Coupling Interactions and Performance: Predicting Team Performance from Thin Slices of Conflict

MALTE F. JUNG, Cornell University

Do teams show stable conflict interaction patterns that predict their performance hours, weeks, or even months in advance? Two studies demonstrate that two of the same patterns of emotional interaction dynamics that distinguish functional from dysfunctional marriages also distinguish high from low-performance design teams in the field, up to 6 months in advance, with up to 91% accuracy, and based on just 15minutes of interaction data: Group Affective Balance, the balance of positive to negative affect during an interaction, and Hostile Affect, the expression of a set of specific negative behaviors were both found as predictors of team performance. The research also contributes a novel method to obtain a representative sample of a team's conflict interaction. Implications for our understanding of design work in teams and for the design of groupware and feedback intervention systems are discussed.

Categories and Subject Descriptors: H.1.2: User/Machine Systems, H.4 Information Systems Applications, H.5.3 Group and Organization Interfaces, J.4 Social and Behavioral Sciences, K.4.3 Organizational Impacts

General Terms: Measurement, Performance, Theory, Design

Additional Key Words and Phrases: Teamwork, intra-group conflict, emotions, team dynamics, team performance, design teams

ACM Reference Format:
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 (May 2016), 32 pages.
DOI: http://dx.doi.org/10.1145/2753767

1. INTRODUCTION

Predicting team performance from team dynamics is difficult. Despite a long history of studies into the relationship between team dynamics and performance, there remains inconsistent evidence about what aspects of a team's dynamics predict its performance [Tausczik and Pennebaker 2013]. While some approaches, such as Linguistic Style Matching [Gonzales et al. 2009], have shown promising predictive power in the laboratory, they fail when tested in the field, e.g., Munson et al. [2014]. With few exceptions, e.g., Curhan and Pentland [2007] and Jung et al. [2012], especially measures that are predictive by just assessing the initial stages of a team's life are lacking. Most studies that examine team dynamics and performance have used data covering the entire lifetime of a project (for example, Fussell et al. [1998], Munson et al. [2014], and Tripathi and Burleson [2012]) rather than using just a small sample of data that is collected before work on a project concludes.

The ability to predict team performance is not only of interest for our understanding of teamwork in general but also for many aspects in the area of human–computer

This work was funded in part by generous support from the Kempe Foundation of Sweden.
Author's addresses: M. F. Jung, Department of Information Science, Cornell University, 206 Gates Hall, Ithaca, NY 14851; email: mfj28@cornell.edu.

DOI: http://dx.doi.org/10.1145/2753767

interaction (HCI). First, teamwork is central to HCI design practice in education and industry. Many current HCI courses involve project work in teams and almost all HCI work in industry involves teamwork to some degree. Developing predictive measures of team performance and especially design team performance informs our understanding of teamwork in HCI, helps us develop better ways to help students learn effective teamwork practices, and informs industry practitioners when managing teams engaged in HCI-related work.

Second, understanding how team dynamics relate to performance is not only important for our conceptual understanding of teamwork and especially of teamwork that involves user-centered design practices but also for designing systems that support teamwork, for example, by giving feedback or intervening into team dynamics. While several studies have demonstrated that team dynamics can be influenced through feedback, none of them were able to show any effect on team performance [DiMicco et al. 2004, 2007; Kim et al. 2008; Leshed et al. 2009]. As Tausczik and Pennebaker [2013] remark "Shaping group dynamics relies on understanding the basic question: Why do some groups of people work well together while others do not?" The lack of robust predictors of team performance and theories about the relationship between team dynamics and performance are a crucial limiting factor in developing such feedback systems to improve performance. Rather than just asking "what *can* we measure," it is important to ask, "what *should* we measure" when designing new feedback systems. In order to build systems that improve team dynamics and performance, measures are needed that not only accurately reflect the quality of a team's interactions but also predict a team's future performance.

Finally, building understanding about the role of team dynamics in shaping team performance informs how we design groupware. Identifying the processes that are influential in shaping team performance is important when rethinking how we structure and design systems that support work in groups. In designing these systems, designers draw from an implicit understanding of what is important for the performance of work in groups. Extending and challenging current assumptions about what is important for teams to perform well is, therefore, crucial when building software that is truly supportive of teamwork.

1.1. Conflict and Performance in Teams

A process that has received considerable attention, as a determinant of team performance, is conflict within groups, typically referred to as intragroup conflict [Jehn 1995, 1997; Amason 1996]. Intragroup conflict is a promising process to focus on, because it permeates all teamwork, and how groups engage in conflict is thought to be an important determinant of performance [De Dreu and Weingart 2003; De Wit et al. 2012].

It can be considered a robust finding that conflict impairs team performance once it becomes "personal" and laden with hostility [Jehn 1995, 1997]. However, despite a few studies, which examined how conflict unfolds at a moment to moment level, e.g., Paletz et al. [2011] and Paletz et al. [2013], not much is known about the role of emotional expressions during conflict at that level of analysis [Weingart et al. 2015], and whether teams show emotional conflict interaction patterns at the onset of a project that are characteristic of later conflicts and predictive of performance. The lack of insight into how conflict and especially emotions play out on a moment-to-moment basis, can be partially attributed to the fact that almost all conflict studies rely on retrospective assessments of conflict [De Dreu and Weingart 2003; De Wit et al. 2012] and past research has not attempted to predict team performance based on sampling interaction dynamics during actual episodes of conflict or by assessing conflict days, weeks, or even months in advance. Additionally, past research on conflict has focused predominantly

on the topic of conflict and specifically whether it is about relationship-related issues or task-related issues [De Dreu and Weingart 2003], thereby largely ignoring a specific focus on affect during conflict despite ample research showing the central role of affect for the performance of teams [Barsade and Gibson 2007].

Efforts to find reliable predictors of team performance based on conflict dynamics have been murky at best. Two highly influential studies by Jehn [1995, 1997] shaped the currently dominant theory that the type of conflict a team engages in predicts its performance (positive for task-focused conflict, and negative for relationship-focused conflict). However, a metaanalysis of a large number of intragroup conflict studies found that the topic of conflict does not distinguish low from high-performing teams [De Dreu and Weingart 2003]. On the other hand, researchers of married couples have been far more successful at predicting outcomes based on interaction dynamics during conflict, and particularly, by focusing on affect. Especially illustrative is a study in which Gottman and Levenson [2000] demonstrated that the fate of a marriage can be predicted years in advance from the emotional interaction dynamics occurring during a couple's conflict interaction—with 93% accuracy. Particularly impressive in this study was that only a 15-minue thin slice of a couple's conflict interaction was needed to make these highly accurate predictions. In another study, divorce could be predicted over a 6-year period based on the emotional interaction dynamics measured during the initial 3-minute slice of a conflict episode with 80% accuracy [Carrere and Gottman 1999]. Using the same thin-slicing approach of measuring emotional dynamics during conflict, it was possible to predict marital outcomes such as satisfaction and divorce across a wide range of studies [Gottman 1994; Levenson et al. 1994; Levenson and Gottman 1983; Levenson and Gottman 1985]. Despite the impressive predictive power of thin slices of conflict in predicting marital outcomes, and despite the importance of conflict for performance in teams, thin slices of conflict have not yet been explored as a predictor of team performance.

1.2. Thin Slicing Conflict to Predict Performance

Thin slicing, the process of making accurate classifications based on small samples, or "thin slices" of expressive behaviors [Ambady and Rosenthal 1993], has not only proven effective in predicting outcomes of marriages but also of doctor–patient interactions, family interactions, interviews, or work-related interactions [Ambady and Rosenthal 1992]. Thin-slicing approaches have also been used to study how people form impressions of online profile information [Stecher and Counts 2008], or to improve debugging and program understanding [Sridharan et al. 2007]. Thin slices used in the studies of behavior typically ranged between 30 seconds and 5minutes. The thin-slicing research showed powerfully that certain behavioral characteristics are stable over time and that only a small-interaction sample is necessary to make meaningful judgments about behavior occurring over longer durations such as hours, or even months. Finding ways to predict team performance from thin slices is particularly relevant for the design of feedback systems because an ability to predictive performance from just a small sample of data would allow the development of diagnostic tools and feedback systems that do not require the continuous monitoring of teams and rather focus on short interaction episodes that can be instrumented more easily.

The current research, therefore, extends thin-slicing work by Jung et al. [2012] as well as Curhan and Pentland [2007] by examining to what extent thin slices of conflict interactions predict team performance in the field. Curhan and Pentland demonstrated that subjective and objective negotiation outcomes can be predicted from just a 5-minute sample of interaction dynamics at the onset of an employment negotiation. Jung et al. [2012] demonstrated that subjective and objective software engineering team performance can be predicted from just a 5-minute interaction sample of a

two-person team at the onset of a programming task. Both of these studies are laboratory studies with dyads rather than larger teams and predictions were made over just a few hours. It is still an open question whether accurate predictions can be made over a timeframe of months for project teams in the field. The studies presented here demonstrate that subjective and objective team performance can be predicted months in advance from the balance of positive to negative affect and a specific set of hostile behaviors occurring during just a 15-minute thin slice of a team's conflict interaction.

The studies presented here are the first to apply theory about conflict in marital interactions to further our understanding of conflict and performance in design teamwork. The article makes three specific contributions. First, this research contributes to our understanding of teamwork in design and intragroup conflict more broadly. The studies found that two critical interaction pattern that distinguish between functional and dysfunctional marriages, also distinguish between high- and low-performance design teams and introduces them as key predictors of team performance: Group Affective Balance (GAB), the balance of positive to negative affect during an interaction, and Hostile Affect, the expression of a set of specific negative behaviors. These findings have direct implications for how we teach designing in teams, as well as how we design groupware and feedback intervention systems. Second, while past research on predicting performance from thin slices of interaction data focused on dyads, working on a laboratory task over a few hours, this research extends that work [Curhan and Pentland 2007; Jung et al. 2012] to larger teams, in the field, and over a timeframe of months. Third, this research introduces a conflict elicitation protocol that generates an intragroup conflict interaction in the lab that is easily instrumentable and highly diagnostic of a team's interaction dynamics.

2. GAB AND PERFORMANCE

Central to the studies on predicting marital outcomes from emotional dynamics during conflict is a balance theory of marriage [Gottman and Levenson 1992]. The theory posits that a couple's ability to regulate the affective balance of positive to negative affect during conflict, such that a surplus of positive affect is maintained, is critical for the quality and long-term outcomes of the relationship. This ability to regulate affect is especially crucial to repair the impact of hostile expressions and to prevent negative affect from escalating further. In other words, couples that are able to consistently produce more positive than negative affect especially during their conflict interactions are more likely to have satisfying relationships and are more likely to stay married [Gottman and Levenson 1992; Gottman 1994]. Implicit in the idea to focus on the positive and negative emotions in relation to each other is also the finding that it is not the presence of negative expressed affect during conflict that is detrimental to a relationship but the absence of any positive affect: "stability in marriage is likely based in the ability to produce a fairly high balance of positive to negative behaviors and not in the exclusion of all negative behaviors" [Gottman and Levenson 1992, p. 232].

Based on the idea that the way in which a couple balances positive and negative affect during conflict tells us a lot about a couple's fate, the assessment of relative amounts of positive and negative affect has been highly informative across many studies [Gottman and Levenson 1992; Gottman 1994; Levenson et al. 1994; Levenson and Gottman 1983; Levenson and Gottman 1985]. A key predictor of marital satisfaction and divorce was a couple's affective balance measured as the relative amounts of positive to negative expressed emotions occurring during a thin slice of a couple's conflict interaction. While the construct of affective balance describes behavioral dynamics of conflict, research has shown that besides an observational operationalization it can also be reliably operationalized by assessing affective experience and physiology during conflict [Levenson and Gottman 1985].

