Using sensor technology to capture the structure and content of team interactions in medical emergency teams during stressful moments

Maaike Endedijk[a*], Marcella Hoogeboom[a*], Marleen Groenier[a], Stijn de Laat[a], Jolien van Sas[a]

[a] University of Twente, The Netherlands.
*both authors contributed equally to this work

Abstract

In healthcare, action teams are carrying out complex medical procedures in intense and unpredictable situations to save lives. Previous research has shown that efficient communication, high-quality coordination, and coping with stress are particularly essential for high performance. However, precisely and objectively capturing these team interactions during stressful moments remains a challenge. In this study, we used a multimodal design to capture the structure and content of team interactions of medical teams at moments of high arousal during a simulated crisis situation. Sociometric badges were used to measure the structure of team interactions, including speaking time, overlapping speech and conversational imbalance. Video coding was used to reveal the content of the team interactions. Furthermore, the Empatica E4 was used to unobtrusively measure the team leader's skin conductance to identify moments of high arousal. In total, 21 four-person teams of technical medicine students in the Netherlands were monitored in a simulation environment while they diagnosed and managed a patient with cardiac arrest. Outcomes of this exploratory study revealed that more effective teams showed greater conversational imbalance than less effective teams, but during moments of high arousal the opposite was found. Also, a number of differences were found for the content of team interaction. Combining sensor technology with traditional measures can enhance our understanding of the complex interaction processes underlying effective team performance, but technological advances together with more knowledge about the simultaneous application of these methods are needed to tap into the full potential of wearable sensor technology in team research.

Keywords: team interaction; video observation; skin conductance; sociometric badges; medical simulation; action teams.

Corresponding authors: A.M.G. Hoogeboom a.m.g.m.hoogeboom@utwente.nl &Maaike Endedijk m.d.endedijk@utwente.nl University of Twente, Drienerlolaan 5 PO Box 217 7500 AE Enschede a.m.g.m.hoogeboom@utwente.nl DOI:*https://doi.org/10.14786/flr.v6i3.353*

1. Introduction

Teams are ubiquitous in organizations. Since the end of the 20th century the focus of work in organizations has shifted from the individual employee to employees as part of a team (Kozlowski & Ilgen, 2006). Alongside this shift we have seen an increase in research focusing on the collaborative processes and outcomes of different types of teams (Vangrieken, Boon, Dochy, & Kyndt, 2017). One specific form of teams is an action team, "…where members with specialized skills must improvise and coordinate their actions in intense, unpredictable situations" (Edmonson, 2003, p. 1421). In other words, it is the task of action teams to quickly establish effective coordination and communication in unexpected situations. As negative outcomes of these team processes can be detrimental for human safety (e.g., in medical teams or aviation teams), it is of utmost importance that these teams are trained well in performing these complex team interaction skills in a realistic environment. Simulation rooms provide excellent, risk-free opportunities to practice both technical and team interaction skills in realistic scenarios that allow team members to experience how they will perform during stressful moments (Kneebone, Nestel, Vincent, & Darzi, 2007).

Not only research but also practice can benefit from exploring team interaction processes during such stressful moments during scenario-based training (Entin & Serfaty, 1999; Lei, Waller, Hagen, & Kaplan, 2016). Traditionally, debriefing sessions with expert debriefers are used to provide feedback in simulation-based learning (Fanning & Gaba, 2007). During these sessions, expert debriefers select specific, observable events in the scenario and stimulate trainees to reflect on their behavior and decisions. The use of expert debriefers to provide feedback is not only very costly and time- and labor-intensive, but research has also shown that how debriefers facilitate the debriefing sessions is highly variable (Tannenbaum & Cerasoli, 2013). In addition, the events they select are not necessarily the moments team members experience as stressful. We argue that a next step is needed to move from traditional human-based observation methods to methods that allow for more objective and timely identification of effective team interaction processes during stressful moments.

Wearable sensors have opened up a new world of research possibilities to detect body signals and analyze speech from team interactions, providing insights into how people respond and interact without interfering with their natural work processes (Fischer & Järvelä, 2014). For example, sociometric badges (e.g., Olguin et al., 2009; Pentland, 2012) are sensors that are worn around the neck, similar to typical ID-badges, and are able to detect various features of social interaction. The Empatica E4-wristband [Empatica Ins, Cambridge, USA] (Garbarino, Lai, Tognetti, Picard, & Bender, 2014) is able to measure skin conductance (also known as electrodermal activity: Boucsein, 2012) in an unobtrusive way. This can be used as an indicator for identifying moments of high arousal (Boucheix, 2017; Christopoulos, Uy, & Yap, 2016), which typically reflect high levels of distress in the context of medical action teams during a crisis situation (Hunziker, Johansson, et al., 2011). When combined, these sensors enable detailed exploration of social interaction in teams during moments of high arousal in action teams.

However, a lot is still unknown about how to use and combine these sensors with more traditional measures to detect effective team interaction processes during moments of high arousal, especially for action teams. Therefore, in line with the purpose of this special issue, the goal of this study is to clarify the methodological approach, added value and pitfalls of using and combining different sensor technologies in combination with video observation. Our study was performed in a medical simulation room for advanced life support (ALS) training. ALS is a complex emergency situation following cardiac arrest of a patient and is characterized by "extreme time pressures, diagnostic uncertainty, and rapidly evolving situations" (Doumouras, Keshet, Nathens, Ahmed, & Hicks, 2012, p. 274; Hunziker, Laschinger, et al., 2011). Research has shown that human factors such as efficient team communication, coordination, and stress especially affect ALS efficiency and performance (Fernandez Castelao, Russo, Riethmüller, & Boos, 2013; Hunziker, Laschinger, et al., 2011). Some have estimated that poor non-technical skills can contribute to 64 to 83% of critical incidents in a medical context or crisis situation, for example, in anesthesia (Arnstein, 1997).

In this study, we combine traditional video observation methodology to systematically analyze the content of team interaction behaviors with innovative sensor technology to explore the structure of team interaction processes in more detail (sociometric badges) and identify moments of high arousal (Empatica E4). The outcomes of this study demonstrate how using a combination of different sensors and traditional measures provides a means to get a rich picture of complex team interactions during moments of high arousal. These

insights advance our knowledge of how to use and combine sensor technology and how this information can be used for optimization of simulation environments to help prospective and current medical professionals to improve not only their medical skills, but also their team interaction skills.

2. Theoretical framework

2.1 Simulation-based medical education

Simulation in medical education is used as an educational technique to improve health outcomes, dating back to the 17th century (Cooke, Irby, & O'Brien, 2010; McGaghie, Issenberg, Petrusa, & Scalese, 2009). The number of medical simulation settings has expanded with the development of complex technologies. Such advanced technologies enable simulations that closely resemble reality, especially when combining them with high-fidelity scenarios built around events that potentially could have serious consequences for the patient (Dias & Neto, 2016; Grenvik, Schafer, DeVita, & Rogers, 2004). Skills acquired in a well-designed medical simulation environment show better transfer to improved real-life patient care compared to traditional, on-the-job medical training (McGaghie, Issenberg, Cohen, Barsuk, & Wayne, 2011).

Simulation is especially useful in training under conditions of uncertainty, ambiguity and rapid situation changes with potentially severe consequences for patient safety (Satish & Streufert, 2002). High-fidelity simulation-based learning is nowadays the standard for training ALS teams who must diagnose and manage a patient in cardiac arrest (Sahu & Lata, 2010). Diagnosing and managing a patient in cardiac arrest requires immediate medical intervention and efficient teamwork; otherwise, a patient might not survive (Hunziker, Johansson, et al., 2011). Successful resuscitation depends on the integrated application of technical skills, such as intubation, chest compressions, and clinical reasoning, and non-technical skills related to working in a team, such as communication, decision-making, and leadership (Hunziker, Tschan, Semmer, Howell, & Marsch, 2010). Effective and efficient teamwork in a resuscitation scenario requires a sequence of actions that is performed in the correct way and at the right time (Hunziker, Johansson, et al., 2011).

