Confidence Boost in Dyadic Online Teamwork: An Individual-Focused Perspective

Anonymous Author(s)

CCS CONCEPTS

• ;

ACM Reference Format:
Anonymous Author(s). . Confidence Boost in Dyadic Online Teamwork: An Individual-Focused Perspective. In *Proceedings of* . ACM, New York, NY, USA, 12 pages.

Individuals are often more confident in their solutions when working in teams than when working on their own. This confidence boost is observed even when it is not accompanied by a corresponding gain in performance, raising the question of what other factors might be responsible. We address this question by developing a large-scale experimental setting in the form of a two-player online game that allows us to track the confidence of individuals in naturally-occurring online collaborative tasks. This setting enables us to disentangle and compare the effects of different components of the collaborative process on the confidence of each team member. We show that a low-confidence individual receives a confidence boost as a direct consequence of interacting with their teammate. Furthermore, the extent of the boost depends more on the confidence, rather than on the competence, of the teammate. We operationalize these insights in a model that predicts whether a low-confidence individual will increase their confidence as a result of working with a teammate. The resulting framework can enhance our understanding of confidence boost as an often overlooked byproduct of online teamwork.

1 INTRODUCTION

One often overlooked byproduct of teamwork is the effect it has on the confidence of those involved. In particular, it was shown that teams are likely to express higher levels of confidence compared to individuals in problem-solving tasks [2, 29, 30], save for a few contrary results [34]. Notably, this boost in confidence is not necessarily justified by a corresponding improvement in performance [15, 22, 28, 32], raising the question of what might account for this phenomenon.

In this work, we take a finer-grained perspective and analyze the effect of team interaction on the confidence of *individual* team members. To this end, we develop a large-scale experimental platform that tracks individual confidence over the course of the interaction, allowing us to disentangle and compare factors mediating individual-level confidence boost.

We first establish through a randomized experiment that low-confidence individuals can get a boost in confidence *as a direct result* of discussing with their teammates. This emphasizes the importance of the nature of team interaction in shaping the confidence of the individuals, and motivates an investigation into interaction-based factors that are predictive of confidence boost.

Drawing inspiration from results reported in non-collaborative settings [24], we hypothesize that low-confidence individuals might be influenced by the confidence of the teammates with whom they interact. We provide support for this hypothesis through a regression analysis, revealing that the confidence of one's teammate is an important predictor of an individual's eventual confidence. Interestingly, we find the teammate's *competence* to be a far less predictive factor, suggesting that low-confidence individuals are not as much influenced by their teammate's knowledge as they are by the confidence their teammates exhibit.

Finally, we show that this new understanding can inform a model that aims to predict whether a low-confidence individual will experience a confidence boost by interacting with his teammate. We find that, beyond the confidence exhibited by the teammate, the language used during the team discussion serves as an additional signal. Notably, this predictive model is able to capture information beyond what can be accounted for by solution quality, contributing to our understanding of the factors mediating *unjustified* confidence boost.

In summary, in this work we:

- Design an online experimental setting that allows us to track changes in confidence exhibited by *individuals* over the course of online teamwork.
- Establish that individuals can undergo confidence boost as a direct result of interacting with their teammates.
- Show that a teammate's confidence (rather than their competence) is an important factor mediating this boost.

- Propose a model that can predict whether a low-confidence individual will experience a boost in confidence based on the interaction with their teammate.

2 BACKGROUND

Performance and confidence in teamwork

Performance of task-oriented teams. People often work in teams with the hope that by exchanging information with others, they might eventually produce higher-quality solutions. It is thus not surprising that substantial attention has been devoted to studying performance of teams.

As a result, we have an increasing understanding of mechanisms that are conducive to better-quality solutions in task-oriented teams, both offline [4, 19, 20, 36, 37] and online [10, 25, inter alia]. Some of these insights have already led to design of platforms to promote more constructive collaborations in online teams. For instance, Salehi et al. (2016) have proposed a system, Huddler, to assemble ad hoc teams while optimizing for fit between team members by harnessing the benefit of member familiarity [26].

Dyadic confidence expression. While less studied than team performance, confidence expressions in teamwork have been investigated in a number of studies in offline, face-to-face settings. The most common tasks studied include general knowledge questions [2, 3, 28], risky shifts choices [33], and perhaps more consequentially, the memory recall tasks [31, 32], which have implications for assessing validity of collaborative testimony.

It is generally observed that teams tend to express higher levels of confidence compared to individuals working on the same task. This trend holds in both experiments applying between-subject designs [2, 32] and those using within-subject designs [3, 28] (although contrary results also exist [33]).

Given that online communication offers different affordances and has characteristics distinct from offline settings, we are drawn to the question of how much this observed trend of confidence boost would generalize to an online setting, and whether we can quantify it at an individual level.

Role of individual confidence in teamwork. The role of an individual's initial confidence in teamwork has been studied in various contexts. For instance, group members' initial confidences are shown to affect the participation levels [30] and the criterion for consensus [6] in offline settings, and initial confidence is also shown to be related to an individual's influence over eventual team decisions [31].

While these studies highlight the role of an individual's *initial* confidence, in many real-life scenarios, a team collaboration does not end with the joint decision. When the actual quality of the team decision remains elusive, individuals may need to rely on their subjective perceptions of the team solution to inform their subsequent actions, making their *final* confidence in the team solution a consequential object of study. These intuitions motivate our platform design, which tracks fine-grained individual-level confidence changes over the course of team collaborations, instead of the team-level, collectively-deliberated confidence judgments that were the objects of previous study.

These considerations lead to our first research question:

RQ1: Can we observe and quantify individual-level confidence boost in online teamwork?

Confidence and related concepts

Confidence has been studied broadly and has branched out to include a number of subfields, such as *overconfidence* (which includes, notably, the finding of the Dunning-Kruger effect [18]), and *confidence realism* [1]. In this section, we review a few concepts most relevant to our study.