There is evidence that the idea of relative amounts of positive and negative affect as predictors of performance generalizes beyond marital interactions to teamwork. Using a thin-slicing approach with professional programmers working in pairs on a day-long programming task, Jung et al. [2012] showed that, analogous to couples, programming teams could be categorized as either regulated or nonregulated based on their affective balance assessed during the first 5 minutes of their interaction. This categorization predicted not only the satisfaction of the programmers with the overall programming experience but also the objective quality of the code they had developed. While this study focused on dyadic teamwork interactions, the idea that abilities to manage conflict and to regulate emotions are crucial for long-term outcomes extends beyond dyads to groups and teams with more than two members. If negative and especially hostile emotions are not regulated, they are likely to initiate a spiraling increase of interpersonal negativity [Andersson and Pearson 1999]. In line with this, Barsade [2002] demonstrated that it only takes one negative team member to affect an entire team and impair conflict processes and performance. Focusing specifically on intragroup conflict, Curseu et al. [2012] found that groups skilled at emotion regulation are more likely to prevent disagreements about tasks from developing into damaging forms of interpersonal conflict. Together these studies support the idea that affective balance during conflict is not only a predictor of outcomes in couples but also in groups.

H1: A team's GAB assessed from a 15-minute thin slice of the team's conflict interaction will be predictive of team performance.

3. STUDY 1: PERFORMANCE PREDICTION BASED ON THE EXPERIENCE OF AFFECTIVE BALANCE

Participants engaged in an 8-month team-based product development project. A 15-minute thin slice of a conflict interaction was video recorded for each team during the project and emotional interaction dynamics were assessed through self-report. Team performance assessed through self-report was used as the dependent variable.

3.1. Participants

Thirty engineering design teams with overall 100 students were recruited over the course of three years from a three quarter long master's level capstone course in user-centered mechatronics design at a large North American university. Most participants had several years of prior industry experience. Team sizes ranged from 2 to 4 students (M: 3.37). 19 teams were mixed in gender, 11 were all male and there was no all female team. Teams were self-formed. In some cases, students knew each other before taking the course. The teams were entirely self-managed and formally leader less. Teams did not designate an explicit leader or project manager and an emergent leadership style was used throughout the class. Overall the teams mimicked the type and structure of small startup teams or self-managed research and product development teams in industry.

The class provided a good context to study teamwork as several aspects of the particular class setting match those of other team projects in field settings other than a classroom: First, with 9 months, the class project was long enough for group dynamics to develop. The class is three quarters or approximately 9-months long of which about 8 months are used for one project. Each team "owned" a space in a large, open space (see Figure 1) for the entire duration of the three quarters. Team members spent a majority of the project time working together in their assigned space.

Second, teams worked on open-ended assignments for which no clearly specified success criteria existed in advance (see Table I for sample project descriptions). Third, the class required a time commitment comparable to that of a project in industry.

Fig. 1. Class-room space that housed the teams for the duration of the three quarters. Each team "owns" a workspace with a table. The classroom also offers meeting spaces, and a small "shop" area with tools and a work table.

Table I. Problem Descriptions From One Cohort of the Product Development Class

Industry	Problem Description
Automotive	Design and build a Human Input Device (HID) that will accommodate driver and passenger needs in the year 2020.
Software	Design a website that introduces the various alternative fuel technologies to the consumer.
Mechatronics	Design and build a camera-projector prototype sensor system usable in mobile robotics.
Government	Develop a solution that protects a person falling down by preventing him/herself from getting hurt.
Software	Design and develop a solution that enables a transfer from mechanical design/manufacturing techniques to the design and construction of buildings.
Consumer Products	Design a new consumer electronics oral care solution that motivates its user to maintain regular oral hygiene and that provides feedback of how effective hygiene has been.
Consumer Products	Develop a platform for blending traditional, physical symbols ("atoms") with what are now relatively separate social practices in the digitally connected world ("bits") as an exciting domain for future wearable, network-capable consumer products.
Software	Design and build a system that allows home office work.
Telecommunication	The development of new services and products in the health area using mobile telephony as a platform of communications to promote the use and development of 3G technologies.

All projects focused on user-centered design and either involved mechatronics or HCI aspects.

Students usually spend 20 to 40 hours per week on this course and up to 50 or more before major deadlines. Fourth, teams had to accommodate real world constraints as the projects were industry sponsored, and teams were responsible to a company liaison, and responsible for a budget of approximately US $15,000. At the end of the class, all project outcomes were presented in front of an academic and industry audience during a large project fair. (The class has been described in detail by Carleton and Leifer [2009].)

3.2. Procedure

Data about GAB and team performance were collected in two steps. First, about one academic quarter, two and a half months before the class ended, I obtained a representative sample of each team's conflict interaction style through an interaction session

Fig. 2. Laboratory setup for the interaction session. A low table was chosen to allow for a more intimate interaction. Four cameras are placed around the table to capture each person's behavior individually.

and I measured each team's affective balance through a recall session. Second, once the class was over, I measured team performance by administering the team diagnostic survey [Wageman et al. 2005].

3.2.1. Group Interaction Task to Obtain an Interaction Sample. The goal of the group interaction task was to elicit a representative sample of a group's conflict interaction style. I chose a conflict scenario because conflict interactions have been found particularly diagnostic of group and marital performance alike. Participating in a discussion of a conflicting topic creates a lot of engagement and the goal of the interaction session was to generate an interaction sample that would be as diagnostic as possible about the affective interaction style of a group. The group interaction session was timed such that it occurred in the week leading up to a major class-deadline in order to sample each team at a time of heightened engagement.

I developed the group interaction session by transferring the dyadic interaction task [Roberts et al. 2007] from a dyadic couples interaction task into a task for small workgroups. The original dyadic interaction task was a core component of the studies investigating the ability to predict a couples future marital satisfaction and likelihood of divorce [Gottman 1994; Levenson et al. 1994; Levenson and Gottman 1983; Levenson and Gottman 1985]. The task elicits emotionally charged conflict episodes that are highly comparable to interactions outside the laboratory, and therefore, has the advantage of a high ecological validity [Roberts et al. 2007]. The main component of the task, the conflict discussion, is composed of three phases: (1) Conflict topic inventory, (2) conflict facilitation, and (3) conflict interaction. The aim of the first phase is to identify potential conflict topics. Both spouses fill out the problem inventory and rate how much they disagree with their partner about a specified set of problem topics. The aim of the second phase is to facilitate a conflict interaction. This is usually done by a facilitator who identifies a conflict topic that partners disagree about using the survey ratings. In a discussion with the couple, the facilitator, then, highlights differences in opinion, and draws out emotions related to the conflict topic thus priming the couple for conflict. Once a suitable topic is identified, the facilitator leaves the room and the third phase begins, which has the aim for the couple to discuss potential solutions for the identified topic. This last phase typically lasts for 15minutes. My aim was to create a similar task (The Group Interaction Task) that would allow me to elicit those behaviors

Fig. 3. Snapshot of a team as recorded during the interaction session. Four cameras captured each person's face and upper body.

and interaction styles that would be most diagnostic of a group's conflict engagement style.

Before the main interaction task and directly upon arrival at the laboratory, groups were greeted by an experimenter, and asked to sit around a circular table in the center of a small room. After giving informed consent each group was asked to discuss their projects' most important requirements for 15 minutes. The purpose of this task was to get participants familiarized with the room and to situate the discussion in the context of their group projects. The requirements task allowed the groups to quickly get into a meaningful discussion about their project, to surface different values and disagreements and to directly contribute to the progress of each group's specific project.

The second task was a group problem-solving task to elicit the desired group conflict interaction sample. This part of the group interaction task mimics the structure of the dyadic interaction task in that it also consists of three phases: Conflict topic inventory, conflict facilitation, and conflict interaction. In the first phase, each group member was given 5 minutes to develop a problem inventory by individually listing issues of disagreement within the team, and then, to order the issues in terms of importance. Participants were encouraged to list issues about the task, the process of how tasks were completed within the group, and interpersonal relationship-oriented issues. This free-form inventory was chosen because topics that aroused conflict interactions were typically highly specific to each team and no standardized conflict topic inventory existed. Once the inventory was completed the second phase started and the experimenter entered the room and asked participants to share an issue they deemed important. The experimenter then facilitated a discussion that allowed the group to converge on one issue that constituted a major area of disagreement within the team and that elicited a comparable level of engagement by all team members. As the third phase, the issue that emerged from this discussion was then given to the group with the task to discuss it for 15minutes toward a possible solution.

Audio and video were recorded during both the requirements discussion and the problem discussion. Four video cameras were installed in the room such that they clearly captured each person's face and upper body (see Figure 3 for an example). The cameras could be adjusted remotely to adapt to subjects changing positions. A microphone was placed in the center of the table to capture speech at high quality.

Fig. 4. Setup for the recall session. Participants simultaneously watched a video record of the problem discussion while continuously rating how they felt using a rating dial placed in front of them. Headphones and blinds were used to minimize participants influencing each others' ratings.

Fig. 5. Rating dial setup used during the recall session. Participants were asked to indicate how they felt during the interaction by turning the arrow toward a position that best described how they felt. The dial is based on a video controller but the spring forcing the dial back to the neutral position was removed as not to bias the ratings toward a neutral rating.

3.2.2. Measuring GAB through a Recall Session. Immediately after the problem discussion, each group completed a recall session (Figure 4). The task for the recall session was modeled after the recall task used in studies on couples [Ruef and Levenson 2007]. The goal of the task is to obtain a continuous self-report measure of each group member's emotional experience as it emerged throughout the conflict discussion by letting participants reexperience the group interaction.

Participants were placed at a table in front of a rating dial (Figure 5) and were asked to watch a recording of the group's 15-minute conflict discussion. Headphones and a visual barrier were used to maintain privacy of each participant's ratings and minimize participants' influence on each other's ratings through reactions (e.g., giggles) to the videos. Participants received the following instructions:

> As you watch the video, please indicate how you perceived the interaction using the rating dial. Try to put yourself back in the situation of the discussion and adjust

the rating dial so that it indicates how you perceived each situation. The dial can be adjusted from "very negative" to "neutral" and "very positive" and to any state in between. Please adjust the dial as often as needed, so that it always indicates how you felt. Hold the dial in the selected position as long as the emotional state endured. For example if you felt slightly positive during a long time then keep the dial slightly inclined to the right for the entire time you had that slightly positive feeling.

As participants were watching the video recording, the rating dial allowed participants to continually indicate how they felt at each moment in time by adjusting the dial on a 14-point scale from "very negative" to "very positive. Participants provided on average 119.1 affect ratings (SD = 94.4, Min = 6, Max = 633) over the 15-minute recall session. A key advantage of the self-report-based assessment of affect is the highly reduced effort in comparison to systematic observation and previous research has validated the effectiveness of this procedure in recalling and assessing emotional dynamics of actual interaction sessions [Gottman and Levenson 1985].

The final GAB measure was operationalized by counting the number of instances the rating dial was turned to one of the upper five positive points (on the 14 point scale) and subtracting the number of instances the rating dial was turned onto one of the lower five negative points during the 15-minute recall session. Then, the mean of these difference scores was calculated for each group.