2.2 Team interaction

A fundamental aspect that can lead to high team performance in a crisis situation, such as a cardiac arrest, are the emergent, interactive processes between medical personnel (e.g., Hunziker, Johansson, et al., 2011). Team interaction is defined as a series of ongoing behavioral processes and actions that occur over time (Lei et al., 2016; Stachowski, Kaplan, & Waller, 2009). For decades, team researchers have been advocating to capture the dynamic nature of such team interactions, as opposed to using static measures (Marks, Zaccaro, & Mathieu, 2000). There are several ways to quantify verbal team interactions. Some researchers focus on the *content* of the interaction, for example, by using observation schema that identify what team members say, such as the number of agreements, suggestions, or opinions in team conversations (e.g., Atwal & Caldwell, 2005; Hoogeboom & Wilderom, 2015). Others focus on the *structure* of the conversation, such as which team member is speaking, the number of interruptions or the degree of turn-taking, regardless of the content (Kim, McFee, Olguin Olguin, Waber, & Pentland, 2012; Koudenburg, Postmes, & Gordijn, 2017; Pugliese, Nicholson, & Bezemer, 2015). For example, when providing teams real-time feedback on conversational balance (i.e. over-participators were stimulated to decrease their participation), this influenced their decision making for either better or worse (DiMicco, Hollenbach, Pandolfo, & Bender, 2007). In the current study, both the content and structure of team interactions will be taken into account.

Previous studies have already identified some content and structure related characteristics of effective team interaction in action teams which are dealing with crisis situations. For example, using video observation and coding, previous studies on the *structure* of interactions of airline teams have shown that effective teams displayed less complex, more homogenous interaction processes (Kanki, Folk, & Irwin, 1991; Zijlstra, Waller, & Phillips, 2012). Other related research on the structure of interactions in nuclear power plant control room

crews showed that effective interactions during crises consisted of fewer actors and less back-and-forth communication (Stachowski et al., 2009). Hence, shorter, less complex, and less reciprocal team interaction, indicating somewhat scripted or standardized forms of team interaction, seem to be more effective in action teams. Regarding the *content* of interactions, Kolbe et al. (2014) showed that effective medical team members more frequently spoke up and aided assistance after implicit action coordination (i.e. team monitoring). Such task-related helping interactions after team monitoring behavior seemed vital for high performance. For the team leader, communicating clear goals and a clear task distribution has proven to reduce the emotional reactions by team members, leading to an increase in performance in stressful situations (Zaccaro et al., 2001; Andersen, Jensen, Lippert, & Østergaard, 2010; Marsch et al., 2004). Moreover, research in emergency command-and-control teams has shown the importance of leader structuring behavior, such as clarifying and summarizing (van der Haar et al., 2017). During resuscitation, the use of closed-loop-communication is advocated to avoid errors (Fernandez Castelao et al., 2013). This clear, structured, and standardized form of communication (Brindley & Reynolds, 2011; Härgestam, Lindkvist, Brulin, Jacobsson, & Hultin, 2013) consists of an initial message (call-out) by the team leader, which should be confirmed or acknowledged by the receiver (check back) and confirmed once again by the team leader (closing the loop) (Davis et al., 2017; Härgestam et al., 2013; Jacobsson, Hargestam, Hultin, & Brulin, 2012; Schmutz, Hoffmann, Heimberg, & Manser, 2015). Hence, in general, various studies from different fields suggest that effective and less effective action teams differ in both the content and structure of their team interaction.

To enhance the learning opportunities in simulation environments, it is important to delineate the effective team interaction processes that are required for high team performance (Goldman, 2014). However, only a small - but growing - number of studies have investigated in detail how teams interact in the daily context of their work and how this contributes to achieving their goals (Humphrey & Aime, 2015). Nowadays, technical and methodological advances allow us to capture team interactions more precisely (Molenaar, 2014). One of the available, but to date less frequently used, devices is the sociometric badge (Kim et al., 2012): a wearable that includes several types of technology, namely Bluetooth, an infrared sensor, an accelerometer and a microphone. Either when used in isolation or when combined with other sensors, the badges allow for fine-grained analysis of the structure of verbal team interactions. For example, when combined with sensors that capture electrodermal activity, the structure of team interactions in high- arousal moments can be compared to moments of low arousal.

2.3   Moments of high arousal: definitions and effects

Identifying moments of high arousal during a medical simulation session can inform us about what happens during moments when a person is not able to process the mental effort or cognitive load required or when the perceived demands of the environment exceed a person's ability to cope with these demands (Berntson & Cacioppo, 2000; Stemmler, 2004; Boucheix, 2017; Lazarus & Folkman, 1984). For example, in the context of medical simulations, higher levels of arousal may ensue when time pressure forces an ALS team member to act quickly. It should be noted that moments of high arousal can also automatically occur due to positive events, such as positive workplace interaction or excitement (Heaphy & Dutton, 2008). Hence, high arousal is not only attributable to distress, but also to excitement (Russell, 1980). In other words, when physiological measures of arousal are used, information is obtained about the intensity of physiological arousal, but not about the psychological state or valence (e.g. distress or excitement) associated with it (e.g., Akinola, 2010). However, even though no general inferences about valence can be drawn from the intensity of arousal (e.g., Boucsein, 2012; Larsen, Diener, & Lucas, 2002), a previous study on self-reported emotions during resuscitation performance has shown that negative emotions (stress or overload) are significantly higher during resuscitation, while positive emotions were highest before resuscitation, decreased during resuscitation, and increased again when the simulated patient was awake again (Hunziker, Laschinger, et al., 2011). Therefore, in our study we interpret high levels of arousal as an indicator of feelings of distress.

Heart rate and skin conductance, physiological responses of the body, are examples of markers of autonomic activity of the nervous system, and concomitants of arousal (Akinola, 2010; Benedek & Kaernbach, 2010). Particularly during social interaction, skin conductance has been found to be the most sensitive indictor of emotional responsiveness or arousal, as opposed to the other physiological markers (Marci, Ham, Moran, & Orr, 2007). This physiological measure captures the intensity of emotions during interactions with others

(Akinola, 2010; Figner & Murphy, 2011). Skin conductance is defined as variations in the eccrine sweat glands (i.e., sweat glands which are present in all bodily parts, with the highest density in the palms and soles (Boucsein, 2012): in response to sweat secretion from the skin (e.g., Benedek & Kaernbach, 2010). The popularity of using skin conductance measures is due to its direct relation with a stimulus or emotional response, which enables us to capture moments of high arousal, as long as the temperature in the environment is kept constant (Boucsein, 2012; Lang, Bradley, & Cuthbert, 1998).

Previous studies have examined the effect of stress on ALS performance, but with conflicting results. On the one hand, Hunziker, Laschinger et al. (2011) and Hunziker, Semmer et al. (2012) found that perceived stress during early resuscitation negatively influenced ALS performance. In the latter study, also physiological measures as indicators for stress were used, but no association with team performance was found, possible due to the fact that the team members were engaged in physical activity what distorted the physiological measurements. In a similar vein, the study of Sandroni, Fenici, et al. (2005) did not find a direct relation between physiological stress measures and individual performance during the ALS scenario as measured by a written multiple-choice test. Other studies which examined the relation between physiological measures of stress and performance in other medical simulation settings found a U-shaped association (also known as the Yerkes-Dodson law, cf. Cohen, 2011): positive relationships between arousal and performance have been reported, whereas extreme levels of arousal were detrimental for performance (e.g., Keitel et al., 2011; Wetzel et al., 2010). To understand these conflicting findings, we suggest that exploring more in-depth how members *interact* during moments of high arousal is an interesting endeavor, as this might better explain performance than solely the level of arousal.