Confidence operationalizations. Confidence has been most commonly studied under the following three operationalizations:[1]

(1) Confidence as an estimation of *quality* of performance. Examples of this operationalization include asking students to self-rate their grades,[2] and prompting individuals to estimate task-completion times [8].
(2) Confidence as an individual's *certainty* in the answer. For instance, participants may be asked to rate their confidence on the solution with the options such as "certain", "fairly certain" and "doubtful" [32].
(3) Confidence as an estimation of one's relative *placement* within a population. With this operationalization, an individual may be asked to rank their performance relative to others [18].

In this work, we employ the first confidence operationalization for the unambiguity in its interpretation and implicit calibration between participant judgments. We further explain the rationale behind this choice in Section 3 (Operationalizations and Concepts).

Factors affecting confidence. Several studies have attempted to analyze potential sources behind people's confidence estimates. A correlational analysis suggests that confidence is dependent on the amount and the strength of supporting (but not contradicting) evidence [17]. Tversky and Kahneman propose that internal consistency is a key factor determining individual confidence judgment, which often results in unwarranted confidence from what they term "the illusion of validity" [35]. In interactive decision-making settings, experiments show that confidence boost may come from the

[1] These operationalizations naturally applies to the study of overconfidence, for which Moore and Healy (2008) has provided a detailed discussion [23].
[2] See Boud and Falchikov (1989) for a review on the operationalization on student self-assessment studies [7].

process of rationale construction [15], instead of other possible theories of "perceived information gain" or "inference certification".

In the context of team collaboration, a study by Stephenson and Wagner (1989) highlights that confidence boost might simply be an effect of the perception that the decision is joint, raising doubts on whether the interaction process in itself has additional value in boosting confidence in team members. This triggers the following question:

RQ2: Does team interaction have a causal role in boosting the confidence of the participants?

Peer influences. Perhaps most related to our work is the study by Moussaïd et al. (2013). They find that in *non-collaborative* settings, an individual may adjust their opinion and confidence in their opinion upon feedback in the form of another person's opinion and corresponding confidence, especially when the feedback is of higher confidence [24].

Viewing the dyadic interaction process as a (potential) way for both parties to either explicitly or implicitly exchange their respective opinions and their confidence in those opinions leads naturally to the following research question:

RQ3: How does the teammate's confidence and/or knowledge impact an individual's confidence?

Related concepts. One related concept that may cause some confusion is the concept of *self-efficacy*, introduced by Bandura (1982), which describes an individual's self-perception of their capacity to handle a prospective situation. This concept naturally generalizes to teams, leading to the finer-grained concepts of *team process efficacy* and *team outcome efficacy* [11]. The former captures a team's confidence in its capability to work together, while the later centers on the team's confidence in its capability to achieve team goals.

Although these concepts are related, they focus on a very different aspect of confidence compared to what we consider in this work. Specifically, while they attempt to capture an individual or a team's confidence in *prospective* events or processes, we aim to analyze confidence about decisions that an individual or team has already committed to. While efficacy-related concepts are also important, we choose to focus on the latter conceptualization as we can more objectively assess the quality of decisions that are already made, compared to assessing capabilities of individuals or teams for future events.

We address RQ1 and RQ3 via observational studies, and tackle RQ2 with a randomized experiment. We further describe the details of these studies in Section 3 (Experimental Design).

3 METHODS

Large-scale Experimental Setting

In order to observe and experimentally intervene on goal-oriented team collaborations at scale, we design a two-player[3] online game in which individuals form ad hoc teams to solve geographical puzzles. The game is a two-player version of the popular single-player game *GeoGuessr*.[4] In our game, players navigate a first-person view of some place in the world, and their task is to identify the exact location of the place shown to them with the help of their teammates. We choose this game design because it allows us to generate a virtually unlimited number of puzzles with solutions that are known to us but not to the players. The game is public online at [URL redacted for anonymity] and receives hundreds of unique weekly visitors, allowing us to collect substantial amount of data without the need for financial incentives and designing payment schemes that might otherwise affect subjective judgment [9, 16]. The experiment was approved by the Internal Review Board of the author(s)' institution(s).

For each game, two players are randomly matched into a team using a lobby system and assigned to one of the thousands of puzzles available. Each game is divided into two rounds. First, in the *solo round*, players navigate independently around a Google StreetView for a maximum of three minutes to find evidence such as street signs and vegetation that could give insight into which location they were placed in. Once the players are ready, they indicate their current (best) independent guesses of the true location by placing a marker on a world map, constituting their individual solution. Players are then prompted to indicate their confidence in their individual solutions (see Table 2 for the exact phrasing of the prompts).

Next, all team members are placed in a *team round* (Figure 1) in which they are allowed to chat with their teammates via a basic textual chat interface for up to five minutes.[5] Players may exchange the clues they found in the solo round and deliberate on a final team solution. Each individual in the team is prompted for an independent confidence rating in the team solution.

An example team chat session is shown in Table 1. It is important to note that neither the individual solutions nor the individuals' confidence in those solutions are automatically revealed to the team, making team interaction—through textual messages and movements of the joint marker on the map—the only way to communicate such information.

[3]Our platform allows up to five players to be matched in a team to facilitate future work. However, we focus on the dyadic game to control for the effect of group size in this study.

[4]https://geoguessr.com/

[5]The game concludes either when time is up or when both team members indicate their agreement with the final solution.

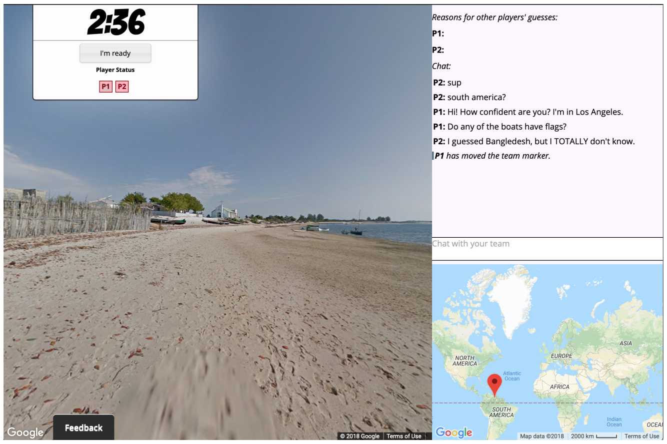

Figure 1: Game interface for the team round. During the team phase, players exchange clues they independently discovered in the solo round, after which no further navigation is possible. They communicate through a chat interface (message pane shown on the top-right corner). The full team chat for this particular game is reproduced in Table 1.