The justification for using the group-level mean of difference scores lie in the construct properties of GAB. GAB can be best described as what Kozlowski and Klein [2000] refer to a configural group construct. Configural group constructs are seen as distinct from shared and global group constructs and "capture the array, pattern, or variability of individual characteristics within a team" [Klein and Kozlowski 2000; p. 215]. For configural team constructs, there is no assumption that individual characteristics of interest are held in common by the members of a team. Since only "shared team properties require the demonstration of within-group consensus or consistency" (ibid, p.18), the validity of the measure does not rest on the consistency of behavior patterns across individuals. Additionally, GAB does not intend to describe a latent psychological property of a group such as group cohesion, or group mood. It rather describes the behavior patterns occurring within a group and especially the balance in the occurrence of positive and negative affect patterns and the experience thereof.

The decision to operationalize affective balance by subtracting positive from negative affect instead of using, for example, a ratio-based approach was made because subtraction-based operationalization of affective balance has been used in several studies of marital interactions that this research is building upon. For example, Carrere and Gottman [1999] used a subtraction-based balance measure to predict whether newly married couples would divorce within 6 years following the measurement. Later Gottman and Levenson et al. [2000] used a subtraction-based balance measure to predict divorce in couples from 15 minutes of video over 14 years. Finally, Gottman and Levenson [1992] compared subtraction-based balance measures with ratio-based balance measures and found that predictive power was comparable between measures. A subtraction-based approach for assessing affective balance is also inherent in the construction of emotional dynamics graphs (called point graphs), which were a central instrument in describing emotional dynamics of couples [Gottman 1994; Gottman and Levenson 2000], and which were also used for the second study described in this article.

3.2.3. Performance Measurement. Team performance was assessed through self-report and operationalized with the Team Diagnostic Survey [Wageman et al. 2005] at about 8 months into the project and after all class deliverables had been completed. The Team Diagnostic Survey is a widely used 26-item self-report measure that assesses

Table II. Descriptive Statistics and Pearson Correlations Among Independent Variables

Variable		1	2	3
(1) GAB	Pearson's r	—	0.153	−0.335
	p-value	—	0.437	0.082
(2) Team Size	Pearson's r		—	−0.200
	p-value		—	0.307
(3) % Female	Pearson's r			—
	p-value			—
M		0.310	3.357	0.241
SD		0.244	0.622	0.195
Minimum		−0.428	2.000	0.000
Maximum		0.646	4.000	0.666

Note: $N = 28$. $^{*}p < 0.05$, $^{**}p < .01$, $^{***}p < 0.001$ (All two-tailed Pearson correlation tests).

Table III. Results of Hierarchical Multiple Regression Analysis with Subjective Team Performance as Dependent Variable

	R^2	ΔR^2	B	$SE\ B$	β
Step 1	0.35***				
Constant			3.31	0.15	
GAB[1]			1.43	0.39	0.59***
Step 2	0.41**	.06			
Constant			2.99	0.57	
GAB[1]			1.65	0.41	0.68***
Team Size			0.02	0.15	0.02
% Females			0.82	0.51	0.12

Note: $^{\text{n.s.}}p$ = not significant; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

team processes that are crucial for team effectiveness and that has been validated with a large number of teams in many kinds of organizations (ibid). The instrument assesses team performance by measuring how a team manages its effort, how a team draws on member talent, and whether a team develops a performance strategy that is well suited to the task and situation. While the instrument does not explicitly ask team members about the perceived performance of the team, previous work has shown that team performance is directly correlated with the successful management of these three processes [Hackman 2002], and therefore, other researchers have used it as a measure of team performance as well (e.g., Hackman and O'Connor [2004]). The global score of the Team Diagnostic Survey, the average of the scores on all three dimensions, was taken as an overall team performance measure (Scale reliability $\alpha = 0.90$). Each item is measured along a five-point scale ranging from "very inaccurate" to "very accurate."

3.2.4. Controls. Team size and percentage of female team members were included as controls. Team size was chosen based on the idea that the more people work on a problem, the better the outcome. Percentage of female team members was chosen as previous research showed that the percentage of females in the team correlates highly with overall group performance [Woolley et al. 2010]. Table II presents an overview of descriptive statistics and correlations for each independent variable.

3.3. Results of Study 1

28 of the 30 groups were included in the analyses. For two groups, no rating dial data were available from which to calculate a GAB score. Hierarchical multiple regression

[1]An alternative, ratio-based, GAB measure (as constructed in Gottman and Levenson, 1992) revealed the same significant correlations, albeit with slightly lower R^2 values.

was performed to investigate the ability of GAB to predict team performance while controlling for team size and percentage of females in each team. Analyses showed that assumptions of normality, multicollinearity, linearity, and homoscedasticity were not violated.

In the first step of the hierarchical multiple regression, only GAB, the main predictor of interest, was entered. This model was statistically significant $F(1, 26) = 13.79$; $p < 0.001$ and explained 35% of the variance in team performance. In the second step, team size and the percentage of female team members were entered in the model. The total variance explained by model 2 was 41% $F(3, 24) = 5.57$; $p < 0.01$. While adding team size and percent females explained an additional 6% of variance in team performance, this addition of explained variance over model 1 was not significant ($\Delta R2 = 0.06$; $F(2, 24) = 1.30$; $p = 0.29$). Additionally, neither team size ($\beta = 0.02$, $p = 0.91$) nor percent females ($\beta = 0.27$, $p = 0.12$) were statistically significant as predictors.

3.4. Discussion of Study 1

The findings provide support that affective balance is not only a strong predictor of marital outcomes but also of team performance. These findings were not only significant but also constituted a surprisingly large effect. The single factor of GAB measured over just 15 minutes explained 35% of the variance in team performance 2.5 months in advance. The 35% explained variance is particularly high given that the model was theory driven rather than fitted post hoc. Interesting is also the finding that neither team size nor the composition of the team (amount of females) mattered as predictors of team performance.

Three limitations should be noted. First, team performance was assessed through self-report. While the team diagnostic survey is a widely accepted instrument and has been shown to correlate with team performance [Hackman and O'Connor 2004], it remains an open question if objective team outcomes can be predicted from emotional interaction dynamics during conflict. Second, with only two and a half months left before project completion, the thin slice of the teams' interaction dynamics was taken relatively late during the project life. This leaves open the possibility that the affective balance measurements assessed groups' implicit perceptions of their own performance. In other words, some groups might have been aware that late in the class that their performance was not on par with their peers, and therefore, exhibited and experienced more conflict and frustration. To control for this potential "self-diagnosis" effect, future studies should collect data much earlier, at a point when it would be more unlikely for a team to already have developed a possibly accurate intuition about their own team performance and when there is more time to intervene. Third, the study used an experiential measure of a team's GAB. It is an open question to what degree the experience of affective balance is reflected in expressed behavior during conflict interactions. Identifying a behavioral measure that could be assessed with unobtrusive automated techniques of a team's emotional interaction dynamics will also be critical for incorporating GAB in feedback systems for teams or automated approaches to assess team dynamics.

4. STUDY 2: PERFORMANCE PREDICTION BASED ON BEHAVIORAL INTERACTION DYNAMICS

The goal of the second study was to extend the first study to gain deeper insights into the interaction dynamics during conflict and to address the limitations listed in the previous paragraph. Several extensions were made: First, an objective performance measure was used in addition to the self-report measure. Second, emotional interaction dynamics were assessed much earlier during the project lifetime at a point when the

teams were less likely to form an accurate intuition about their performance. Third, a behavioral measure of GAB was used that allowed a far more granular lens on the ensuing interaction dynamics during conflict. Consequently, I extended the first hypothesis as follows:

H2: A team's GAB assessed from a thin slice of the team's conflict interaction will be predictive of subjective as well as objective team performance.

In order to develop system interventions that can improve team performance it is important to identify those behaviors that are particularly harmful for a team's performance. Of the behaviors monitored in couples interactions four behaviors were found to be particularly corrosive for marriages: contempt, criticism, stonewalling, and defensiveness. Contempt, as defined in the studies of marital interactions [Coan and Gottman 2007], refers to behaviors that belittle, hurt, or humiliate another party. This includes mockery, direct insults, sarcasm, but also subtle entirely nonverbal behaviors such as eye-rolls or dimplers [Ekman and Friesen 1982]. Contempt might be the most damaging behavior of the four. Criticism refers to statements that highlight another party's personality as inherently defective. Studies of intragroup conflict have highlighted these negative judgments of personality as a typical characteristic of relationship conflict. For example, Jehn [1997, p. 542] cites this statement as characteristic of relationship conflict: "Her attitude just stinks. It's a personality conflict in the first place. . . ." Stonewalling refers to behaviors that "communicate an unwillingness to listen or respond [Coan and Gottman 2007, p. 279]." A typical form of stonewalling is for one person to seemingly evaluate their fingernails while being talked to. Finally, defensiveness categorizes behaviors that reflect an intention to deflect responsibility or blame. Defensiveness can even take the form of counterattacks. Due to their corrosiveness, these four behaviors have been called "horsemen of the apocalypse" in studies of marital interactions [Gottman 1994]. Hostile behaviors also have been theorized as one of the major reasons why relationship conflict is harmful for team performance [Weingart et al. 2015] as they are highly likely to escalate into an increasing spiral of negativity [Andersson and Pearson 1999]. I therefore hypothesize that

H3: The number of hostile affect expressions (horsemen) made during a conflict sample predicts subjective and objective team performance.

4.1. Participants

Student teams were recruited from the same class as in study 1. Nine teams, with overall 36 students participated in the study. Each team had four members. Seven teams were mixed in gender, two had only male team members.

4.2. Procedures

Procedures were identical with those in the first study except the interaction session was conducted at 2 months into the project and 6 months before the conclusion of the class. The interaction session took place in the week leading up to the first major deadline in the second quarter of the course. The percentage of female team members was included as control.

4.2.1. Measuring GAB. GAB was measured experientially as in study 1. In addition, systematic observation of behavior of the problem discussion session was used to construct a behavioral GAB measure [Bakeman and Gottman 1997; Weingart et al. 2004]. To code the videos, I used a version of the Specific Affect Coding System (SPAFF) that was adapted for breast cancer support group interaction [Giese-Davis et al. 2005]. SPAFF is a mutually exclusive and exhaustive coding system that generates a

Fig. 6. Screenshot of the VCode video coding environment for a three-person team observed during a pilot-study. The timeline shows a 14-second segment of SPAFF codes for the person on the upper left. Each line represents a different emotional expression.

continuous stream of behavioral data for each coded participant [Coan and Gottman 2007]. This particular version of SPAFF distinguishes between 23 categories of emotion expression and is more sensitive in capturing subtle changes in affect. SPAFF was chosen for two specific qualities that distinguish it from other behaviorally based categorization systems for emotions. First, SPAFF [Coan and Gottman 2007] captures four quadrants of behavior: Facial muscle movement, speech prosody or tone of voice, verbal content, and body posture and movement. This distinguishes it from other coding systems for emotions such as the commonly used Facial Action Coding System (FACS) [Ekman and Friesen 1978]. Second, in comparison to other behavioral emotion coding systems that distinguish affective behaviors on the level of movement, SPAFF distinguishes between affective behaviors on the level of the emotional meaning a particular behavior has in a specific context. This second characteristic makes SPAFF extremely powerful but also difficult to apply.

Videos were analyzed by 12 coders using the VCode software (Figure 6) [Hagedorn et al. 2008]. Coders were kept blind to the hypotheses and to the performance of the teams. Each coder went through a 5-week-long extensive training that followed the three steps outlined in Coan and Gottman [2007]: First, coders were sensitized to SPAFF constructs and indicators and trained in basic people watching skills. Second, coders were trained in a subset of FACS, which concluded with a short FACS exam. Third, coders learned the set of SPAFF codes through reading materials, video examples, self-enactments, and coding exercises.