3   The present study

In this paper, we focus on how we can use the combination of sociometric data, physiological data, and video data to explore if more effective teams – compared to less effective teams - alter the content and structure of their team interactions during moments of high arousal. We expect that sensor technology will have added value in addition to self-report measures to objectively identify moments of high arousal and analyze the structure of the team interactions. As these measures are relatively new to the field of social science and specifically educational science, this exploratory study sets out to uncover what the added value, pitfalls and hurdles are when using and combining different sensor technologies with more traditional measures in order to better understand team interactions. To better understand the added value of sensor technology, we have translated the aim of this study into the following research question: *How do more versus less effective medical action teams differ in content and structure of team interactions during scenario-based training, and how do their interactions differ during moments of high arousal versus outside these moments of high arousal?* We will answer the research question by step by step testing the differences in content and structure of team interactions between

1) moments of high arousal and non-high arousal moments (without taking effectiveness into account);

2) more and less effective teams (without taking arousal into account); and

3) the combination of both: differences between high arousal versus non-high arousal moments separately for more and less effective teams.

4 Methodology

4.1 Participants and design

All 95 first-year master's students (comprising 24 teams) who enrolled in the course 'Advanced Life Support (ALS)' in the master study program 'Technical Medicine' at University of Twente were invited to participate in the study. Ninety-two students gave written informed consent to participate in the study. The three students who did not give consent and their team members were excluded from the study, resulting in a data set of 21 four-person teams and one three-person team. To avoid an unequal situation for this three-person team, one person from another team was added to this team during the assessment. To ensure comparability across teams, this three-person team was left out from the analysis ($N = 84$). On average, the students were 22.4 years old ($SD = 1.1$) and 44% were male. During the ALS course, students learned to diagnose and manage a patient with cardiac arrest and perform cardiopulmonary resuscitation (CPR) on an advanced human patient simulator in scenarios of varying complexity. A multimethod design was adopted which included four different sources of data: (1) Video coding of the content of team interactions, (2) sociometric measurement to capture the structure of team interactions, (3) the Empatica E4-wristband to capture skin conductance/arousal, and (4) teacher surveys to assess team effectiveness.

4.2 Procedure

Prior to data collection, the study was approved by the ethical committee of the university as well as by the teachers involved in the ALS course. During the first introductory lecture of the course, the students were informed about the study. The data were collected during the final assessment of the course. All teams were assessed on the same day. Two rooms were used for the ALS simulation scenarios, both had a regulated temperature with 0.1-degree Celsius temperature tolerance. The temperature for both rooms was kept constant at 20.5 degrees Celsius. Before the start of the scenario, all four students of each team were randomly assigned to one of the four fixed team roles: *team leader*, responsible for task distribution, monitoring team performance, creating an overview of the situation, and patient handover at the end of the scenario; *medication nurse*, responsible for drug administrations and connecting devices to the patient; and two *CPR administrators*, responsible for chest compressions and airway management. All students had practiced each role at least one time during the course, so they knew what was expected from them at the assessment.

In the ALS context the physical activities that are required by team members in the role of medication nurse and CPR administrator (such as intubation or chest compressions) influence their physiological arousal (i.e., producing distorted or biased physiological measurement: Berntson & Cacioppo, 2000; Stemmler, 2004). The team leader in the ALS situation is less physically active, so we decided to measure physiological arousal of the team leader to identify moments of high arousal. The researchers distributed a sociometric badge to all team members and secured the Empatica E4-wristband on the non-dominant wrist of the student in the role of team leader.

As Hunziker, Laschinger, et al. (2011) described in their paper on stress and team performance during a simulated resuscitation, simulated Advanced Life Support (ALS) scenarios usually follow a specific pattern. Similar to the Hunziker study, the ALS scenarios started with a short briefing about the patient's history by one of the teachers (maximum 90 seconds), immediately followed by a resuscitation period during which CPR is performed. The scenarios ended with a handover of the patient to another team or specialist. However, contrary to the Hunziker study, in the majority of the scenarios in our study the patient was still in a critical condition when the scenario was completed; the patient could breathe on its own, but did not have a steady pulse or sinus rhythm. Given that the patient still was in a critical condition, the patient handover at the end is a crucial phase in the scenario. The duration of the scenarios was on average 22.00 minutes ($SD = 4.83$). Previous studies have shown that team behaviors and interactions change as the scenario progresses, with a stronger focus on leadership and coordination skills at the beginning of a scenario (Tschan et al., 2006, 2014), and agreement on a shared diagnosis and treatment plan, as well as accurate handover at the end of the scenario. Especially in assessment settings the end is perceived as a stressful part of the scenario (Sandroni et al., 2005). Therefore, although we recorded the complete scenarios, for the analysis we focused on the beginning and end of the scenario, which we defined as the first 16.7% minutes (1st time window) and last 16.7% minutes (2nd

time window). The duration of these 16.7%-windows was on average 3 minutes and 40 seconds (*SD* = 49.6 s) and varied from 2 minutes and 24 seconds for the shortest video to 4 minutes and 52 seconds for the longest video.

4.3 Instruments

4.3.1 Sociometric badges

Sociometric badges were used to assess the structure of team interactions (Kim et al., 2012). Sociometric badges are wearables distributed by Humanyze (a spin-out of the MIT Media Lab) that measure proximity, body movement and speech features using respectively Bluetooth, infrared, an accelerometer and microphones. Data were uploaded from the badges using Sociometric Datalab Research edition (version 3.1.3029) and subsequently exported to excel files using the 'structured meeting' setting which disregards the Bluetooth and infrared data in order to provide a dataset in which each team member is assumed to be in each other's proximity during the entire session, which was the case in our setting. Also, the setting 'noisy environment' was used, which filters out additional environment noise, such as the beeping of the heart monitor or the speech from teachers not wearing a badge. The exported microphone data shows per second whether a participant was talking or silent.

From this data, the following four metrics were calculated at the team level: 1) the proportion of the time that one or more of the team members was speaking (i.e., *proportion of speaking time*), 2) the proportion of the speaking time when at least two team members were speaking at the same time (i.e., *proportion of overlapping speech*), and 3) the distribution of how each team member contributed to the overall speaking time (i.e., *conversational imbalance in speech*). *The conversational imbalance* was calculated as the standard deviation of the amount of time that each team member spoke (regardless of whether another member was speaking at the same time), corrected for the duration of the relevant time window. The higher this standard deviation, the greater the variation in speech among members, that is, the greater the conversational imbalance. In addition, also 4) the *proportion of speaking time of the team leader* was calculated, measured as the proportion of the time the team leader was speaking, regardless of whether another member was speaking at the same time. All measures were calculated using R (R Core Team, 2014).

4.3.2 Empatica E4-wristband

The Empatica E4 (hereafter referred to as E4), a relatively unobtrusive wristband, including 8 mm silver-plated electrodes, was used for the skin conductance recording. Continuous measures of physiological arousal at a sample rate of 4 Hz were collected using this device. The wristband was placed on the wrist of the team leader's non-dominant hand. Continuous decomposition analysis was used to get the relevant phasic electrodermal activity parameters[1]: Skin conductance responses (i.e., the number of peaks for certain periods of time which represents the fast-varying phasic component of skin conductance; the amplitude threshold for the extraction of skin conductance responses was .01 micro Siemens (µS)) and amplitude of the skin conductance responses (i.e., the height of a single skin conductance response). Using this analysis, the classical trough-to-peak parameters such as number of skin conductance responses and amplitude of the skin conductance responses can be obtained (Benedek & Kaernbach, 2010).