Operationalizations and Concepts

Confidence operationalization. Our setting allows a natural operationalization of an individual's confidence as their estimation of how precise their solution (guessed location) is. This operationalization has the advantage that all participants are expected to share the same understanding of what each confidence level indicates, even in the absence of explicit explanation and clarification from the experimenters, which makes it fairer to compare between participants.

We prompt each individual to self-report a confidence level on their individual solution (henceforth, $\text{conf}_{\text{start}}$, i.e., the confidence at the start of the team interaction) and on the team solution (conf_{end}, i.e., the confidence at the end of the team interaction). Table 2 lists the choices users are presented with when reporting confidence, as well as the quality levels that directly correspond to these confidence levels.[6]

(Un)justified confidence boost. To measure the degree of change in terms of an individual's confidence, we subtract an individual's initial perception from their final confidence (conf_{end} - $\text{conf}_{\text{start}}$). An individual undegoes a confidence boost if this value is strictly positive.

[6]We note that despite being ostensibly categorical, our reported confidence categories are designed to fit on a quantitative scale. In fact, the quantized confidence level from 1 to 4 tracks very strongly ($R^2 = 0.975$) with the *average* log of the distance from individuals' initial guesses to their respective correct locations (a log scale is established as being natural for these types of estimation problems [21]). Hence, the amount of improvement from level 1 to 2 is similar to the amount of improvement from level 2 to 3 and from 3 to 4, making it reasonable to compare between numerical levels and to run standard significance tests on our confidence data.

Table 1: An example team chat for the puzzle shown in Figure 1. In their solo rounds, both players had guessed the wrong continent. P1 indicated a country-level confidence ($\text{conf}_{\text{start}}$ = 3), and P2 correctly claimed that it could be anywhere ($\text{conf}_{\text{start}}$ = 1). At the end of the discussion, the players agreed on a solution indicating the Caribbean coast of Venezuela (qual_{end} = 1, as the correct solution was on a different continent, in Madagascar). While P1 maintained the same level of confidence, P2 experienced an unjustified confidence boost as P2 now estimated the team solution to be in the right region (conf_{end} = 4).

P1: sup
P1: south america?
P2: Hi! How confident are you? I'm in Los Angeles.
P2: Do any of the boats have flags?
P2: I guessed Bangledesh, but I TOTALLY don't know.
P1: no
P1: hm there is church on the beach
P2: The dude on the sand looks like not African.
He also is wearing a wedding ring.
So probably a former colony of some sort.
P2: There is trash on the beach, too.
P1: i think should be south america
P1: and maybe venesuela?
P2: Oh, and a whitish/Jewish looking guy with a baseball cap.
Definitely a tourist destination.
Probably not Bangledesh.
P2: South America sounds reasonable.
P2: Venezuela?
P1: probably caribean region
P1: yeah my guess is cenezuela
P1: venezuela
P1: i think flag is with blue color like boats :D
P2: The skiffs have large outriggers.
A place definitely historically dependent on fishing.
But that's like half the world.
P2: I'll take your guess.
P2: Venezuela for the win!
P1: ok lets try then
P1 moved the team marker [to the Caribbean coast of Venezuela]

Our discretization of confidence is designed to correspond directly to levels of solution quality. For each guess (either by an individual or by a team), we are able to obtain the ground truth location through reverse geocoding, and thus establish an unambiguous mapping between confidence levels and solution quality levels (Table 2). This correspondence allows us to objectively assess whether a confidence boost is *justified* or not by an improvement in quality.

We will refer to the quality of an individual's solo solution as $\text{qual}_{\text{start}}$, the quality of the team's solution as qual_{end}, and an individual's change in quality as the difference between

Table 2: When self-reporting confidence, players need to click on one of four buttons (first column), corresponding to four levels of confidence (second column). The average zoom level in players' map interfaces increases with higher confidence (third column), validating that the reported confidence corresponds to perceived levels of guess precision. This discretization is designed to correspond directly with levels of solution quality (last two columns).

reported confidence	confidence level	zoom level	solution quality	quality level
"Could be anywhere"	1	4.35	wrong continent	1
"I know the continent!"	2	4.55	correct continent, wrong country	2
"I know the country!!"	3	5.35	correct country, wrong region	3
"I know the region!!!"	4	7.16	correct region	4

the two values (i.e., $\text{qual}_{\text{end}} - \text{qual}_{\text{start}}$). A confidence boost is regarded as *unjustified* if we only observe an increase in confidence but not quality: $\text{conf}_{\text{end}} > \text{conf}_{\text{start}}$, yet $\text{qual}_{\text{end}} \leq \text{qual}_{\text{start}}$, as is the case in the example in Table 1.

Experimental Design

Observational Study. For RQ1, we track and compare confidence data collected from individuals before and after team collaboration to find general patterns of confidence boost. The large-scale setting further offers us the ability to analyze other observed covariates (while maintaining considerable sample size) and identify factors contributing to such changes. Specifically, we focus on analyzing the relation between an individual's eventual confidence in the team solution and the confidence and competence of their teammate (RQ3).

From November 30, 2017 to July 29, 2018, we collected a total of 1,288 two-player team games involving 2,576 solo rounds,[7] played by 1,172 unique players.

Validating the confidence operationalization. To validate that the self-reported confidence values are meaningful, we conduct two tests, one for solo round confidence and one for team round confidence. First, since players with higher confidences are expected to have more precise locations in mind as their solutions, we would expect them to zoom in more on the map to locate their solution.[8] Indeed, we find that more confident players do have higher zoom levels (Spearman's $\rho = 0.23$, average zoom levels for each confidence level is shown in Table 2, Column 3), suggesting they are making guesses that they perceive as being more precise.