Training was partially conducted by an SPAFF trainer who was experienced in training SPAFF coders for more than 15 years. To facilitate the training and the coding, a 30 page code manual was created that described each code in detail with examples and behavioral indicators. The curriculum included basic training in FACS to sensitize coders to subtle changes in facial muscle movement as well. Coding was done for each person shown in the video separately. Coding took about 6–8 hours per person (∼20–30 hours per team).

To allow for reliable coding, three steps were taken. First, a baseline set of coded videos was created with the help of a professional SPAFF coder. The baseline videos were recorded from a previous class cohort and were not part of the study dataset.

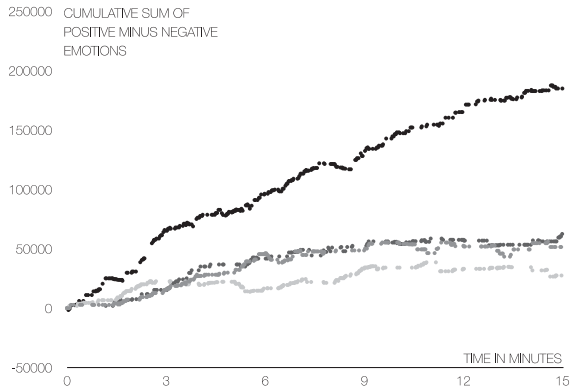

Fig. 7. Point graph for one of the nine teams. A point graph plots the cumulative sum of positive minus negative emotions over time. Each line represents one of the four members of the team. An upward slope, for example, indicates that a member of the team was able to consistently produce more positive than negative expressions of affect. The average of all four point graph slopes was taken as a measure of GAB.

Each coder's performance was then tested against that baseline and coders were only allowed to code videos from the actual study dataset once they showed an interrater reliability score of at least (Kappa = 0.6) with the baseline dataset. Second, weekly meetings were held during which two coded files were compared code by code in order to clarify confusions and to prevent reliability decay. Third, two independent coders double coded 9 of the 35 videos (25%) to allow interrater agreement testing. From the videos that were coded twice, one was selected at random to be included in the analysis.

The behavior-based GAB measure was operationalized in analogy to the operationalization used in marital studies that is based on the construction of point graphs [Gottman and Levenson 1992; Gottman and Levenson 2000]. Point graphs plot the cumulative sum of positive minus negative affect over time and have been a key instrument in visualizing and analyzing not only the emotional interaction dynamics of couples [Gottman and Levenson, 1992] but also of programming teams [Jung et al. 2012]. To build point graphs, first, the millisecond-duration of each SPAFF code was multiplied with "+1" for positive codes, "–1" for negative codes, and "0" for the neutral and the tension codes and the resulting values were plotted cumulatively as a point graph over time (see Figure 7 for an example of a point graph). Then, as a second step, linear regression analyses were performed on all point graphs to determine their slope. Third, the average slope of the point graphs for each team was taken as each team's GAB measure.

4.2.2. Measuring Hostile Affect. The team-based hostile affect measure was operationalized by taking the average count of occurrences of the emotions contempt and defensiveness during the conflict discussion. "Criticism" the third of the behaviors called "the four horsemen of the apocalypse" was captured by the contempt code in the version of SPAFF used and no occurrences of "stonewalling" were coded during the interactions.

4.2.3. Controls. As in study 1, percentage of female team members as included as control. Team size was constant at four members per team and was, therefore, not included as a control variable. Table IV presents an overview of descriptive statistics and correlations for each independent variable. The strong positive and significant correlation between the behavioral and self-report-based GAB measure confirms the alignment of experiential and behavioral measures of interaction dynamics in the marital studies

Table IV. Descriptive Statistics and Pearson Correlations among Independent Variables

Variable		1	2	3
(1) GAB (behavioral)	Pearson's r	—	0.758*	0.291
	p-value	—	0.018	0.448
(2) GAB (self-report)	Pearson's r		—	−0.185
	p-value		—	0.634
(3) % Female	Pearson's r			—
	p-value			—
M		0.038	0.325	0.250
SD		0.033	0.265	0.177
Minimum		−0.017	−0.084	0.000
Maximum		0.090	0.693	0.500

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (All two-tailed Pearson correlation tests).

[Gottman and Levenson 1985; Levenson and Gottman 1885], and further validates the use of a self-report-based GAB score as a proxy for a behavior-based measure.

4.2.4. Performance Measurement. Team performance was measured subjectively as in study 1 with the Team Diagnostic Survey [Wageman et al. 2005] at about 8 months into the project and after all class deliverables had been completed. Team performance was also assessed objectively through the final grade each team received for their main prototype deliverable, which typically consists of a functional technical system. The grade is given by the teaching team, a group of professors, teaching assistants, and industry consultants.

4.3. Results

All nine teams were included in the analysis. One team had only three members present during the group interaction session. Coder agreement was assessed using Cohen's Kappa [Cohen 1960] and ranged between $\kappa = 0.51$ and $\kappa = 0.67$ (M = 0.59, SD = 0.07). According to Landis and Koch [1977], this is a moderate to substantial level of agreement. Between 132 and 484 emotional expressions were coded for each participant (M = 291.89, SD = 85.13). Table V shows how often each expression occurred on average per participant.

The objective performance measure, $D(30) = 1.66$, $p < 0.05$, was significantly non-normal, and therefore, Kendall's tau correlation coefficients were used to analyze correlations with this measure. Subjective and objective performance measures were not correlated ($\tau = -0.02$, $p = 0.87$, two-tailed), which is surprising but emphasizes the importance of an objective performance measure.

4.3.1. GAB and Performance. All nine groups were included in the analyses. Hierarchical multiple regression was performed to investigate the ability of GAB to predict subjective and objective team performance while controlling for percentage of females in each team.

In the first step of the hierarchical multiple regression, only the behavior-based GAB, the main predictor of interest, was entered. This model was marginally statistically significant $F(1, 7) = 4.14$; $p = 0.08$ and explained 37% of the variance in subjective team performance. In the second step, the percentages of female team members were entered in the model. The total variance of subjective performance explained by model 2 was 38% $F(2, 6) = 1.84$; $p = 0.24$. Adding percent females explained only an additional 1% of variance in team performance, which did not constitute a significant improvement ($\Delta R2 = 0.01$; $F(1, 6) = 0.09$; $p = 0.78$). Additionally, percent females ($\beta = 0.1$, $p = 0.78$) was not a significant predictor. Therefore, model 1 was chosen as the final model, but

Table V. Descriptive Statistics Showing the Number of Times Each Emotion was Coded Per Person. Most Frequent Expression was Neutral, Followed by Tension and Validation. Hostile Behaviors (Horsemen) are Underlined

SPAFF Codes	*M*	*SD*	min	max
Low affection	0.6	1.1	0	5
High affection	0.1	0.5	0	3
Validation	47.5	21.7	15	86
Interest	5.5	5.5	0	23
Excitement	0.8	2.1	0	10
Humor	2.6	3.4	0	11
Neutral	89.5	37.0	39	208
Tense humor	5.3	4.4	0	21
Tension	71.4	23.1	27	123
Low fear	0.1	0.2	0	1
High fear	N/A	N/A	N/A	N/A
Low sadness	0.6	1.6	0	9
High sadness	N/A	N/A	N/A	N/A
Direct anger	N/A	N/A	N/A	N/A
Constrained anger	12.9	20.0	0	113
Micro-moment contempt	8.3	10.4	0	54
Contempt	2.1	2.8	0	11
Domineering	3.5	4.6	0	18
Belligerence	0.8	2.6	0	15
Defensiveness	3.3	3.4	0	14
Whining	N/A	N/A	N/A	N/A
Disgust	0.1	0.3	0	1
Stonewalling	N/A	N/A	N/A	N/A

Table VI. Hierarchical Multiple Regression with the Behavior-Based GAB Score as Independent and Subjective and Objective Team Performance as Dependent Variables

	R^2	ΔR^2	*B*	*SEB*	β
Dependent Variable: Subjective Team Performance:					
Step 1	$0.37^{\dagger}$				
Constant			3.23	0.27	
GAB (behavior)			11.30	5.56	$0.61^{\dagger}$
Step 2	0.38	0.01			
Constant			3.17	0.36	
GAB (behavior)			10.77	6.23	0.58
% Females			0.34	1.15	0.12
Dependent Variable: Objective Team Performance:					
Step 1	0.01				
Constant			4.07	0.17	
GAB (behavior)			−0.64	3.46	−0.07
Step 2	0.40	$0.39^{\dagger}$			
Constant			4.28	0.18	
GAB (behavior)			1.11	3.04	.12
% Females			−1.11	0.56	$-0.66^{\dagger}$

Note: $^{\text{n.s.}}p$ = not significant; $^{\dagger}p < 0.01$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

due to the lack of significance H2 received no support. The self-report-based GAB score was neither able to predict subjective ($F(1, 7) = 2.69$; $p = 0.15$) nor objective ($F(1, 7) = 1.19$; $p = 0.31$) team performance.

4.3.2. Hostile Affect and Performance. First, hierarchical multiple regression was performed to investigate the ability of the hostile affect score to predict subjective team performance while controlling for percentage of females in each team. As Table III

Table VII. Hierarchical Multiple Regression with Hostile Affect as Main Independent and Subjective and Objective Team Performance as Dependent Variables

	R^2	ΔR^2	B	$SE\ B$	β
Dependent Variable: Subjective Team Performance					
Step 1	0.00				
Constant			3.73	0.44	
Hostile Affect			−0.04	0.27	−0.06·
Step 2	0.10	0.10			
Constant			3.56	0.50	
Hostile Affect			−0.12	0.29	−0.17
% Females			1.11	1.41	0.32
Dependent Variable: Objective Team Performance					
Step 1	0.80***				
Constant			4.50	0.10	
Hostile Affect			−0.31	0.06	−0.89***
Step 2	0.91***	0.12*			
Constant			4.59	0.08	
Hostile Affect			−0.27	0.05	−0.77***
% Females			−0.61	0.22	−0.36*

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

shows, a model predicting subjective team performance based on hostile affect was not significant ($F(1, 7) = 0.03$; $p = 0.88$) even when controlling for the amount of female members in the teams ($F(2, 6) = 0.32$; $p = 0.74$).

Second, hierarchical multiple regression was performed to investigate the ability of the hostile affect score to predict objective team performance while controlling for percentage of females in each team. In the first step of the hierarchical multiple regression, only hostile affect, the main predictor of interest, was entered. This model was highly statistically significant $F(1, 7) = 25.59$; $p < 0.001$ and explained 80% of the variance in objective team performance. In the second step, the percentages of female team members were entered in the model. The total variance of subjective performance explained by model 2 was 91%, $F(2, 6) = 31.59$; $p < 0.001$. Adding percent females to the model explained an additional 12% of variance in team performance, which did constitute a significant improvement ($\Delta R2 = 0.12$; $F(1, 6) = 8.00$; $p < 0.05$). Additionally, percent females ($\beta = -0.36$, $p = 0.05$) was significant as predictor. Therefore model 2 was chosen as the final model that provided strong support for H3. Figure 8 shows a plot of the relationship between hostile affect and objective team performance.

4.4. Discussion of Study 2

The second study extended the first study by providing more insight into the role of specific emotional expressions in predicting team performance. Additionally the second study addressed methodological limitations of the first study. Interaction samples were collected almost 6 months before the project deadline instead of 2.5 months ahead. The early assessment time gives more opportunities for successful interventions. Performance was assessed objectively through the prototype grade. A behavioral measure of GAB was used and a measure of hostile affect was introduced.

Most importantly, hypothesis 3, which states that those behaviors that have been found particularly corrosive for marital interactions (and therefore been referred to as the four horsemen of the apocalypse) are also predictive of performance, was partially supported. Hostile behaviors predicted 91% of objective team performance but not subjective team performance. Even without controlling for female group members, hostile affect explained a surprisingly high 80% variance in objective team performance.