To detect the moments of high arousal, skin conductance amplitudes for each individual team leader were computed (Hamaker, 2012; Wiesenfeld, Whitman, & Malatesta, 1984). Information about peak amplitude of individual skin conductance responses can be used to detect moments of high arousal (Bach, Flandin, Friston, & Dolan, 2009). The highest skin conductance amplitude in the beginning and end of the scenarios was determined for each team leader. On the basis of the highest skin conductance amplitude a single segment of 30 seconds was selected both in the beginning and the ending of the scenario. This resulted in two 30-second segments for each team: one in the beginning (first 16.7%) and one in the end (last 16.7%) of the

---

[1] This analysis also yields an estimation of tonic activity: Skin conductance level (i.e., the slowly varying skin conductance activity: Benedek & Kaernbach, 2010).

scenario[2]. The 30-second segment was based on the thin-slices theory of social interaction (Curhan & Pentland, 2007). Because several studies have reported an average delay of one to four seconds between a stimulus and a skin conductance response (e.g., Dawson et al., 2007; Weis & Herbert, 2017) we used a segment starting 5 seconds before and 25 seconds after the detected highest amplitude. The highest amplitude peak (in micro Siemens, µS) of each team leader in both the beginning ($M = 2.34$, $SD = 2.67$) and end ($M = 3.63$, $SD = 3.26$) was compared with the mean amplitude of the beginning ($M = 1.65$, $SD = 2.01$) and end ($M = 2.76$, $SD = 2.65$). This indicated that both in the beginning ($t (15) = 3.19$, $p < .01$) and in the end ($t (15) = 4.67$, $p < .01$) the highest amplitude peak was significantly higher than the mean amplitude in that segment.

4.3.3   Video

The content of the team interactions was measured via video recordings. The video cameras were ceiling-mounted, fixed cameras that minimized obtrusiveness and reactivity of the team members. The content of the team interactions was coded for the first and last 16.7% minutes of each scenario. The codebook was specifically designed to capture behaviors that occur frequently in action teams and is rooted in earlier theorizing on teams and action teams (Lei et al., 2016; Stachowski et al., 2009; Zijlstra et al., 2012). Additionally, on basis of the theory described in the theoretical framework, two items were added in order to code closed-loop-communication, more specifically: (1) check-back (by a team member), and (2) closing the loop (by the team leader); see Table 1 for an overview of definitions and examples of the coded behaviors. A distinction was made between behaviors of the team leader, and of the other team members (followers). Exhaustive coding was applied, meaning that all behaviors were coded and a non-observable category was used for behaviors that were not understandable or not relevant. The "Observer XT" software program (Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000; Spiers, 2004) was used to code the videos. One of the coders parsed all sessions, that is: segmented the videos in speaker utterances (Klonek, Burba, Kauffeld, & Quera, 2016), using as a unit of analysis "a sentence or part of a compound sentence that can be regarded as meaningful in itself, regardless of the meaning of the coding categories" (Strijbos, Martens, Prins, & Jochems, 2006, p. 37). Subsequently, all segmented scenarios were systematically coded by two independent, trained coders, who were not informed about the moments of high arousal. Overall, an inter-rater agreement of 83.1% (Cohen's kappa = .80; Cohen, 1960) was established. After coding the videos, the behavioral codes "Not observable", "External communication", and the infrequent behaviors "Laugh" and "Apologies", were merged into one category "External / other behavior".

The frequencies of the behaviors of a team are highly influenced by the total duration of that video. Therefore, all coded behaviors were standardized according to the shortest video using the following formula: standardized frequency of a certain behavior of team X = coded frequency of the behavior of team X * (duration of the shortest video / duration of video team X). This resulted in a time-standardized behavior which enabled direct comparisons of the frequencies of the team members' behaviors across the different teams. In addition, to enable comparison of the frequencies between the 30-second segment and the rest of the time window, the percentages of every leader and follower behavior were calculated relatively to all of the leader or follower behavior in the relevant time interval.

---

[2] By following this procedure, we detected the highest moment of arousal of the team leader during the beginning and end of the scenario. This is a relative high arousal moment compared to the rest of the scenario, but could still be rather low in an absolute sense. Detecting a state of high arousal in an absolute sense is only possible if a baseline measurement is carried out, which was not possible in the setting of our study. So, for readability, we speak of moments of high arousal, but this should be read as "moments of relatively high arousal compared to the rest of the time window".

Table 1

Examples of coded video behaviors.

Behavioral code	Team member	Description	Examples
Team leader behavior			
Command	TL	The team leader gives an individual a specific assignment of responsibility (addressed call out).	*"Do you want to turn on the ECG?"* *"You can administer it directly."*
Summary	TL	The team leader produces a summary or discussion of the current situation, diagnoses and/or provides information to other team members on what to expect in the next stage. Any repetition of what was discussed with a bystander is also coded as *summary*.	*"We expect something like hyperaemia…"* *"We will evaluate the patient on visible symptoms."*
Confirmation	TL	The team leader answers a question by giving a confirmation.	*"Yes."*
Inquiry	TL	The team leader requests factual information, statement, or analysis from one or more individuals.	*"Is the patient breathing?"* *"Is the airway unobstructed?"*
Questioning	TL	The team leader requests confirmation or rejection of a statement from one or more individuals.	*"Shall we both have a look at the screen?"*
Suggestion	TL	The team leader suggests a future action without delegating it to a specific team member (call out not addressed).	*"Maybe we can ask for an ultrasound of the abdomen."* *"In 30 seconds, we need to do a heart rhythm check."*
Opinion	TL	The team leader makes a statement to express a personal view.	*"I think it is hyperaemia."* *"I agree."*
Observe	TL	The team leader recognizes or notices a fact or occurrence.	*"I can see a heartbeat."* *"I can see an asystole."*
Closing the loop	TL	The team leader closes the communication loop by confirming the follower's check back.	*"Great, thank you."* *"Ok."*
External communication*	TL	The team leader directs communication at someone outside the CPR-team. This may include a specialist, doctor, nurse, or relative of the patient. Communication to someone outside of the simulation (i.e., the teacher) is also coded as external communication.	*"Is a family member present?"* *"Did the patient have complaints before he was brought in?"*
Laugh*	TL	The team leader laughs or makes a clearly humorous remark.	*"Haha."*
Apologies*	TL	The team leader makes an apology.	*"Oh, sorry."*
Not observable*	TL	The team leader says something but the content is not understandable or not relevant. Code only when the verbal behavior is incomprehensible due to half sentences, simultaneous speaking, or background noise (e.g., beep-sound from the patient monitor), or not relevant to the research.	*""Robert, do you eh…"*
Follower behavior			
Check back	F	A follower responds to a call out by the TL (i.e., command, suggestion, question, or inquiry) in the form of a confirmation, answer or observation.	*"Yes, I will."* *"Shock administered."*
Other	F	The follower makes a statement that is not a check back.	

Note. TL = team leader; F = follower. * after coding, these behaviors were merged into the category external / other behavior.

4.3.4 *Teacher ratings of performance*

To assess team effectiveness, four team effectiveness items by Gibson, Cooper, and Conger (2009) were used: "This team is a consistently well performing team", "This team is effective", "This team makes few mistakes", and "This team delivers high quality work". The items were directly scored after the scenario on a Likert scale from 1 (strongly disagree) to 7 (strongly agree) by two teachers. Each teacher scored twelve teams. The internal consistency of the scale was high: Cronbach's alpha was .97. The teams were categorized as 'more' or 'less' effective on the basis of a median split, which was 5.75. The 10 less effective teams had a

mean score of 4.28 (*SD*= .98), as rated by the teachers (on a scale of 1 to 7); the 12 more effective teams, including those with the median score, scored on average 6.23 (*SD* = .52) on team effectiveness.

