandits for Identifying Effective Team Structures. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173682