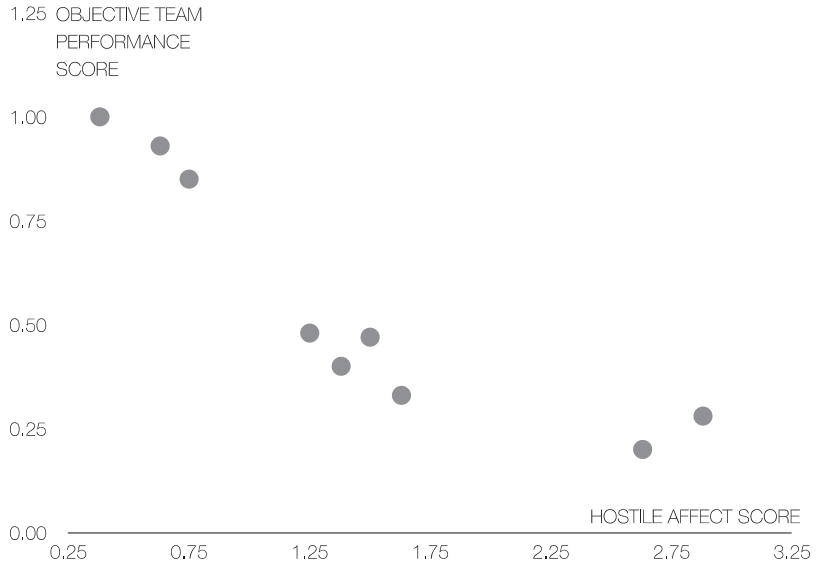

Fig. 8. Plot visualizing the correlation between the amount of expressions of hostile affect (during 15min) and objective team performance (final prototype grade 6 months later).

Figure 8 depicts a clear trend that with increasing expressions of hostility the objective team performance (as measured per final prototype grade) declines.

Worth noting is also the finding that the percentage of female team members was correlated negatively with objective team performance. There was a positive correlation with subjective team performance, but this was not significant. Considering that the first study found no correlation between female team membership and performance, these results are different from what other recent studies found in which a correlation between percentage of female team membership and team performance was demonstrated (e.g., Woolley et al. [2010] and Apesteguia et al. [2012]). These differences in findings speak to the complexity of gender dynamics and more broadly diversity within a team. Both of the studies cited stressed the mere composition of a group as an important factor for performance and the findings presented here raise the question whether, over the long term, the impact of a group's composition might be outweighed by a group's interaction dynamics that develop over time. In other words, the relationship between composition structure and development of helpful or harmful interaction patterns is unclear. Since the studies cited above relied on shorter timeframes, future studies could shed light how team composition and especially gender dynamics influence the development of interaction patterns over time.

The second study could not replicate the findings about the predictive power of GAB on performance. While the effect size for GAB in predicting subjective team performance was similar to the first study, the effect turned out not to be statistically significant. The lack of statistical significance at the 5% level might be due to the relatively small sample size of the second study.

An important limitation of the second study is its small sample size of nine teams. Assuming an effect size of 0.39 as it has been found in Ambady and Rosenthal's [1992] metaanalysis of thin-slicing studies, to achieve 80% power the recommended sample size for a study is 23 for one predictor and 28 for two predictors according to G*Power [Faul et al. 2009]. Based on this effect size estimate, the power of the statistical test is assumed to be 0.37 for one predictor, and 0.24 for two predictors. The low power of

the test has two implications for the interpretation of the findings. First, in general terms, this means that the likelihood of detecting an effect when there is in fact one is only 37% and 24%, respectively. Given the low power, a nonsignificant test does not mean that there could not be an effect. This might also explain why the second study could not replicate the findings about GAB and performance of the first study. The second implication of the low power is a high likelihood of inflated effect sizes. Therefore, the 91% and 80% effect sizes might not adequately reflect the true effect size in the population and the actual effect size for the relationship between hostile affect and performance might be closer to what Ambady and Rosenthal found in their metaanalysis.

5. OVERALL DISCUSSION

The findings across both studies provide support for the idea that the same emotional interaction patterns that distinguish early on between functional and dysfunctional marriages also distinguish between high- and low-performing teams: Affective balance and hostile affect. The first study examined only one measure but with its larger sample size provides support that the findings may generalize to a broader population (statistical prerequisites for generalizability were met). The second study adds an additional predictive pattern: Hostile Affect.

The studies demonstrated that team performance can be predicted from emotional interaction patterns occurring during just 15 minutes of a team's conflict conversation up to 6 months before a project concludes. The same behaviors that earned the title "four horsemen of the apocalypse" due to their corrosiveness for marriages seem to have similarly corrosive effects on team outcomes. The findings were not only significant but also constituted large effects rarely found in this kind of research. GAB explained up to 35% of the variance in team performance and a model containing hostile affect explained 91% of variance in team performance. The findings also have to be interpreted in the light that the models were not fitted post hoc to the data but rather constructed based on prior work and theory.

Way conflict is assessed in this article is different from previous intragroup conflict studies (for a review, see De Dreu and Weingart [2003] and De Wit et al. [2012]) in two important ways: first, almost all previous studies of intragroup conflict assessed conflict through self-report measures. Most studies employed an intragroup conflict scale developed by Jehn [1994] that distinguished task and relationship conflict as key predictors of performance. The assessment of intragroup conflict in this study is different in that it assesses conflict as it actually occurs through a moment-to-moment analysis of emotional behavior. Second, the distinction of task and relationship conflict in the literature is based on the topic of conflict, or what conflict is about (i.e. relationship-related issues such as clashes of values, or task-related issues, such as disagreements about project ideas). The assessment of intragroup in this study distinguishes conflict patterns based on their emotional dynamics, rather than based on topic. Overall these two key differences distinguish the conflict assessment from previous conflict assessment approaches and they allow a new perspective on conflict and how it evolves on a moment-to-moment basis. Future studies should address how the conflict dynamics described here relate to the established dimensions of task and relationship conflict.

An important limitation of this research is that due to the nature of the studies, no causal claims can be made about the relationship between emotional interaction dynamics and performance. However, the strong parallels to the findings on marital interactions invite the speculation that it is the ability to regulate affect that determines not only marital satisfaction and divorce but also team performance. There are several studies that support this idea. Studying specifically conflict in groups, Curseu

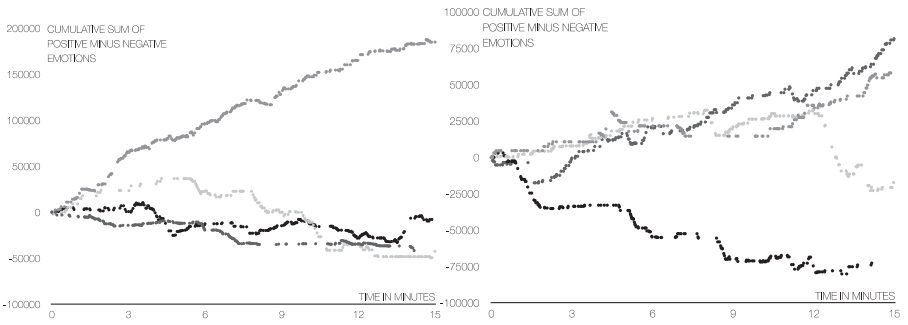

Fig. 9. Left: Point graph for a team with one positive "outlier" and three downwards-oriented point graphs. Each line represents one member's emotional expressions over time. Right: Point graph for a team with one negative "outlier" and three upwards-oriented point graphs.

et al. [2012] argued that groups with a high ability in regulating emotions are more likely to maintain productive forms of conflict and avoid harmful forms of conflict. In related work on negotiations, Carnevale and Isen [1986] found that inducing positive affect improves negotiation outcomes. While not focusing specifically on conflict, Barsade [2002] demonstrated in a study in which a confederate worked on a problem-solving task with other team members, that the emotional interaction dynamics and performance of teams can be influenced through the emotional behavior of one team member alone. More broadly, emotion and emotion regulation have been argued to be crucial processes in determining team performance [Barsade and Gibson 2007].

5.1. Limitations of the Construct and Operationalization of GAB

As a construct, GAB describes a team's ability to regulate the affective balance of positive to negative expressed affect during conflict, such that a surplus of positive affect is maintained, by assessing the dynamics of emotional behaviors occurring in the interactions of a team. Analogous to many of the marital interaction studies, this construct was operationalized through a single behavioral or experiential summary statistic. While using a single group-mean summary statistic to operationalize the affective balance of a team's conflict style is simple and effective, there are important limitations that I will discuss.

First, a single summary statistic does not fully capture the temporal dynamics of a team's emotional behaviors. While a subtraction-based score, a ratio-score, or even the slope of a point graph capture some aspect of the temporal dynamics inherent in a group's emotional behavior, potentially important aspects might be missed. For example, a team with a steady surplus of positive over negative emotional expressions might have the same GAP score as a team that has emotional ups and downs, if their overall counts of positive and negative behaviors are the same. Similarly, the current operationalization of GAP does not account for differences in startup dynamics: Rather than the overall relationship between positive and negative affect it might be important to distinguish how teams start their conflict conversation. Studies on couples found that the startup of a conflict conversation is particularly informative [Carrere and Gottman 1999]. Point graphs, such as the ones shown in Figure 9, provide a rich perspective on the temporal dynamics of affect in groups. An alternative to using a single summary statistic to describe interaction dynamics could therefore be approaches to model the particular shape of a point graph, or to examine behavior sequences through sequential analysis techniques [Bakeman and Gottman 1997]. The latter could also be used to study the temporal impact of specific hostile behaviors, such as contempt, within a group.

Second, a summary statistic based on the group mean, as used for the operationalization of GAP, does not adequately capture the group's specific emotional configuration. For example, two of the point graph slopes constructed in Study 2 showed a pattern in which one team member's expression of emotion diverged substantially from that of the other group members (Figure 9). A mean-based summary statistic does not capture this diverging behavior of a single individual. However, capturing such divergence might be important, as several studies have shown that the behavior of a single individual shows contagious properties within a group and can impact a group's emotional interaction dynamics and performance in both positive and negative ways [Barsade 2002]. For example, several researchers have argued that it takes only one "bad apple" in the team to poison a team's interpersonal dynamics and ultimately its performance [De Jong et al. 2014; Felps et al. 2006]. Having a measure that distinguishes teams that have a "bad apple" from those that don't, could be very helpful. Alternative operationalizations of GAP might, therefore, employ statistics that capture the variance, convergence or temporal alignment of point-graph slopes, affect ratios, or subtractions rather than using their group level averages. Another alternative could be to distinguish between different types of interaction dynamics based on a set of theory-driven rules as has been done in some of the marital studies. For example, Gottman and Levenson [1992] distinguished between regulated and nonregulated couples based on a couple's point graph slope configuration. Operationally, for a couple to be "regulated" both point-graph slopes had to be significantly positive—if only one point graph slope is negative, a couple is classified as nonregulated. The same categorization has also been used to categorize pair programming teams [Jung et al. 2012] and this categorization approach could possibly be extended to work for teams with more than two members.

Finally, capturing the balance of positive and negative affect to assess a group's ability rests on the assumption that a "good" group interaction is the result of successful regulation and GAP does not explicitly capture a group's ability to regulate affect. An alternative approach to measure a group's ability to regulate affect might, therefore, be to assess the likelihood for a group to recover after a particularly negative interaction episode, or after the occurrence of hostile behaviors within the group.