4.4 Analysis

4.4.1 Selection of the data and dealing with missing data

All recorded data were checked for missing values. Video recordings and skin conductance data of all 21 teams were successfully obtained. Mechanical issues with five sociometric badges resulted in the loss of the data of 20 participants among 13 different teams. This resulted in a total of 18 teams from which sociometric data of the team leader was intact, and nine teams from which complete data of the sociometric badges was available of all team members. The demographic data from these subsamples was comparable with the reported data from the 21 teams; no significant differences were found.

4.4.2 Synchronization of the data

To combine the data from the E4, sociometric badges and video observations, the data had to be synchronized. Coded video behaviors, sociometric data and skin conductance amplitude data were synchronized on the basis of a mutual timeline. The internal clocks in the E4, sociometric badges, and video recording devices were used for synchronization purposes employing custom Python code (represented by Unix time: number of seconds from 1-1-1970 in Coordinated Universal Time: UTC). The data sources could only be synchronized if the clock time of the E4 biosensor and the timestamp of the video recording device were exactly aligned. On the basis of the Python code, the clock times of the E4, video, and sociometric badges could not be matched for five teams. Possibly, the clock time of the E4 was not synchronized (e.g., with a computer or laptop that had an accurate clock time); therefore, differences in the clock times of the video and sociometric badge on the one hand and the E4 on the other hand might be present. However, it should be noted that it is difficult to pinpoint the exact cause of the synchronization issue. Due to these synchronization issues and the earlier mentioned malfunctioning of five sociometric badges, a total of 16 teams were available for further analysis of the video data combined with the skin conductance amplitude data; 13 teams from which the sociometric data of the team leader could be matched to the skin conductance measures and video data, and seven teams from which the sociometric data was available of all team members and could be linked to the skin conductance measures and video data.

4.4.3 Analysis of relations between variables

To answer the first research question of whether content and structure of team interactions were different for moments of high arousal versus non-high arousal, a series of repeated measures MANOVAs were conducted. For each time window (beginning and end), a separate repeated measures MANOVA was conducted for the dependent variables describing the content of team interaction (team leader and follower behavior) and the variables describing the structure (proportion speaking time, proportion overlapping speech, conversational imbalance). In addition, a dependent sample t-test was conducted to test for differences in proportion speaking time of the team leader. This last variable could not be grouped into one of the MANOVAs due to a different sample size. In all of these analyses the level of arousal (high versus non-high) was used as the within-subject variable. The second research question about the differences between more and less effective teams was also answered with a series of MANOVAs and independent sample t-tests with the same dependent variables, but this time the level of effectiveness (more versus less) was used as the between-subject variable. To answer the third research question, the data was split into two groups (more versus less effective teams) to explore differences between those two groups in team interaction during moments of high arousal and non-high arousal. Due to the low sample size, we had to refrain from conducting MANOVAs as this resulted in not enough residual degrees of freedom. Therefore, separate independent t-tests were conducted to test for differences in structure and content of team interaction.

5. Results

5.1 Comparison of structure and content of team interactions between high arousal moments and rest of the time window

The results of the comparison of structure and content of team interactions between high arousal moments and rest of the time window is displayed in Table 2. As can be seen in Table 2, team members were speaking for approximately half of the time, and during the end of the scenario almost two third of the time. The team leader was speaking for roughly a quarter of the time. The repeated measures MANOVAs and dependent t-test showed no differences for the structure of team interactions on team level, both for the first ($F(12, 1) = .54$, $p = .667$) and second time window ($F(12, 1) = .22$, $p = .878$). Also, no differences were found in proportion of speaking time of the leader ($t (12) = .41$, $p = .686$ first time window, $t (12) = 1.08$, $p = .303$ second time window).

Regarding the content of the team interaction, in the beginning of the scenario, team leader communication is characterized by many commands, while in the end of the scenario, when teams have to reach a diagnosis and have to handover the patient, there is more external communication. The repeated measures MANOVA showed a significant effect of arousal on the content of team interaction, $F(10, 6) = 17.72$, $p = .001$ (first time window) and $F(10, 6) = 4.30$, $p = .044$ (second time window). Further inspection of the univariate ANOVAs showed significant differences between high arousal moments and the rest of the time window in the first time window for the team leader behaviors *Questioning* ($F(1, 15) = 45.52$, $p < .001$) and *Opinion* ($F(1, 15) = 8.32$, $p = .011$). This demonstrated that, on average, team leaders showed relatively less questioning ($M = 0.0\%$) and opinion ($M = 0.0\%$) behavior during moments of high arousal than during the rest of the first time window ($M = 3.3\%$ and $M = 0.8\%$). For the second time window, the univariate ANOVAs showed significant differences for *Suggestion* and *Opinion*, with $F(1, 15) = 10.12$, $p = .006$ and $F(1, 15) = 6.96$, $p = .019$ respectively. This indicated that, on average, team leaders showed relatively less suggesting ($M = 2.8\%$) and opinion ($M = 0.0\%$) behavior during moments of high arousal compared to the rest of the second time window ($M = 8.1\%$ and $M = 1.9\%$).

Table 2

Comparison in structure and content of team interactions between moments of high arousal and outside these moments.

	1st time window: first 16.7%				2nd time window: last 16.7%			
	During moment of high arousal		Outside moment of high arousal		During moment of high arousal		Outside moment of high arousal	
			Content of team interactions[a]					
Leader behavior (n = 16)								
Command	2.2	(45.8%)	9.9	(30.5%)	0.7	(11.6%)	2.8	(9.4%)
Summary	0.4	(6.1%)	2.4	(7.1%)	0.2	(3.4%)	1.3	(5.3%)
Confirmation	0.4	(5.2%)	3.1	(9.6%)	0.8	(15.5%)	4.3	(15.5%)
Inquiry	0.1	(1.0%)	1.1	(3.1%)	0.2	(3.3%)	1.7	(6.1%)
Questioning	0.0	(0.0%) **	1.1	(3.3%) **	0.3	(5.3%)	1.0	(3.4%)
Suggestion	0.5	(12.0%)	4.2	(12.8%)	0.2	(2.8%) **	2.2	(8.1%) **
Opinion	0.0	(0.0%) **	0.3	(0.8%) **	0.0	(0.0%) *	0.4	(1.9%) *
Observe	0.3	(4.7%)	1.8	(5.3%)	0.7	(11.7%)	2.4	(8.3%)
Closing the loop	0.3	(4.4%)	2.3	(6.7%)	0.5	(8.5%)	1.6	(5.6%)
External / other behavior	0.8	(20.8%)	6.9	(20.8%)	1.6	(37.8%)	8.9	(36.3%)
Total	5.0	(100%)	33.1	(100%)	5.2	(100%)	26.6	(100%)
Follower behavior (n = 16)								
Check back	1.1	(21.3%)	7.5	(28.8%)	0.9	(18.3%)	3.6	(12.9%)
Other	3.6	(78.7%)	18.5	(71.2%)	4.6	(81.7%)	24.2	(87.1%)

	Total	4.7	(100%)	26.0	(100%)	5.5	(100%)	27.8	(100%)
					Structure of team interactions				
Team (n = 7)									
Proportion of speaking time (%)		46.0		54.7		64.5		67.7	
Proportion of overlapping speech (%)		24.5		25.4		29.9		27.7	
Imbalance in speech (*SD*)		0.10		0.09		0.11		0.09	
Leader (n = 13)									
Proportion of speaking time (%)		26.2		27.9		29.5		33.5	

[a]Standardized frequencies (see Method section).