Second, we compare the between-team variance[9] and within-team variance for final confidence. We expect within-team variation to be smaller than between-team variation for final confidence, since unlike between-teams individuals, within-teams individuals estimate the quality of the same guess after having discussed with each other. In contrast, we expect no difference when comparing within-team and between-team initial confidence, given that individuals make those decisions independently before interaction in both settings. Indeed, we find a significant difference for final confidence (t-test $p < 0.001$) and no significant difference for initial confidence.

Randomized Experiment. To disentangle the effects of team discussion from non-interactional effects of joint decision-making [32] (RQ2), we conduct a randomized experiment in which the treatment heavily impairs the discussion.

Under the treatment condition, teams are assigned to a "lightning" version of the game where they only have 30 seconds to discuss. For the rest of the time they experience the full functionality of the team phase, save for the discussion. While the control group experiences the combined effects of team discussion and joint decision-making, the treatment group effectively only reaps the benefits of the latter. Since teams are randomly assigned to one of the two conditions and are identical in expectation in all other aspects, the differences between these groups are thus indicative of the relative importance of the two factors.[10]

From June 8, 2018 to August 14, 2018, we randomly and independently assign teams to the "lightning" treatment before the start of the team round with 50% probability, collecting a total of 392 treatment games and 388 control games. If a dyad is selected for the lightning treatment, we inform each member immediately after completion of the solo phase via a pop-up box with the message "Lightning round! You only have 30 seconds to discuss!" After both players dismiss the box, the 30-second timer begins. Players are notified about the 30-second time limit once again in the chat box, and they

[7] Not counting games played by the authors, and only counting games for which we received complete confidence feedback and could obtain complete information about solution quality.

[8] Approximately, with zoom level 1, the map shows world-level detail; zoom level 5 corresponds continent-level detail, and one might expect to observe city-level detail at zoom level 10.

[9] We obtain between-team pairs by pairing up individuals who played the same puzzle separately in different teams, such that both teams had solutions of the same quality.

[10] We choose this operationalization over disallowing chats completely (which would completely isolate the effect of joint decision-making) in order to keep our setting natural: players may feel that eliminating team discussion removes a crucial mechanic of the game, whereas lightning rounds are a common variation of other popular games.

receive another warning when only 15 seconds remain. After the 30 seconds run out, the chat box is grayed out to indicate that chatting is disabled, and both players receive the following notification in the chat box: “The discussion period has ended. You can now adjust the map marker to finalize your guess.” The players can then use the remainder of their 5-minute team phase to finalize their guess and indicate their respective confidences in this guess.[11] Individuals assigned to the control group experience no additional notifications or changes to the game interface compared to games before this experimental period.

Manipulation check. To verify our randomized assignment, we confirm that the average solo-phase solution quality and confidence levels are not significantly different between groups. Furthermore, we did not find the response rate to be an issue, with 74% of individuals in the lightning group finishing the game and reporting confidence at the end, compared to 60% for the control group, before filtering to ensure teams have complete information ($n_1 = 776$, $n_2 = 784$).[12]

We validate that our experimental design is effective in limiting discussions by comparing the length of discussions from treatment groups to that from control groups.[13] We use the length of the discussion in terms of number of words uttered as our proxy, and find that lightning-round team discussions tend to be much shorter, at 4.3 words on average per individual versus 17.7 words on average for regular games, after filtering out games with incomplete information ($n_1 = 414$, $n_2 = 520$, t-test $p < 0.001$), confirming the treatment effect.

4 RESULTS

RQ1: Individual confidence boost

By prompting team members to reflect on their respective confidence levels before and after team interaction, we are able to track how individual confidence may change through the course of interaction. This allows us to revisit the phenomenon of confidence boost at the individual level and at a larger scale than in previous work, which has surveyed confidence of whole teams rather than individuals [28, 32]. In this section, we establish that the overall confidence boost previously observed in teams is also expressed at the individual level. Next, we zoom into different subpopulations to understand which groups of individuals are more prone to a confidence boost.

Previous work has shown that group work generally makes teams as a whole more confident when they as a group deliberate on their collective confidence [2, 28, 29]. Echoing these findings on the individual level, we find that post-interaction individual confidences are on average significantly higher than values reported prior to interaction (t-test $p < 0.01$). Furthermore, out of the individuals that do report a change in confidence (38%), a large majority (74%) report an increase.

Considering that an individual’s degree of confidence change is naturally constrained by the range of confidence values they are allowed to report,[14] it is not surprising that subgroups with different initial confidences exhibit different patterns of change (Figure 2), highlighting the need to control for one’s initial confidence level in further analysis. We focus on the subgroup of individuals who start with the lowest possible confidence level as the main subjects of our study; this group is particularly interesting as it accounts for more (53%) cases of confidence boost than all the other confidence levels combined.

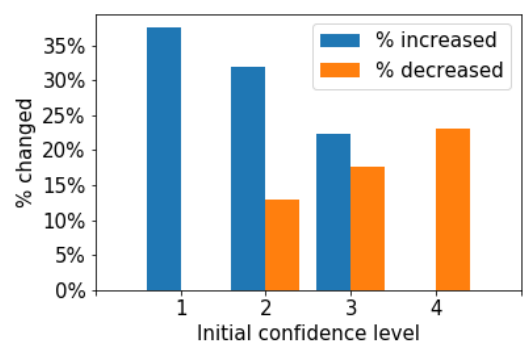

Figure 2: Individuals who start at the lowest confidence level are more likely to experience confidence boost, potentially due to larger room for improvement. We also note that the asymmetrical nature of the two types of confidence changes (compare blue bars with orange bars) further suggests that the population on average shows an increase in confidence after interaction.