5.2. Conflict Elicitation Protocol

With the Group Interaction Task, this work also contributes a conflict elicitation protocol to sample interaction dynamics of teams in the field, but under controlled laboratory conditions. It was designed based on the conflict elicitation protocol for couples [Roberts et al. 2007] and it *"compresses the as-lived experience of group into a sort of in vitro laboratory-based 'espresso' suitable for various approaches to instrumentation and data collection."*[2] The central idea of the approach is that conflict episodes are particularly informative about the quality of a team's interactions, which is also why so many studies of teams have focused on conflict as a core process, albeit almost exclusively only through self-report assessments (for reviews, see De Dreu and Weingart [2003], Weingart et al. [2015], and De Wit et al. [2012]). Rather than continuously shadowing a team with recording equipment and waiting for conflict to occur naturally, this procedure allows sampling an episode of conflict at a time and location specified by the experimenter. Also the elicitation approach allows for an optimal recording setup, which would not be possible in the field.

5.2.1. Novelty of the Approach. The ability to study intragroup or team conflict in the lab is novel. To my knowledge, no study has so far described or used a similar

[2]I am quoting here a comment of one of my reviewers, who summarized the contribution of this approach better than I have been able to.

thin-slicing approach to intragroup conflict that allows studying groups engaged in a conflict episode under laboratory conditions. The only observational accounts of intragroup conflict episodes are those made during long ethnographic observations (e.g., Jehn [1997]) and such qualitative accounts are not amenable to a study of moment-to-moment temporal dynamics of affect described here. Emotionally volatile conflict episodes are also almost impossible to elicit in the laboratory with established laboratory tasks such as the desert survival task [Lafferty and Eady 1974] or negotiation tasks. Therefore, this method could open exciting new avenues for research on intragroup conflict, emotional dynamics in teams, and for designing and testing conflict intervention and feedback systems.

5.2.2. Ecological Validity. Eliciting conflict between people who have an established relationship also makes findings about the interaction dynamics of teams more ecologically valid for two reasons, as argued by Roberts et al. [2007]: First, the elicited conflict relies on an ongoing emotional relationship between individuals who have developed a specific conflict engagement style that surfaces across contexts. Second, by standardizing the elicitation task by asking each team to discuss a topic that generates the greatest amount of disagreement, each team discusses an equally meaningful topic leading to interactions are comparable across interaction settings and likely to interactions outside the laboratory. In addition to these arguments, there is evidence that the patterns of emotional interaction dynamics found in conflict interactions generalize to regular nonconflict interactions. For example, Driver and Gottman [2004] found that the finding of the importance of balancing positive and negative effect in conflict interactions in the lab generalizes to regular dinner interactions in the home.

As intended, the Group Interaction Task was particularly successful in establishing a high level of emotional engagement during the main problem discussion session. Despite the four cameras and the microphone on the table, some groups engaged in interactions that were surprisingly vulnerable. The following transcript was made from one team's problem discussion (Emotions coded with each utterance are listed in brackets) and shows an example of how an interaction escalates in negativity from expressions of domineering and frustration to expressions of contempt and belligerence. The group is discussing participant A's failure to complete a part of the work in time.

Team members A, B, and C are discussing A's failure to deliver on time.

(1) A *Oh yeah, ok.*
(2) B *And you said you hadn't started yet until Tuesday.*
[Domineering, frustration]
(3) A *Ok, yeah. So that was a mistake. Uhm. Yeah, I should have had that done earlier.*
[Tension, validation]
(4) B *But I don't . . . . I really think the gant chart is just an example and it's not, it's indicative of your actions in general.*
[Domineering]
(5) A *Rolls eyes.*
[Micromoment Contempt]
(6) B *I don't think we need to make excuses for each particular thing. I think we need to talk about what we need to do to change this pattern of behavior. Period.*
[Domineering]
(7) A *Ok. So. Let me tell you why I didn't do it Saturday!*
[Defensiveness]
(8) B *We know. I don't think excuses are a good idea, period.*
[Frustration, domineering]

(9) A *I, I feel it would be helpful for this discussion.*
[Empathy]
(10) C *Well if he, if he wants to tell us.*
[Neutral, interest]
(11) B *I don't. You remember what Bernie said?*
[Belligerence]
(12) A *Laughs.*
[Contempt]
(13) B *Reasons are bullshit.*
[Domineering, Contempt]

While a transcript can only partially reflect the vividness of the conflict interaction the statements the participants make combined with the affect labels give an insight into the kind of conflict interactions and emotions that could be observed based on the group interaction task.

5.2.3. Thin-Slicing Effort. It is important to distinguish the effort to generate a 15-minute conflict sample from the effort to analyze it. Early thin-slicing literature predominantly employed lay people making brief and quick judgments on a set of predefined categories. Thin slicing in combination with a careful and detailed analysis of behavior is consistent with more recent uses of the term that focus on the length of the slice as a distinguishing criteria rather than the granularity of the analysis. For example, Curhan and Pentland [2007], referred to thin slicing when generating detailed behavioral accounts of behavioral interaction dynamics of negotiations using automated measures. Also Gottman's research on marital dynamics, which uses a similar effort in analyzing video, is often included in reviews of "thin slicing" (e.g., in Curhan and Pentland [2007]). Further, Ambady and Rosenthal [1992, 1993] introduced thin slicing as making accurate "judgments about people on the basis of brief exposures to them" [Ambady and Rosenthal 1992] and the research presented here fits within the criteria of brief exposure: Given the assumption that each team spent roughly 10 hours per week together (a conservative estimate given that teams spent up to 40 hours per week ahead of deadlines), a 15-minute interaction sample only constituted only about 0.08% of the total time team members spent together over the 8-month duration of the project. Despite the relatively small effort to generate such a thin slice of only 15 minutes, these interactions turned out to be highly informative. The brevity of the exposure to a team's interaction dynamics further distinguishes this study from others that collected data about team dynamics over the entire duration of a team's life (e.g., Tausczik and Pennebaker [2013]).

A key limitation of the approach used here was the effort to analyze the thin slices of conflict behavior using systematic observation of behavior. It took up to 30 hours per team to generate the interaction data, which is far too long for any immediate practical application. However, the controlled setting of the interaction task makes an automated instrumentation more accessible through sensor-based technology as employed in studies by Curhan and Pentland [2007] who used a wearable sensor badge or by Won et al. [2014] who used two Kinect movement sensors to capture interaction dynamics. Especially studies that rely on sensors that need to be installed in the environment such as the Kinect sensors can benefit from the use of an elicitation protocol as it lets researchers select the exact time and place of observation. Even without sophisticated sensor-based technology, the rating dial procedure demonstrated that emotional dynamics of the video can be assessed with just 15minutes of effort.

5.2.3. Reflective Impact. Finally, the interaction session was not only successful in providing an accurate sample of a team's conflict solving style it also had immediate

benefits for the team as it helped them to reflect about their interactions. The teams utilized the interaction session to discuss a wide range of different topics covering disagreements on process, their relationship, and the task they were working on. Most teams reported the interaction session to be directly valuable as a learning opportunity and experienced it as immediately useful since several teams asked whether they could continue their discussion beyond the 15 minute limit.

5.3. Implications for the Design of Team Feedback Systems

By highlighting the role of GAB and hostile affect, the findings from the two studies have direct implications for systems that support teamwork whether it is online or face-to-face. The majority of team feedback systems have relied on providing information about participation patterns without focusing directly on emotional interaction dynamics. For example, the conversation clock [Bergstrom and Karahalios 2007] visualized speaking time patterns to participants. DiMicco et al. [2004] used a shared display to influence Group participation patterns by visualizing whether a group member under or over participated. Another study by Kim et al. [2008] introduced the meeting mediator to shape participation patterns by showing the degree of interactivity and balance in group members' interactions. Other systems that focus on giving feedback about participation patterns include "Babble" by Erickson et al. [1999], "Chat Circles" by Viégas and Donath [1999] or the "Participation Tool" by Janssen et al. [2007]. While some feedback systems have focused on affect to some degree (e.g., Tausczik and Pennebaker [2013], Nowak et al. [2012], Leshed et al. [2009], and Kim et al. [2008]), none of these systems have considered giving feedback about hostile affect or measuring and presenting the balance of positive and negative affect. The work presented here, however, suggests that it might be effective for feedback tools to focus on balance between positive and negative affect as well as on a small set of specific hostile behaviors.

In implementing automated feedback systems, the difficulty of automated assessment of the patterns described here has to be considered. While the techniques used here to generate the hostile affect assessment are highly time consuming (while it required only 15 minutes for participants to self-code their interactions for a GAB assessment it required 6–8 hours per participant to code just 15 minutes of data using the modified SPAFF coding scheme), robust toolkits are available that measure positive and negative affect from facial or prosodic changes. Even detection of specific affect expressions is becoming increasingly robust as demonstrated in studies by Black et al. [2013] on emotion expression recognition in marital conflict interactions. With an automated assessment of affect, and based on the findings presented here that how emotions are managed during conflict interactions is highly diagnostic of future team performance three recommendations can be made.

First, designers of feedback tools should consider how feedback can be used to support long-term emotion regulation skills development for teams. A recent study by Finkel et al. [2013] showed that interventions to build emotion regulation skills can lead to long term improvements in interactions. The authors demonstrated a causal link between the ability to regulate affect and marital outcomes over time. A 15-minute emotion regulation intervention had a measureable impact on preventing relationship quality from deteriorating three years later. For example, a feedback system could, therefore, monitor the emotional dynamics in a team and propose specific emotion regulation strategies that group members can learn to employ over time.

Second, as this study, and a previous study [Jung et al. 2012] showed, characteristic patterns in a team's emotional interaction dynamics that foreshadow a team's future performance are often visible early on during teamwork. Feedback systems

should, therefore, consider particularly focusing on providing feedback in early phases of teamwork when team members shape their interaction styles.

Finally, feedback systems might benefit from targeting conflict interactions as opposed to general team meetings. As demonstrated by recent studies, feedback systems can productively intervene in intragroup conflict through direct, and targeted interventions. For example, Hoffman et al. [2015] demonstrated that a robot could positively influence how couples engage in conflict by provoking an empathic response towards the robot at moments of heightened emotional arousal. Particularly interesting about this study is the peripheral role of the robot, demonstrating that technology can moderate conflict interactions through subtle almost ambient interventions. Another study by Jung et al. [2015] explored whether a robot could respond productively to expressions of hostility, thereby, de-regulating their negative impact. The authors showed that the robot could regulate intragroup conflict processes and specifically the perception of conflict and the emotion experience by intervening after a team member's hostile remark. Together these two studies are a promising start into thinking about the role of technology in actively shaping conflict processes and socio-emotional interaction dynamics more broadly.

5.4. Implications for Our Understanding of Design Teamwork

This study contributes to our understanding of designing in teams as socially mediated activity [Bucciarelli 1988; Minneman 1991] as it highlights the importance of emotional interaction dynamics during design teamwork and especially during conflict. While other studies have investigated the relationship between social dynamics and performance in design teams, these studies focused predominantly on cognitive or task-related processes such as question asking, e.g., Eris [2004], gesturing, e.g., Tang [1989] and Tang [1991], process changes, e.g., Frankenberger and Auer [1997], prototyping, e.g., Dow et al. [2011], negotiation [Minneman 1991], and others (for a review, see Finger and Dixon [1989a, 1989b]). However, not much has been done to look at the role of emotions in designing, their dynamics in design teams, and how the emotions designers express, or feel, shape subsequent performance relevant outcomes. The past research emotions in design has almost exclusively looked at how we design products to elicit certain emotions, e.g., Desmet [2003] and Norman [2004]. This is especially surprising given a growing literature that highlights the importance of emotions for the performance of teamwork, e.g., Barsade and Gibson [2007] and Kelly and Barsade [2001] and creative problem solving in general, e.g., Isen et al. [1987].