$^*p < .05.$ $^{**}p < .01.$

5.2 Comparison in structure and content of team interactions between more and less effective teams

To determine whether the structure and content of interactions differed between more and less effective teams, MANOVAs (see Table 3) were conducted to compare differences for the beginning (first time window) and the end (second time window). For both time windows, the MANOVAs did not show significant effects of team effectiveness on the structure of team interactions ($F(5, 1) = 2.67$, $p = .220$ and $F(5, 1) = 8.94$, $p = .052$ respectively). However, as the analysis of the second time window was approaching significance, we further inspected the outcomes of the univariate ANOVAs, which showed significant differences in the *conversational imbalance* between the more and less effective teams. Both in the beginning ($F(1, 5) = 11.94$, $p = .018$) and in the end ($F(1, 5) = 11.17$, $p = .020$), the more effective teams showed greater imbalance (respectively $M = 0.12$ and $M = 0.13$) than the less effective teams ($M = 0.06$ for both time windows). In other words, in more effective teams one person was more dominant in terms of speaking time, while in less effective teams the team members contributed more equally; see Table 3. For content of team interactions two MANOVAs were conducted that showed no significant effect of team effectiveness on content of team interactions for both time windows ($F(10, 5) = .58$, $p = .785$ and $F(10, 5) = .52$, $p = .822$ for the first and second time window respectively). Separate univariate ANOVAs confirmed no significant effects.

Table 3. *Comparison in structure and content of team interactions between more and less effective teams.*

	1st time window: first 16.7%				2nd time window: last 16.7%			
	More effective teams		Less effective teams		More effective teams		Less effective teams	
Content of team interactions[a]								
Leader behavior[b]								
Command	12.7	(33.3%)	11.0	(31.1%)	4.0	(11.0%)	2.6	(8.3%)
Summary	2.9	(7.5%)	2.5	(6.6%)	1.5	(4.7%)	1.5	(5.3%)
Confirmation	3.3	(8.8%)	3.8	(10.4%)	5.6	(15.8%)	4.3	(15.6%)
Inquiry	1.2	(3.0%)	1.1	(2.8%)	2.1	(6.1%)	1.4	(5.1%)
Questioning	1.0	(2.7%)	1.2	(3.1%)	1.5	(4.2%)	0.7	(2.5%)
Suggestion	4.9	(12.7%)	4.5	(12.3%)	2.3	(6.3%)	2.6	(9.4%)
Opinion	0.3	(0.7%)	0.3	(0.7%)	0.3	(1.0%)	0.6	(2.6%)
Observe	2.4	(5.8%)	1.6	(4.5%)	3.4	(8.7%)	2.6	(9.3%)
Closing the loop	2.7	(6.8%)	2.2	(6.0%)	2.4	(6.9%)	1.6	(4.9%)
External / other behavior	6.9	(18.7%)	8.7	(22.5%)	10.9	(35.3%)	9.8	(36.9%)
Total	38.3	(100%)	36.9	(100%)	34.0	(100%)	27.7	(100%)
Follower behavior[b]								
Check back	9.5	(29.9%)	6.9	(24.8%)	5.4	(15.6%)	2.9	(9.5%)
Other	22.0	(70.1%)	22.2	(75.2%)	28.6	(84.4%)	29.4	(90.5%)
Total	31.5	(100%)	31.1	(100%)	34.0	(100%)	32.3	(100%)
Structure of team interactions								
Team[c]								
Proportion of speaking time (%)	56.4		50.5		63.3		69.9	
Proportion of overlapping speech (%)	22.8		26.4		23.6		31.7	
Imbalance in speech (*SD*)	0.12	*	0.06	*	0.13	*	0.06	*
Leader[d]								
Proportion of speaking time (%)	30.4		22.0		34.4		29.2	

[a]Standardized frequencies (see Method section). [b]$n = 10$ for more effective teams and $n = 6$ for less effective teams. [c]$n = 3$ for more effective teams and $n = 4$ for less effective teams. [d]$n = 9$ for more effective teams and $n = 4$ for less effective teams.
*$p < .05$. **$p < .01$.

5.3 Comparison between moments of high arousal and rest of the time window separately for more and less effective teams

In addition to the overall differences between more and less effective teams, we were also interested to see whether more and less effective teams were different in how they changed their content and structure of the team interaction between moments of high arousal and outside these moments. Therefore, separately for the less and more effective teams, we tested what the differences were between the 30-second segment and the rest of the corresponding time window using dependent t-tests.

On average, more effective teams (see Table 4a) were less imbalanced during moments of high arousal ($M = .09$) than during the rest of the second time window ($M = .14$) ($t (2) = 5.6$, $p = .030$). Regarding the content of team interaction, in the more effective teams, team leader behavior was characterized by more *Commands*, both during moments of high arousal (54%), and in the rest of the time window ($M = 30.7\%$).

During the moment of high arousal, team leaders showed relatively less *Confirmation* (*M* = 1.7%), less *Questioning* (*M* = 0.0%), less *Opinion* (*M* = 0.0%) and less *Closing the loop* (*M* = 3.1%) than during the rest of the first time window (respectively *M* = 9.5%, *M* = 3.0% and *M* = 0.8% and *M* = 7.2%; *t* (9) = -4.77, *p* = .001, *t* (9) = -5.17, *p* = .001 and *t* (9) = -2.33, *p* = .045 and *t*(9) = -4.08, *p* = .020). In the end of the scenario, the moment of high arousal was characterized by relatively less *Summary* behavior (*M* = 1.4%) when compared to the rest of the second time window (*M* = 5.3%), *t* (9) = -3.36, *p* = .008.

Also, for *less effective teams* (see Table 4b) some differences were found regarding the structure and content of team interaction during the high arousal moments, compared to rest of the corresponding time window. Contrary to the more effective teams, we found that less effective teams, on average, were more imbalanced during moments of high arousal (*M* = .11) than during the rest of the window (*M* = .06), *t* (3) = 4.1, *p* = .026); yet, this difference was only significant for the first time window. Moreover, in the less effective teams, team leader behavior was characterized by a high percentage of *Commands* (*M* = 32.0% and *M* = 30.2%). In addition, in the beginning of the scenario, team leaders of less effective teams showed relatively less *Questioning* (M = 0.0%) and *Inquiry* (*M* = 0.0%) during the moments of high arousal compared to the rest of the first time window (respectively *M* = 3.7% and *M* = 3.3%; *t* (5) = -4.19, *p* = .009 and *t* (5) = -2.92, *p* = .033). In the end of the scenario, during the moment of high arousal also relatively less *Inquiry* behavior (M = 0.0%) and less *Suggesting* behavior (*M* = 3.3%) was exhibited compared to the rest of the second time window (respectively *M* = 5.9% and *M* = 10.4%; t (5) = -3.40, *p* = .019, and *t* (5) = -3.21, *p* = .024). In addition, a difference in follower behavior was found; in less effective teams; the followers showed relatively less check back behavior during moments of high arousal than during the rest of the time window (*M* = 15.6% vs. *M* = 25.4%, *t* (5) = 2.59, *p* = .049).

Table 4A. *Comparison in structure and content of team interactions between moments of high arousal and outside these moments within more effective teams.*