It is also important to note that the confidence boost we observe may not be justified: only 30.4% of individuals who report an increase in confidence actually have team solutions

[11] We preserve the 5-minute time limit used in the regular game mode in order to avoid introducing time-stress into the players, and to make the formats of the two team rounds as similar as possible. Though the difference in time-stress between the two discussion formats may still be a factor behind our results, this is a fundamental limitation shared by previous work (e.g. [32]).

[12] It’s worth noting that the difference in response rates between the two groups is statistically significant ($p < 0.01$). However, we work around the most obvious problem posed by a difference in response rate, the fact that one member of the dyad leaving the group precludes the possibility of further discussion, by only considering dyads in which both individuals completed the game and post-game confidence survey.

[13] In subsequent analysis, we only compare games with complete information about confidence (all team members reported confidence levels) and solution quality.

[14] For instance, we would not be able to observe increase in confidence from an individual who is already at the highest confidence level. Such ceiling issues are shared by all prior operationalizations of confidence.

that are of higher quality (compared to their pre-interaction individual ones). In addition, out of the 230 individuals who end up with team solutions *worse* than their initial solutions, there are still 18.6% of them who report a higher confidence,[15] suggesting that there are quality-independent factors that contribute to this increase.

RQ2: Effect of team interaction

Noting that confidence boost happens in online teamwork even when it is unjustified by a corresponding increase in solution quality, it is natural to wonder why and when it occurs. Compared to the individual problem-solving process, collaborative decision-making bears two additional components: the process of interacting with other team members to exchange information, and the perception that one is participating in a joint decision, either of which might be responsible for the confidence boost observed. Stephenson and Wagner (1989) attempted to compare the relative effect of these two components in memory recall tasks, and concluded that the very perception of making a joint decision may create "an assumption of confidence" in the group.

However, with their experimental design, the separation between interaction and making a joint decision is not complete: the dyads are still able to talk and discuss while in the condition that tries to capture only the act of making a joint decision. Consequently, their design gives no conclusive result on the role of interaction in confidence formation. Our complete control over the online interface, on the other hand, provides us the ability to better separate the two components of collaboration, putting us in a better position to evaluate if interaction in itself may have an impact on such confidence boost.

We first note from this observational data that the *extent* of interaction is correlated with confidence boost. While our game is designed to be collaborative and encourages conversation, discussions happen at varying lengths; at the very extreme, there are teams which complete the task without exchanging any messages (henceforth *silent teams*). If confidence boost is truly independent of discussion, we should expect that the length of discussion should not have any correlation with the extent of confidence boost.

Instead, we observe significantly greater confidence boost in teams that have exchanged messages compared to silent teams, with an average increase of 0.33 versus 0.13 respectively ($n_1 = 2,148$, $n_2 = 846$, $p < 0.001$). This overall difference in confidence boost between the two groups persists regardless of whether we consider the subset of individuals in which the solution quality improves ($n_1 = 439$, $n_2 = 125$, $p = 0.01$), remains the same ($n_1 = 1,516$, $n_2 = 639$, $p < 0.001$), or decreases ($n_1 = 193$, $n_2 = 82$, $p < 0.01$), suggesting that the confidence boost afforded by discussion is not solely a consequence of an improvement in the objective quality of the produced solution.[16]

These observations can not, however, account for a self-selection bias. To exclude this alternative explanation, we conduct a randomized experiment in which we manipulate the degree of interaction (see Section 3, Experimental Design, for further details).

Experiment result. We compare 520 individuals from lightning games with 414 individuals in two-player regular games from the same time frame.

We find that 26% of individuals improve in confidence in lightning games, which is significantly lower than the 33% of individuals who improve in regular games (t-test on proportions $p < 0.05$), confirming that confidence boost comes as a direct consequence of team discussion, rather than simply being a consequence of making a joint decision.[17]

Notably, most of this difference is accounted for by the individuals who have the lowest level of confidence in their individual solutions ($\text{conf}_{\text{start}} = 1$), showing that low-confidence individuals are most susceptible to confidence boost through discussion (Figure 3).

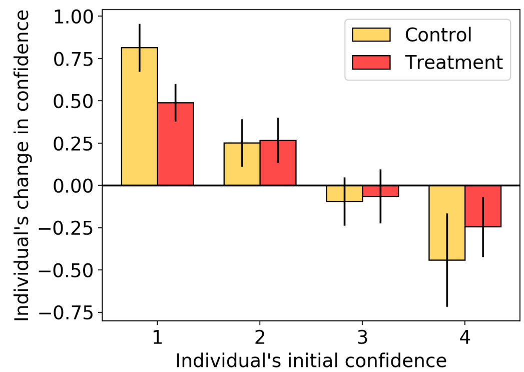

Figure 3: Comparing treatment and control games, we see that individuals with the lowest level of confidence account for most of the overall reduction in confidence boost from the lightning manipulation (t-test $p < 0.01$).

[15] 25.6% of individuals who stay at the same quality level report higher confidence, while 44.2% of individuals who improve in quality report higher confidence.

[16] Confidence boost is also most pronounced among low-confidence individuals: individuals of confidence 1 increase the most in confidence through discussion, on average (0.37 for individuals in silent teams, 0.71 otherwise, $p < 0.001$), while individuals of confidence 3 and 4 do not increase significantly on average.

[17] We find that this difference still holds even if we only consider individuals whose solution quality remains unchanged ($n_1 = 472$, $n_2 = 372$), where any increase in confidence would be unjustified.

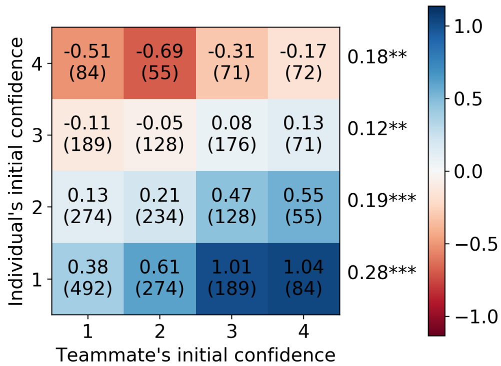

Figure 4: Heat values indicate average change in confidence for dyads with the given initial confidence composition. We observe that, for the low-confidence individuals ($\text{conf}_{\text{start}} = 1$), the more confident one's teammate is, the greater one's own confidence tends to become. Trends for high-confidence individuals are less pronounced. Correlation between one's change in confidence and the teammate's initial confidence is computed for each row and shown to the right (*p<0.05; **p<0.01; ***p<0.001).