If we accept the role of emotional interaction dynamics and a team's behavior in shaping those dynamics as important for design practice, it is important to consider these dynamics when teaching design. The insights gained about the importance of a group's affective balance and hostile affect pattern can be used to help students learn how to manage conflict in design teamwork effectively. To build capacities in teams to effectively manage or regulate emotions during conflict two things are important: first, recognition of dysfunctional patterns, and second, development of intervention strategies. Students can learn to recognize hostile affect (especially contempt, criticism, defensiveness, and stonewalling) and develop an awareness of those patterns as harmful. With the ability to recognize dysfunctional emotional behaviors, students can then be taught to develop effective intervention and repair strategies to prevent negative behaviors from their tendency to escalate [Anderson and Pearsson 1999].

5.5. Socio-Emotional Dynamics Support for CSCW Systems

This research has implications for the design of groupware and CSCW applications more broadly. How emotions are managed or regulated is not only central for the quality of marital interactions but also for teamwork. Yet, most systems that support

teamwork and collaboration, whether it is in face to face or distributed contexts, focus on supporting task oriented processes rather than socio-emotional team processes. An example of an existing body of the theory that has motivated many collaborative work support systems is the coordination theory [Malone and Crowston 1990; Malone and Crowston 1994]. The Coordination theory has shaped how we build and think about building systems that support team-based crowdwork [Kittur et al. 2013], teamwork with robots, Johnson et al. [2014], cross team collaboration systems [Dabbish et al. 2010], to distributed collaboration systems [Nomura et al. 2008; Kittur et al. 2009]. According to Malone and Crowston [1990, p. 359], in the Coordination theory *"the common problems have to do with coordination: How can overall goals be subdivided into actions? How can actions be assigned to groups or to individual actors? How can resources be allocated among different actors? How can information be shared among different actors to help achieve the overall goals?"* This perspective does not conceptualize emotion and emotion regulatory processes as central to collaborative task achievement, and therefore, systems drawing from the coordination theory have focused predominantly on shaping how tasks task division, achievement tracking, and decision support.

In contrast to the coordination theory that places the coordination of tasks, resources, and goals at the center of teamwork, the findings of this study place the coordination of affect at the center of teamwork. By coordination of affect, I refer to a team's coordination efforts in shaping the affective quality of an interaction, for example, by regulating an interaction such that is characterized by a surplus of positive behaviors or by repairing negative and hostile behaviors to prevent them from escalating. Consider line (10) in the transcript shown in Section 5.2.2. In this example, group member [C] attempts to repair the increasing hostility expressed in interactions of [A] and [B] by expressing empathy to one of the team members "*Well if he, if he wants to tell us".* Repairing negativity has been shown to be crucial as it prevents it from escalating [Anderson and Pearsson 1999].

Coordinating the emotional quality of an interaction is important, because according to the Emotions as Social Information model [Van Kleef 2009], emotional expressions affect people interpersonally in two important ways. First, emotional expressions trigger direct affective reactions in others thereby influencing not only another person's likely reaction (such as the escalation of conflict through a defensive or hostile reaction to a perceived attack), but also cognitive processes involved in decision making and creative problem solving, e.g., Carnevale and Isen [1986] and Daubman et al. [1987]. Second, emotional expressions trigger inferential processes in others, as people build an understanding about a situation, another person's intentions, or how people are related to each other based on the emotional meaning inherent in a person's behavior. CSCW systems in support of groupwork, or groupware such as workflow management systems, group chat applications, or even crowdsourcing applications that employ groups, e.g., Retelny et al. [2014] mediate both pathways: The systems shape how people perceive each other and how they can respond to each other based on their reactions to and inferences about the other persons in the group. For example, a typical workflow management system that structures the interactions between multiple people in submitting, checking, and approving a reimbursement does not allow participants to accurately assess the emotional states of the other participants involved in the group process nor does it allow for effective intervention and repair. One person's frustration likely remains unnoticed, and therefore, unrepaired by other group members, which might lead to an escalation of frustration [Andersson and Pearson 1999] and likely impair future collaborations. This relationship between the quality of interpersonal interactions and the quality of future collaborations has been shown, for example, in the context of negotiations. Curhan et al. [2010] found that the social and emotional

quality of a negotiation predicted not only subjective but also objective negotiation performance in future negotiations, while objective negotiation performance was unrelated with future performance. Given this relationship between the socio-emotional quality of an interaction and its future performance it is important for designers of CSCW applications to consider whether and how group members are able to not only assess each other's emotional behaviors but also whether they are able to intervene and repair when negative behaviors occur.

6. CONCLUSION

To my understanding, the studies presented here are the first to apply theory about conflict in marital interactions to further our understanding of conflict and performance in teamwork. The finding that two specific patterns of conflict that distinguish functional from dysfunctional marriages also distinguish functional from dysfunctional teams has broad implications for how we study teams and how we build interactive systems to support them.

To conclude, I want to reiterate a quote from Tausczik and Pennebaker's [2013, p. 459] article: "*Shaping group dynamics relies on understanding the basic question: Why do some groups of people work well together while others do not? Despite the substantial amount of research on groups, there is surprising little consistent evidence for which group processes promote good group outcomes*." I hope this article sheds new light onto this question.

AUTHOR STATEMENT

None of the empirical finding presented in this article have been published before.

The research most closely related to this work is a previous study on emotional interaction dynamics in pair programming teamwork (see citation below). The research of this article directly builds upon this work by extending the theoretical argument and an entirely new dataset and analysis.

Jung, M., Chong, J., & Leifer, L. Group hedonic balance and pair programming performance: affective interaction dynamics as indicators of performance. In Proc. CHI 2012, ACM Press (2012), 829–838.

Also, some aspects of the methodological approach have been published before:

Jung, M. and L. Leifer (2011). A method to study affective dynamics and performance in engineering design teams. In Proceedings of the International Conference on Engineering Design ICED'11. Copenhagen, Denmark.

None of this work is currently under review at any other venue.

ACKNOWLEDGMENTS

This work would not have been possible without the inspiration and support from Larry Leifer and Janine Giese-Davis. I also want to thank Pamela Hinds, James Gross, Clifford Nass, and Martin Steinert for their helpful comments on earlier versions of this article. Further, I want to acknowledge the countless discussions with Neeraj Sonalkar, and Ade Mabogunje that helped shape this work. Several students helped code the video data and I would like to thank especially Julia Tang and Daniel Lopez for their help. Finally, I want to thank the editor Gloria Mark, and three anonymous reviewers for their invaluable comments and feedback.

REFERENCES

A. C. Amason. 1996. Distinguishing the effects of functional and dysfunctional conflict on strategic decision making: resolving a paradox for top management teams. *Acad. Manag. J.* 39, 1, 123–148.

N. Ambady and R. Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychol. Bull.* 111, 2 (1992), 256–274.

N. Ambady and R. Rosenthal. 1993. Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *J. Pers. Soc. Psychol.* 64, 3 (1993), 431–441.

L. M. Andersson and C. M. Pearson. 1999. Tit for tat? The spiraling effect of incivility in the workplace. *Acad. Manag. Rev.*, 24, 3, 452–471.

J. Apesteguia, G. Azmat, and N. Iriberri. 2012. The impact of gender composition on team performance and decision making: evidence from the field. *Manag. Sci.*, 58, 1, 78–93.

R. Bakeman and J. M. Gottman. 1997. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.

S. G. Barsade. 2002. The ripple effect: emotional contagion and its influence on group behavior. *Administrative Sci. Quarterly* 47, 4, 644–675.

S. G. Barsade and D. E. Gibson. 2007. Why does affect matter in organizations? *Acad. Manag. Perspect.*, 21, 1, 36–59.

T. Bergstrom and K. Karahalios. 2007. Conversation clock: visualizing audio patterns in co-located groups. In *Proceedings of the HICSS'07: 40th Annual Hawaii International Conference on System Sciences*, Waikoloa, Big Island, HI. 78–86.

M. P. Black, A. Katsamanis, B. R. Baucom, C. C. Lee, A. C. Lammert, A. Christensen, and S. S. Narayanan. 2013. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Commun.* 55, 1, 1–21.

L. L. Bucciarelli. 1988. An ethnographic perspective on engineering design. *Des. Stud.* 9, 3, 159–168.

T. Carleton and L. Leifer. 2009. (March). Stanford's ME310 course as an evolution of engineering design. In *Proceedings of the 19th CIRP Design Conference–Competitive Design*. Cranfield University Press.

P. J. Carnevale and A. M. Isen. 1986. The influence of positive affect and visual access on the discovery of integrative solutions in bilateral negotiation. *Organizational Behav. Human Decision Processes* 37, 1, 1–13.

S. Carrere and J. M. Gottman. 1999. Predicting divorce among newlyweds from the first three minutes of a marital conflict discussion. *Family Process* 38, 3, 293–301.

J. A. Coan and J. M. Gottman. 2007. The specific affect coding system (SPAFF). In *Handbook of Emotion Elicitation and Assessment*, J. A. Coan and J. J. B. Allen (Eds). Series in affective science. Oxford University Press, New York, NY, US, 267–285.

J. A. Cohen. 1960. Coefficient of agreement for nominal scales. *Educational Psychol. Measurement* 20, 37–46.

J. R. Curhan, H. A. Elfenbein, and H. Xu. 2006. What do people value when they negotiate? Mapping the domain of subjective value in negotiation. *J. Pers. Soc. Psychol.* 91, 3, 493–709.

J. R. Curhan and A. Pentland. 2007. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *J. Appl. Psychol.* 92, 3, 802–811.

P. L. Curseu, S. Boros, and L. A. Oerlemans. 2012. Task and relationship conflict in short-term and long-term groups: the critical role of emotion regulation. *Int. J. Conflict Manag.* 23, 1, 97–107.

L. A. Dabbish, P. Wagstrom, A. Sarma, and J. D. Herbsleb. 2010. Coordination in innovative design and engineering: observations from a lunar robotics project. In *Proceedings of the Group'10*. ACM Press, 225–234.

C. K. De Dreu and L. R. Weingart. 2003. Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis. *J. Appl. Psychol.* 88, 4, 741–749.

J. P. De Jong, P. L. Curşeu, and R. T. A. Leenders. 2014. When do bad apples not spoil the barrel? Negative relationships in teams, team performance, and buffering mechanisms. *J. Appl. Psychol.* 99, 3, 514–522.

F. R. De Wit, L. L. Greer, and K. A. Jehn. 2012. The paradox of intragroup conflict: a meta-analysis. *J. Appl. Psychol.* 97, 2, 360–390.

P. Desmet. 2003. A multilayered model of product emotions. *Design J.* 6, 2, 4–13.

J. M. DiMicco, A. Pandolfo, and W. Bender. 2004. Influencing group participation with a shared display. In *Proceedings of the CSCW 2004*. ACM Press, 614–623.

J. M. DiMicco, K. J. Hollenbach, A. Pandolfo, and W. Bender. 2007. The impact of increased awareness while face-to-face. *Hum.-Comput. Interact.*, 22, 1, 47–96.

S. Dow, J. Fortuna, D. Schwartz, B. Altringer, D. Schwartz, and S. Klemmer. 2011. Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, (May 2011), 2807–2816.

J. L. Driver and J. M. Gottman. 2004. Daily marital interactions and positive affect during marital conflict among newlywed couples. *Family Process* 43, 3, 301–314.

P. Ekman and W. V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, CA.

Ö. Eris. 2004. *Effective Inquiry for Innovative Engineering Design*. Kluwer Academic Publishers, Norwell, MA.