More effective teams	1st time window: first 16.7%				2nd time window: last 16.7%			
	During moment of high arousal		Outside moment of high arousal		During moment of high arousal		Outside moment of high arousal	
Content of team interactions[a]								
Leader behavior ($n = 10$)								
Command	2.3	(54.0%)	10.4	(30.7%)	0.6	(8.9%)	3.4	(11.2%)
Summary	0.2	(3.3%)	2.7	(7.8%)	0.1	(1.4%) **	1.4	(5.3%) **
Confirmation	0.1	(1.7%) **	3.2	(9.5%) **	1.0	(16.2%)	4.6	(15.2%)
Inquiry	0.1	(1.6%)	1.1	(3.0%)	0.3	(5.3%)	1.8	(6.2%)
Questioning	0.0	(0.0%) **	1.0	(3.0%) **	0.4	(8.5%)	1.1	(3.7%)
Suggestion	0.4	(10.0%)	4.5	(13.1%)	0.2	(2.5%)	2.1	(6.8%)
Opinion	0.0	(0.0%) *	0.3	(0.8%) *	0.0	(0.0%)	0.3	(1.2%)
Observe	0.4	(6.2%)	2.0	(5.5%)	0.7	(10.8%)	2.7	(8.2%)
Closing the loop	0.2	(3.1%) *	2.5	(7.2%) *	0.7	(11.7%)	1.7	(5.9%)
External / other behavior	0.5	(20.0%)	6.4	(19.3%)	1.7	(34.8%)	9.2	(36.3%)
Total	4.2	(100%)	34.1	(100%)	5.7	(100%)	28.3	(100%)
Follower behavior ($n = 10$)								
Check back	1.2	(28.6%)	8.3	(30.5%)	1.2	(20.3%)	4.2	(15.1%)
Other	3.0	(71.4%)	19.0	(69.5%)	4.7	(79.7%)	23.9	(84.9%)
Total	4.2	(100%)	27.3	(100%)	5.9	(100%)	28.1	(100%)
Structure of team interactions								
Team ($n = 3$)								
Proportion of speaking time (%)	31.4		61.0		55.8		64.8	
Proportion of overlapping speech (%)	18.4		23.8		22.2		23.9	
Imbalance in speech (*SD*)	0.09		0.13		0.09 *		0.14 *	
Leader ($n = 9$)								
Proportion of speaking time (%)	25.8		31.0		28.2		35.6	

[a]Standardised frequencies (see Method section).
*$p < .05$. **$p < .01$.

Table 4B. *Comparison in structure and content of team interactions between moments of high arousal and outside these moments within less effective teams.*

Less effective teams	1st time window: first 16.7%				2nd time window: last 16.7%			
	During moment of high arousal		Outside moment of high arousal		During moment of high arousal		Outside moment of high arousal	
Content of team interactions[a]								
Leader behavior (*n* = 6)								
Command	2.0	(32.0%)	9.0	(30.2%)	0.8	(16.1%)	1.7	(6.5%)
Summary	0.7	(10.8%)	1.9	(5.8%)	0.3	(6.7%)	1.1	(5.1%)
Confirmation	0.8	(11.1%)	2.9	(9.8%)	0.5	(14.4%)	3.8	(16.0%)
Inquiry	0.0	(0.0%) *	1.1	(3.3%) *	0.0	(0.0%) *	1.4	(5.9%) *
Questioning	0.0	(0.0%) **	1.2	(3.7%) **	0.0	(0.0%)	0.7	(2.9%)
Suggestion	0.7	(15.3%)	3.8	(12.3%)	0.2	(3.3%) *	2.4	(10.4%) *
Opinion	0.0	(0.0%)	0.3	(0.8%)	0.0	(0.0%)	0.6	(3.2%)
Observe	0.2	(2.1%)	1.5	(4.9%)	0.7	(13.3%)	1.9	(8.5%)
Closing the loop	0.3	(6.7%)	1.9	(6.0%)	0.2	(3.3%)	1.4	(5.2%)
External / other behavior	1.2	(22.1%)	7.6	(23.2%)	1.5	(42.8%)	8.3	(36.3%)
Total	5.9	(100%)	31.2	(100%)	4.2	(100%)	23.3	(100%)
Follower behavior (*n* = 6)								
Check back	0.8	(15.6%) *	6.0	(25.4%) *	0.3	(6.9%)	2.5	(9.2%)
Other	4.5	(84.4%) *	17.7	(74.6%) *	4.5	(93.1%)	24.9	(90.8%)
Total	5.3	(100%)	23.8	(100%)	4.8	(100%)	27.4	(100%)
Structure of team interactions								
Team (*n* = 4)								
Proportion of speaking time (%)	57.0		50.0		71.0		69.8	
Proportion of overlapping speech (%)	29.0		26.5		35.7		30.5	
Imbalance in speech (*SD*)	0.11 *		0.06 *		0.13		0.06	
Leader (*n* = 4)								
Proportion of speaking time (%)	27.3		21.1		32.3		28.9	

[a] Standardized frequencies (see Method section).
*$p < .05$. **$p < .01$.

6. Discussion

In our study, we combined video data and ratings of team effectiveness with skin conductance measures (Empatica E4-wristband), and measures of speech features (sociometric badges) to analyze the structure and content of team interactions in medical action teams. The context of our study, a medical simulation room, was highly relevant for the aim of our study as this enabled us to closely observe medical action teams during a simulated crisis situation (ALS for a patient with cardiac arrest). By studying the differences between more and less effective teams in the content and structure of their interactions, both during moments of high arousal and outside these moments, we were not only able to add to the understanding of effective team interactions during crisis situations, but also of the added value of combining these various sensors with traditional data. In the following paragraphs, we discuss (1) the insights from the study on structure and content of team interactions of action teams during stressful moments, (2) the added value and

challenges of combining information from various sensors to contribute to a more fine-grained understanding of complex team interactions, (3) the limitations of our study and (4) future directions for research using sensor technology to study team interactions.

6.1 Discussion of findings

Understanding team interactions in ALS teams is crucial because ALS teams "need to be organized in such a way that the individual skills of the team members can be used efficiently and effectively" (Cooper & Wakelam, 1999; p. 27). Each team member needs a clear understanding of "how decisions are made within the group; what resources are needed and how they are to be utilized; how leadership is exercised; and how staff new to the situation are integrated into the group." (Cooper & Wakelam, 1999; p. 27). We analyzed the structure and content of team interactions of medical action teams during a simulated crisis situation and compared team interaction at moments of high arousal with team interaction outside these high arousal moments; as well as differences in team interaction between more and less effective teams.

First, when we compared team interaction during moments of high arousal with team interaction outside these moments without taking effectiveness of teams into account, we found differences in the *content* of the team interaction only. In this specific context, it turned out that the team leader gave no opinions during moments of high arousal, both at the beginning and end of the scenario. Moreover, the team leader asked no questions (in the beginning of the scenario) and made less suggestions (in the end). This is in line with previous research stating that during moments of crises fast coordination and clear decision making is needed (Tschan et al., 2006). In other words, there is no room for behavior that needs further interpretation or clarification from team members. However, no significant differences were found that indicated which behavior was more frequently present during the moments of high arousal.

Second, we inspected the differences between more and less effective teams without distinguishing between the moments of high arousal and the rest of the corresponding time windows. Overall, no main effects of effectiveness were found on the content and structure of team interactions. Of course, we have to consider that we had a very small sample size and thus too low power to detect small differences. The only difference that we found was in the conversational imbalance; more effective teams showed greater imbalance than the less effective teams, both at the beginning and the end of the scenario. In other words, in more effective teams, one person was more dominant in terms of speaking time, while in less effective teams the speaking time was more equally distributed. Although equal member contribution is generally seen as positive for high team effectiveness - as all opinions can be taken into account (DiMicco et al., 2007) - this is different for medical action teams during cardiac arrest where swift decisions and clear and effective coordination of the team leader are necessary (Andersen et al., 2010; Marsch et al., 2004). Our study points into the same direction, namely that a greater contribution of one person (greater imbalance) contributes to team performance. However, in extreme contexts effective leaders are receptive to the input of team members (Hannah, Uhl-Bien, Avolio, & Cavarretta, 2009) and team members of effective medical teams more frequently speak up and provide help (Kolbe et al., 2014). Therefore, dominancy of the team leader should not be interpreted as no room for other team members to contribute.