RQ3: Effect of teammate confidence

Having seen that low-confidence individuals can undergo an (unjustified) boost in confidence as a direct consequence of interacting with their teams, we now seek to explore social factors explaining this increase. Inspired by previous work in non-collaborative settings [24], we focus on the effect of the teammate's confidence (RQ3). We hypothesize that low-confidence individuals may be influenced by, or even adopt, the confidence of their more confident teammates.

We first explore this hypothesis by comparing the confidence change of individuals who are paired with teammates of varying confidence levels. We find that the confidence boost of low-confidence individuals is more pronounced when they interact with more confident teammates (Figure 4, bottom row): the more confident one's teammate is, the larger the increase in confidence (Spearman correlation $\rho = 0.26$, $p < 0.001$). This teammate effect is less pronounced for more confident individuals (Figure 4, top three rows), suggesting that low-confidence individuals are particularly susceptible to teammate influence. Importantly, the same pattern holds (with equivalent levels of statistical significance) even if we only consider individuals with unjustified confidence boost.

To quantify this effect while accounting for other correlates, we fit a linear regression model with the individuals' final confidence as the dependent variable. For independent variables, we consider the teammate's confidence, along with the quality of the individual's solo round solution and that of their teammate's solo round solution (approximating their respective knowledge on the particular puzzle), the quality of the team's final solution (the object of the confidence estimation), and the number of words and marker movements exchanged during the discussion (approximating the level of interaction). Even after adjusting for these factors, the association between teammate's confidence and the outcome variable remains significant (Table 3, column 1).[18] The strength of this variable is on par with that of the actual quality of the team's solution, the very object of the confidence estimation.

Notably, the teammate's competence—approximated as the quality of their individual solution—appears to be a far less important factor than their confidence. This could be attributed to the difficulty of distinguishing actual knowledge from perceived or expressed competence, especially for the low-competence individuals [14, 18]. To explore this further, we compare the subgroup of low-confidence individuals that

Table 3: While one's teammate's pre-interaction confidence is predictive of one's confidence in the eventual team solution for low-confidence ($\text{conf}_{\text{start}}$ = 1) individuals (Column 1), we also note that this association may be dampened by increasing knowledge (Column 2 vs. Column 3).

	Dependent variable:		
	post-interaction confidence		
	(1) conf = 1	(2) conf = qual = 1	(3) conf = 1, qual = 3
individual solution quality	−0.05 (0.04)		
teammate confidence	0.20^{***} (0.03)	0.23^{***} (0.04)	0.15 (0.08)
teammate solution quality	0.03 (0.04)	−0.03 (0.07)	−0.02 (0.10)
team solution quality	0.20^{***} (0.05)	0.22^{**} (0.07)	0.44^{***} (0.12)
total marker moves	−0.02 (0.01)	−0.02 (0.02)	0.02 (0.05)
total words	0.004^{***} (0.001)	0.003^{**} (0.001)	0.01^{***} (0.003)
Constant	0.81^{***} (0.09)	0.80^{***} (0.11)	−0.04 (0.37)
Observations	1,039	468	142
R^2	0.14	0.16	0.22
Adjusted R^2	0.13	0.15	0.19
Residual Std. Error (degrees of freedom)	0.85 (1032)	0.77 (462)	0.96 (136)
Note:			*p<0.05; **p<0.01; ***p<0.001

[18] A linear mixed effects model with individual and puzzle as random effects produce similar results.

show little knowledge in the solo round ($\text{qual}_{\text{start}} = 1$, Table 3, column 2) with those that show more knowledge, and thus are under-confident ($\text{qual}_{\text{start}} = 3$, Table 3, column 3)[19]. We note that the latter group is indeed less susceptible to the influence of the teammate's confidence, presumably demonstrating better discernment of actual knowledge (in spite of their expressed lack of confidence). This observation opens an interesting avenue for future work on identifying under-confident individuals based on their interactional behavior.

5 PREDICTING CONFIDENCE BOOST

We operationalize our understanding of the factors responsible for confidence boost in a predictive model. This model aims to determine whether a low-confidence individual will experience a boost in confidence as a result of interacting with their teammate.

Beyond comparing the relative predictive value of the team's features, this operationalization also allows us to further explore how the nature of the discussion is tied to confidence boost. Specifically, we consider the following sets of features:

Confidence features: the confidence of the teammate prior to interaction (CONF).

Discussion features: the total length of the discussion in terms of number of words (LEN), as well as unigram counts for both the individual (BOW_{IND}) and their teammate (BOW_{TM}).

Quality features: the quality of both team members' individual solutions, as well as that of the final dyadic solution (QUAL). Note that this set of features is generally unattainable in real-life scenarios, and we consider them primarily as reference points for comparison.

We train logistic regression models on subsets of these features using 10-fold group-level cross validation, where each dyad forms a group, ensuring that individuals in the same team are either both assigned to train or both to test. Given that the task is imbalanced, we compare our models using AUC ROC and provide accuracies and minority class F1 scores as additional references. The MAJORITY baseline always predicts the "no confidence boost" label. We compare model performance via t-tests, bootstrapping performance across 10 iterations of cross-validation, for a total of 100 folds; in light of potential type I error concerns, we note quantitatively similar results with both a single iteration of cross-validation and with the 5x2cv t-test [12].

Consistent with our earlier findings, the teammate's confidence is predictive of an individual's confidence boost in this setting (Table 4, line 2). Additionally, linguistic cues of confidence boost can also be extracted directly from the text each of the teammates contributes to the discussions, significantly

Table 4: Performance of different models at predicting whether an individual will experience confidence boost (reported scores averaged over 10 cross-validation iterations). Focusing on AUC, all modes significantly outperform the majority baseline ($p < 0.001$); ††† further denotes significant improvement over CONF, while *** denotes significantly better than QUAL, both with $p < 0.001$.