T. Erickson, D. N. Smith, W. A. Kellogg, M. R. Laff, J. T. Richards, and E. Bradner. 1999. Socially translucent systems: social proxies, persistent conversation, and the design of 'babble'. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Pittsburgh, PA. 72–79.

F. Faul, E. Erdfelder, A. Buchner, and A. G. Lang. 2009. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160.

W. Felps, T. R. Mitchell, and E. Byington. 2006. How, when, and why bad apples spoil the barrel: negative group members and dysfunctional groups. *Res. Organizational Behav.* 27, 175–222.

S. Finger and J. R. Dixon. 1989a. A review of research in mechanical engineering design. Part I: descriptive, prescriptive, and computer-based models of design processes. *Res. Eng. Des.*, 1, 1, 51–67.

S. Finger and J. R. Dixon. 1989b. A review of research in mechanical engineering design. Part II: representations, analysis, and design for the life cycle. *Res. Eng. Des.* 1, 2, 121–137.

E. J. Finkel, E. B. Slotter, L. B. Luchies, G. M. Walton, and J. J. Gross. 2013. A brief intervention to promote conflict reappraisal preserves marital quality over time. *Psychol. Sci.* 24, 8, 1595–1601.

E. Frankenberger and P. Auer. 1997. Standardized observation of team-work in design. *Res. Eng. Des.* 9, 1, 1–9.

S. R. Fussell, R. E. Kraut, F. J. Lerch, W. L. Scherlis, M. M. McNally, and J. J. Cadiz. 1998. Coordination, overload and team performance: effects of team communication strategies. In *Proceedings of the CSCW 1998*. ACM Press, 275–284.

J. Giese-Davis, K. A. Piemme, C. Dillon, and S. Twirbutt. 2005. Macrovariables in affective expression in women with breast cancer participating in support groups. In *The New Handbook of Methods in Nonverbal Behavior Research,* J. A. Harrigan, R. Rosenthal, and K. R. Scherer (Eds.). Oxford University Press, Oxford, U.K, 397–446.

A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker. 2009. Language style matching as a predictor of social dynamics in small groups. *Commun. Res.* 37, 1, 3–19.

J. M. Gottman. 1994. *What Predicts Divorce?* Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

J. M. Gottman and R. W. Levenson. 1985. A valid procedure for obtaining self-report of affect in marital interaction. *J. Consulting Clin. Psychol.* 53, 151–160.

J. M. Gottman and R. W. Levenson. 1992. Marital processes predictive of later dissolution: behavior, physiology, and health. *J. Pers. Soc. Psychol.* 63, 2, 221–233.

J. M. Gottman and R. W. Levenson. 2000. The timing of divorce: predicting when a couple will divorce over a 14-year period. *J. Marriage Family* 62, 3, 737–745.

J. R. Hackman. 2002. *Leading Teams: Setting the Stage for Great Performances*. Harvard Business School Press, Boston.

J. R. Hackman and M. O'Connor. 2004. What makes for a great analytic team? Individual vs. *Team Approaches to Intelligence Analysis*. Intelligence Science Board, Office of the Director of Central Intelligence, Washington, DC.

J. Hagedorn, J. Hailpern, and K. G. Karahalios. 2008. VCode and VData: illustrating a new framework for supporting the video annotation workflow. In *Proceedings of the AVI 2008*. ACM Press, 317–321.

G. Hoffman, O. Zuckerman, G. Hirschberger, M. Luria, and T. Shani-Sherman. 2015. Design and evaluation of a peripheral robotic conversation companion. In *Proceedings of the HRI'15*. ACM Press.

A. M. Isen, K. A. Daubman, and G. P. Nowicki. 1987. Positive affect facilitates creative problem solving. *J. Pers. and Soc. Psychol.* 52, 6, 1122–1131.

J. Janssen, G. Erkens, G. Kanselaar, and J. Jaspers. 2007. Visualization of participation: does it contribute to successful computer-supported collaborative learning? *Comput. Education* 49, 4, 1037–1065.

K. A. Jehn. 1994. Enhancing effectiveness: an investigation of advantages and disadvantages of value-based intragroup conflict. *Int. J. Conflict Manag.* 5, 3, 223–238.

K. A. Jehn. 1995. A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Sci. Quarterly* 40, 2, 1–28.

K. A. Jehn. 1997. A qualitative analysis of conflict types and dimensions in organizational groups. *Administrative Sci. Quarterly*, 42, 530–555.

M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis. 2014. Coactive design: designing support for interdependence in joint activity. *J. Hum.-Robot Interact.* 3, 1, 43–69.

M. Jung, J. Chong, and L. Leifer. 2012. Group hedonic balance and pair programming performance: affective interaction dynamics as indicators of performance. In *Proceedings of the CHI 2012*. ACM Press, 829–838.

M. F. Jung, N. Martelaro, and P. J. Hinds. 2015. Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the HRI'15*. ACM Press, 229–236.

J. R. Kelly and S. G. Barsade. 2001. Mood and emotions in small groups and work teams. *Organizational Behav. Hum. Decis. Processes* 86, 1, 99–130.

T. Kim, A. Chang, L. Holland, and A. S. Pentland. 2008. Meeting mediator: enhancing group collaboration using sociometric feedback. In *Proceedings of the CSCW 2008*. ACM Press, 457–466.

A. Kittur, B. Lee, and R. E. Kraut. 2009. Coordination in collective intelligence: the role of team structure and task interdependence. In *Proceedings of the CHI'09*. ACM Press, 1495–1504.

A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton. 2013. The future of crowd work. In *Proceedings of the CSCW'13*. ACM Press, 1301–1318.

K. J. Klein and S. W. Kozlowski. 2000. From micro to meso: critical steps in conceptualizing and conducting multilevel research. *Organizational Res. Methods* 3, 3, 211–236.

S. W. J. Kozlowski and K. J. Klein. 2000. A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*, K. J. Klein and S. W. J. Kozlowski (Eds.). Jossey-Bass, San Francisco, CA, US, 3–90.

J. C. Lafferty and P. M. Eady. 1977. *The Desert Sur Vival Problem.* Experimental Learning Methods, Plymouth, MI.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics* 33, 1, 159–174.

G. Leshed, D. Perez, J. T. Hancock, D. Cosley, J. Birnholtz, S. Lee, P. L. McLeod, and G. Gay. 2009. Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. In *Proceedings of the CHI 2009*. ACM Press, 537–546.

R. W. Levenson, L. L. Carstensen, and J. M. Gottman. 1994. Influence of age and gender on affect, physiology, and their interrelations: a study of long-term marriages. *J. Pers. Soc. Psychol.* 67, 1, 56–68.

R. W. Levenson and J. M. Gottman. 1983. Marital interaction: physiological linkage and affective exchange. *J. Pers. Soc. Psychol.* 45, 3, 587–597.

R. W. Levenson and J. M. Gottman. 1985. Physiological and affective predictors of change in relationship satisfaction. *J. Pers. Soc. Psychol.* 49, 1, 85–94.

T. W. Malone and K. Crowston. 1990. What is coordination theory and how can it help design cooperative work systems? In *Proceedings of the CSCW 1990*. ACM Press, 357–370.

T. W. Malone and K. Crowston. 1994. The interdisciplinary study of coordination. *ACM Comput. Surv.* 26, 1, 87–119.

S. L. Minneman. 1991. The social construction of a technical reality: empirical studies of group engineering design practice. Doctoral dissertation, Mechanical Engineering, Stanford University.

S. A. Munson, K. Kervin, and L. P. Robert. 2014. Monitoring email to indicate project team performance and mutual attraction. In *Proceedings of the CSCW 2014*. ACM Press, 542–549.

S. Nomura, J. Birnholtz, O. Rieger, G. Leshed, D. Trumbull, and G. Gay. 2008. Cutting into collaboration: understanding coordination in distributed and interdisciplinary medical research. In *Proceedings of the CSCW'08*. ACM Press, 427–436.

D. A. Norman. 2004. *Emotional Design: Why We Love (Or Hate) Everyday Things*. Basic Books, New York, NY.

M. Nowak, J. Kim, N. W. Kim, and C. Nass. 2012. Social visualization and negotiation: effects of feedback configuration and status. In *Proceedings of the CSCW 2012*. ACM Press, 1081–1090.

S. B. Paletz, C. D. Schunn, and K. H. Kim. 2013. The interplay of conflict and analogy in multidisciplinary teams. *Cognition*, 126, 1, 1–19.

S. B. Paletz, C. D. Schunn, and K. H. Kim. 2011. Intragroup conflict under the microscope: micro-conflicts in naturalistic team discussions. *Negotiation Conflict Manag. Res.* 4, 4, 314–351.

D. Retelny, S. Robaszkiewicz, A. To, W. S. Lasecki, J. Patel, N. Rahmati, and M. S. Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. ACM, 75–85.

N. A. Roberts, J. L. Tsai, and J. A. Coan. 2007. Emotion elicitation using dyadic interaction tasks. In *Handbook of Emotion Elicitation and Assessment*, J. A. Coan and J. J. B. Allen (Eds.). Series in affective science. Oxford University Press, New York, NY, US, 106–123.

A. M. Ruef and R. W. Levenson. 2007. Continuous measurement of emotion: The affect rating dial. In *Handbook of Emotion Elicitation and Assessment*, J. A. Coan and J. J. B. Allen (Eds.). Series in affective science. Oxford University Press, New York, NY, US, 286–297.

P. Salovey and J. D. Mayer. 1989. Emotional intelligence. *Imagin., Cognit. Pers.*, 9, 3, 185–211.

M. Sridharan, S. J. Fink, and R. Bodik. 2007. Thin slicing. In *ACM SIGPLAN Notices*, ACM, 42, 6(Jun. 2007), 112–122.

K. B. Stecher and S. Counts. 2008. Thin slices of online profile attributes. In *ICWSM*. (Mar. 2008).

J. C. Tang. 1989. Gesturing in design: a study of the use of shared workspaces by design teams. Ph.D. dissertation, Stanford University

J. C. Tang. 1991. Findings from observational studies of collaborative work. *Int. J. Man-Mach. Stud.* 34, 2, 143–160.

Y. R. Tausczik and J. W. Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the CSCW 2013*. ACM Press, 459–468.

P. Tripathi and W. Burleson. 2012. Predicting creativity in the wild: experience sample and sociometric modeling of teams. In *Proceedings of the CSCW 2012*. ACM Press, 1203–1212.

G. A. Van Kleef. 2009. How emotions regulate social life the emotions as social information (EASI) model. *Curr. Directions Psychol. Sci.*, 18, 3, 184–188.

F. B. Viégas and J. S. Donath. 1999. Chat circles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, PA. 9–16.

R. Wageman, J. R. Hackman, and E. Lehman. 2005. Team diagnostic survey development of an instrument. *J. Appl. Behavioral Sci.* 41, 4, 373–398.

L. R. Weingart, M. Olekalns, and P. L. Smith. 2004. Quantitative coding of negotiation behavior. *Int. Negotiation* 9, 3, 441–456.

L. R. Weingart, K. J. Behfar, C. Bendersky, G. Todorova, and K. A. Jehn. 2015. The directness and oppositional intensity of conflict expression. *Academy of Management Review* 40, 2 (2015), 235–262.

A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Sci.* 330, 6004, 686–688.

A. S. Won, J. N. Bailenson, S. C. Stathatos, and W. Dai. 2014. Automatically detected nonverbal behavior predicts creativity in collaborating dyads. *J. Nonverbal Behav.* 38, 3, 389–408.

Received March 2015; revised March 2016; accepted April 2016