	Model	AUC	Accuracy	F1
1	MAJORITY	0.50	62.4	0.00
2	CONF	0.54	62.3	0.23
3	CONF + LEN + BOW_{IND} †††	0.59	64.2	0.42
4	CONF + LEN + BOW_{TM} †††	0.57	62.8	0.39
5	QUAL	0.55	63.8	0.29
6	QUAL + CONF + LEN + BOW_{IND} ***	0.60	64.9	0.44
7	QUAL + CONF + LEN + BOW_{TM} ***	0.58	63.7	0.41

improving the performance of the classifier (lines 3-4). Thus, these linguistic signals have the potential to add additional nuances of how the confidence is expressed and interpreted by the participants. While a deep lingusitic analysis is beyond our scope, we note some cursory examples of such signals: surveying the top 30 features in terms of coefficient magnitude in our model, the use of "dude" and "nice" correspond to increase in confidence, whereas "ill" ("I'll") and "dont" ("don't") are correspond to no increase, suggesting that positively including the teammate in the discussion might be indicative of one's own confidence boost.

Notably, a model that considers only confidence and discussion features outperforms a model that has access to information about the actual quality of the solutions (Figure 4, compare line 3 with line 5). Moreover, combining all features further significantly improves upon the quality-based model ($p < 0.001$). These comparisons bring additional evidence supporting the hypothesis that the boost in confidence can be only in part attributed to the actual knowledge gain that teamwork might bring, being instead mediated by the nature of the interaction.

While useful in comparing the predictive power of different factors, the performance of our models is far too low for practical applications. Future work might consider accounting for additional factors, such as the historical malleability of individuals, and more complex linguistic and conversational features.

6 DISCUSSION

In this work, we design a large-scale online experimental platform to study how teamwork affects the confidence of individual participants. Our full control over the platform, as well as its scale, allows us to track confidence change at a finer-grained level than prior work (which mostly took a team-level perspective) and to disentangle between mediating

[19] We focus on this subgroup instead of individuals at the highest level of competence ($\text{qual}_{\text{start}} = 4$) for the larger data size.

factors. We find that team interaction directly results in a boost in confidence for low-confidence individuals, and that this boost is explainable as an adoption of the confidence, but not necessarily of actual knowledge, from the teammate with whom they interact.

While our framework and results hint at potential implications for team management, a transfer into real world application would require substantial future work. Next we highlight a few aspects of the current study that limit the interpretation of the results and demand further investigation.

Subjective experience in an online setting. In contrast with previous work which has studied the phenomenon in face-to-face collaborations, we focus on online collaboration. Due to the respective difference in affordances, one expects confidence expressions to be realized differently: for example, body-language [13] and pitch [27] can evoke confidence in person, but are missing in text-based interactions. Similarly, biases based on gender and looks of the individuals might be dampened in an anonymous setting like our own.

From this perspective, our work can be seen as confirmation that the trends observed in face-to-face studies carry over to an online collaborative setting. Our finding that people may receive a confidence boost from online teamwork provides a reassuring sign that, beyond improving team performance [25, 37], online teamwork can also have a positive effect on the subjective experience of the participants.

Affordances of the interface. In our setting, participants communicate primarily through text messages.[20] While this design choice allows us to directly analyze discussion contents automatically, it does raise the questions of how people would react and behave if the interface offers a richer set of affordances.

Future work may thus look into online collaborative platform with audio or video capabilities, either of which would reduce the degree of anonymity and enhance the social presence of participants. It would be interesting to analyze different types of signals of confidence boost (e.g. speech or visual cues), as well as the implications of better social presences on confidence expressions in teamwork.

Confidence operationalizations. We make the explicit choice of operationalizing confidence as one's perception of solution quality, since it provides automatic calibration across participants and unambiguous interpretation. This type of confidence, however, might be communicated and perceived differently than one that considers instead one's certainty in the solution.

With both operationalizations, participants may exchange and justify their respective solutions to help form the eventual team solution. However, while exchanging solutions is likely to reveal (at least in part) one's evaluation of their qualities,[21] it does not necessarily convey one's certainty in the solution. In fact, Stephenson and Wagner (1989) has noticed that dyads in their study center their discussions primarily on facts and tend to neglect to exchange confidences in their opinions [32], which might in part explain why the effect of interaction seems weaker in their reports.

Task format. Our problem-solving task follows a specific format which, while common to many online collaboration settings, is not necessarily applicable to all of them. For example, it requires participants to make one single decision. This is in contrast to some previous studies in which participants complete a *set* of questions and provide confidence ratings for each question [3, 28]. Under those settings, individual and dyadic confidences are generally computed as the averages across the entire set. Since our framework operates at an itemized level, it is not immediately transferable to aggregate confidence measures.

Design implications. Our work has a number of implications for designing interfaces to better facilitate and support online teamwork. We have demonstrated the feasibility of quantifying, tracking and even predicting confidence boost for individuals working in teams. In cases when increased confidence is desirable—certainly not always the case, since overconfidence can be problematic [23] —our findings suggest a number of intervention strategies. First, before teams are formed, pre-collaboration confidences of individuals may inform team-matching procedures that optimize for the greatest potential gain in confidence. Second, further work may develop techniques for more accurate and robust confidence detection, enabling platforms to incorporate automatic confidence tracking which relies only on user activity traces. Such platforms may signal to team managers cases in which individuals lack confidence in team decisions.

More generally, we bring attention to an aspect of online collaboration that is often overlooked. Beyond performance, collaboration platforms should consider a more diverse set of individual-level objectives. Confidence is only one example of subjective dimensions of teamwork, along with, for example, satisfaction with one's contributions or perception of fairness in the decision-making process. Even though team performance might still remain the default objective in most cases, there are scenarios in which it may not be the only, or even the most important, goal. In fact, in many real-world cases, the actual quality of the decision might remain unknown to the participants. A better understanding of more subjective aspects of teamwork is thus crucial to the design of online platforms that are flexible in accommodating different collaborative needs.

[20] Although marker movements, which are mutually visible, may also be seen as a form of communication.

[21] For instance, in our setting, "I think it is in France" indicates both an individual's guess as well as a country-level confidence.

REFERENCES

[1] Joe K Adams and Pauline Austin Adams. 1961. Realism of confidence judgments. *Psychological review* 68, 1 (1961), 33.

[2] Carl Martin Allwood and Carl-Gustav Björhag. 1990. Are two judges better than one? On the realism in confidence judgements by pairs and individuals. In *Advances in Psychology*. Vol. 68. Elsevier, 443–463.

[3] Carl Martin Allwood and Pär Anders Granhag. 1996. The effects of arguments on realism in confidence judgements. *Acta Psychologica* 91, 2 (1996), 99–119.

[4] Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. 2010. Optimally interacting minds. *Science* 329, 5995 (2010), 1081–1085.

[5] Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American psychologist* 37, 2 (1982), 122.

[6] David M Boje and J Keith Murnighan. 1982. Group confidence pressures in iterative decisions. *Management Science* 28, 10 (1982), 1187–1196.

[7] David Boud and Nancy Falchikov. 1989. Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher education* 18, 5 (1989), 529–549.

[8] Roger Buehler, Dale Griffin, and Michael Ross. 1994. Exploring the" planning fallacy": Why people underestimate their task completion times. *Journal of personality and social psychology* 67, 3 (1994), 366.

[9] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[10] Derrick Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1139–1152.

[11] Catherine G Collins and Sharon K Parker. 2010. Team capability beliefs over time: Distinguishing between team potency, team outcome efficacy, and team process efficacy. *Journal of Occupational and Organizational Psychology* 83, 4 (2010), 1003–1023.

[12] Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10, 7 (1998), 1895–1923.

[13] Katrien Fransen, S Alexander Haslam, Niklas K Steffens, Norbert Vanbeselaere, Bert De Cuyper, and Filip Boen. 2015. Believing in âĂIJusâĂİ: Exploring leadersâĂŹ capacity to enhance team confidence and performance by building a sense of shared social identity. *Journal of experimental psychology: applied* 21, 1 (2015), 89.

[14] Liye Fu, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2017. When Confidence and Competence Collide: Effects on Online Decision-Making. In *Proceedings of WWW*.

[15] Chip Heath and Rich Gonzalez. 1995. Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes* 61, 3 (1995), 305–326.

[16] P Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk Gabriele Paolacci Advanced School of Economics, CaFoscari University of Venice Jesse Chandler Woodrow Wilson School of Public and International Affairs, Princeton University. *Judgment and Decision Making* 5, 5 (2010).

[17] Asher Koriat, Sarah Lichtenstein, and Baruch Fischhoff. 1980. Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory* 6, 2 (1980), 107.

[18] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77, 6 (1999), 1121.

[19] Patrick R Laughlin and John Adamopoulos. 1980. Social combination processes and individual learning for six-person cooperative groups on an intellective task. *Journal of Personality and Social Psychology* 38, 6 (1980), 941.

[20] Patrick R Laughlin, Erin C Hatch, Jonathan S Silver, and Lee Boh. 2006. Groups perform better than the best individuals on letters-to-numbers problems: effects of group size. *Journal of Personality and social Psychology* 90, 4 (2006), 644.

[21] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108, 22 (2011), 9020–9025.

[22] Peter R Meudell, Graham J Hitch, and Petty Kirby. 1992. Are two heads better than one? Experimental investigations of the social facilitation of memory. *Applied Cognitive Psychology* 6, 6 (1992), 525–543.

[23] Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review* 115, 2 (2008), 502.

[24] Mehdi Moussaïd, Juliane E Kämmer, Pantelis P. Analytis, and Hansjörg Neth. 2013. Social Influence and the Collective Dynamics of Opinion Formation. *PLOS One* 8, 11 (2013), e78433. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078433

[25] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. (2016), 568–578.

[26] Niloufar Salehi, Andrew McCabe, Melissa Valentine, and Michael Bernstein. 2016. Huddler: Convening stable and familiar crowd teams despite unpredictable availability. *arXiv preprint arXiv:1610.08216* (2016).

[27] Klaus R Scherer, Harvey London, and Jared J Wolf. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality* 7, 1 (1973), 31–44.

[28] Jonathon P Schuldt, Christopher F Chabris, Anita Williams Woolley, and J Richard Hackman. 2017. Confidence in dyadic decision making: The role of individual differences. *Journal of Behavioral Decision Making* 30, 2 (2017), 168–180.

[29] Janet A Sniezek. 1992. Groups under uncertainty: An examination of confidence in group decision making. *Organizational Behavior and Human Decision Processes* 52, 1 (1992), 124–155.

[30] Janet A Sniezek and Rebecca A Henry. 1989. Accuracy and confidence in group judgment. *Organizational behavior and human decision processes* 43, 1 (1989), 1–28.

[31] Geoffrey M Stephenson, Dominic Abrams, Wolfgang Wagner, and Gillian Wade. 1986. Partners in recall: Collaborative order in the recall of a police interrogation. *British Journal of Social Psychology* 25, 4 (1986), 341–343.

[32] Geoffrey M Stephenson and Wolfgang Wagner. 1989. Origins of the misplaced confidence effect in collaborative recall. *Applied Cognitive Psychology* 3, 3 (1989), 227–236.

[33] James Arthur Finch Stoner. 1961. *A comparison of individual and group decisions involving risk*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[34] R Scott Tindale. 1989. Group vs individual information processing: The effects of outcome feedback on decision making. *Organizational Behavior and Human Decision Processes* 44, 3 (1989), 454–473.

[35] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[36] Wendy M Williams and Robert J Sternberg. 1988. Group intelligence: why some groups are better than others. *Intelligence* 12, 4 (1988), 351–377.

[37] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330,

6004 (2010), 686–688.