Finally, we analyzed separately for more and less effective teams if there were differences in structure and content of team interaction when comparing moments of high arousal with the remaining time in the begin and end of the scenario. This revealed an interesting regarding the *structure* of the team interaction: when we did not split up between moments of high arousal and the rest of the corresponding time window, we saw more conversational imbalance for more effective teams. However, after splitting up between moments of high arousal and the remaining time of the window, it became clear that both more and less effective teams show differences in conversational imbalance during moments of high arousal, but in opposite direction. The four less effective teams showed more imbalance, i.e. greater dominancy of one person during moments of high arousal, and were more balanced outside these moments (in the beginning of the scenario). On the contrary, in the three more effective teams team members contributed more equally during the high arousal moment (in the end of the scenario), and they were more imbalanced in their contributions during the rest of the time window. An interesting hypothesis for future research is to test whether a certain degree of conversational *imbalance* indeed contributes to higher team performance in action teams, while more equal contributions are needed when the team leader gets into a state of high arousal.

To better understand these outcomes, it is worth to have a look at the differences in the *content* of interaction. First, Table 4a showed that during moments of high arousal more than half of the team leader communication of the more effective teams consisted of commands. In addition, significant differences were found in other behaviors that occurred less frequently during the moments of high arousal: *Conformation*, *Questioning*, *Opinion*, and *Closing the loop*. For less effective teams (Table 4b), it seems that team leader behavior during moments of high arousal was more similar to their behavior outside these moments. Differences in team leader behavior in less effective teams were only found in the behaviors *Inquiry* and *Questioning*. In addition, the less effective teams showed differences in *follower* behavior, namely less check back behavior, which is a crucial step in closed-loop-communication to confirm an initial message from the team leader and to avoid mistakes and misunderstandings (Davis et al., 2017; Härgestam et al., 2013; Jacobsson et al., 2012; Schmutz et al., 2015). It could thus be the case that the finding that the team leader is more dominant during stressful moments in less effective teams, is related to less check back behavior from their followers: the team leader has to repeat or rephrase often because the followers do not confirm (check back) his or her commands which leads to a relatively greater contribution of the team leader. Again, we have to take the low sample size into account here while interpreting the findings, but what our results in general show is that a more fine-grained analysis of high arousal moments of a scenario enhances our understanding of what is effective team interaction during stressful moments. In addition, simultaneously exploring the structure and the content of the interactions can provide a more in-depth understanding of what constitutes effective team behavior.

6.2 Added value and challenges of using sensor technology to gain insights in team interactions

6.2.1 Added value and challenges of using the Empatica E4-wristband

In our study, we used the E4 to detect of moments of high physiological arousal that would not be observable with other methods or measures (Boucheix, 2017). Unveiling the highest level of physiological activation is shown here to lead to advanced insights, as the structure and content of team interaction that was displayed during these moments somewhat differed from what teams displayed outside these high arousal segments. Verbal reports or perceptual recall from participants about when they experienced higher stress or arousal are often distorted and do not accurately reflect the actual stressful moment (cf. Hunziker, Laschinger, et al., 2011). It should be noted that our results have to be interpreted with caution. First, when collecting data in the field (and not in a controlled laboratory environment) increases or fluctuations in skin conductance might be caused not only by higher mental effort, but also by general arousal or body movements (Berntson & Cacioppo, 2000; Cacioppo & Tassinary, 1990; Stemmler, 2004). This means that the assumed link between physiological arousal and psychological, behavioral or interaction processes requires careful interpretation (Akinola, 2010). As mentioned in the theoretical framework, the E4 captures the level of physiological states of arousal, but does not distinguish between excitement and distress. Although other studies using self-reported measures of stress (Hunziker, Laschinger, et al., 2011) indicate that during ALS especially the negative emotions are high and positive emotions are low (which is why we interpreted in our study high arousal as distress), additional validation measures would be preferable. Also, in an ALS context only the team leader's skin conductance can be validly measured, as other members would produce distorted or biased physiological measurement because of their activity concerning the compressions and administration of drugs. In order to understand how this affects the team as a whole, more research is needed to study the physiological concordance: to what extent are the physiological processes aligned and how does that influence the dynamics within the team (Marci et al., 2007)? Finally, even when we interpret arousal as stress, the trigger for higher levels of arousal is still unknown. In the beginning of the scenarios high arousal might be caused by the need to respond quickly to an overload of information, the diagnostic uncertainty and rapidly evolving situations (Doumouras et al., 2012; Hunziker, Johansson, et al., 2011), while at the end of the scenario the team leaders might experience higher arousal because they have to handover the patient and decide on a final diagnosis. Both triggers could give rise to the same amount of arousal, but might result in different kinds of team leader behavior.

6.2.2 Added value and challenges of using sociometric badges

The sociometric badges provided further insight into team interaction processes during these high-arousal moments that adds to the behavioral aspects shown by the video coding. Our results showcase how

the differences found between the behavior of the more and less effective teams are enriched by the information from the sociometric badges: a difference in their conversational balance became visible. This highlights the importance of also exploring the structure of the verbal interactions, besides looking at the content of the interactions. The sociometric badges have been proven to be reliable and accurate enough to study high-level team interactions, such as participation in conversations and total speaking time (Chen, & Miller, 2017). However, in our study, we experienced that the hardware in the sociometric badge or E4 might not always function well, resulting in missing data. Using equipment from a specific manufacturer also means that you are dependent on a commercial party whenever the equipment is malfunctioning. Despite extensive testing, it appeared that the hardware was not functioning properly and technical support was lacking as the researchers support platform for the sociometric badges was discontinued. Especially when collecting data from teams with the sociometric badges, the malfunctioning of one badge obstructs the computation of all team interaction dynamics (e.g., one participant could have a great influence on conversational balance). In this study, the failure of two badges resulted in the loss of data for almost half of the teams: all badges were used continuously during the day, meaning that each badge was used approximately five times that day. As downloading the data takes a substantial amount of time, it was not possible to do this during the experiment. As a consequence, it was discovered only afterwards that two badges malfunctioned, resulting in the loss of data for ten participants. In addition, in this specific study the interpretation of the sociometric measures is still on a superficial level: we do have information about the proportion of overlapping speaking time, but we do not know the specifics. For example, we might know that participants spoke more at the same time (overlap), but not whether this consisted of a lot of brief interruptions (e.g., a confirmatory 'yes') or fewer long interruptions (which might be experienced as more disrupting than brief confirmations). More advanced algorithms could result in additional measures that provide more insight into the effects we found, for example, in relation to the length and number of interruptions.

6.2.3 Added value and challenges of combining sensor technology with traditional measures

Although this triangulation of data sources to capture team dynamics in simulation environments results in rich data, there are several challenges that should be noted when adopting such a research design. First, a great benefit of combining video, sociometric and physiological data is that it offers continuous measures of behavior, interaction and physiological intensity on a temporal scale. Using this sensory triangulation enables the use of physiological arousal as a process-tracing method (Figner & Murphy, 2011). This means that it can provide information about behavioral processes, such as decision making, because such physiological data can be measured and collected continuously. At the same time, we experienced the difficulty of synchronizing the data. By using multiple sensors in combination with traditional measures, the risk that one device is malfunctioning or does not align with one of the other devices is substantial, resulting in a much smaller number of teams due to missing data. For future research, we recommend to always have a back-up procedure for this. For both the E4 and the sociometric badges it is possible to use behavioral markers: you push a button on the device and a time-stamp is saved to the data. When you do this in front of the camera you can always synchronize the video with the sensor device. For our study, it was not possible to include this additional step to the procedure, as we were collecting data in an assessment situation in which we already asked a lot from the participants. Although our study shows the potential of combining sensor technologies and the added value for team learning research, further research is necessary to validate and ground these methods. Each of the team interaction measures that are used in this study, whether observational, self-reported or technology-based, has its limitations. Conflicting information about team interaction from these measures needs to be explained and sources of discrepancies need to be understood. This would require not only validation studies, but also transfer studies from the simulation environment to the field.

6.3 Limitations and future research

Next to the problems and limitations that were related to the technology that was adopted, our study design had some limitations that we also have to acknowledge. To begin with, the small sample size ($N = 22$ teams) and even smaller sample size for the sociometric data resulted in limitations regarding the statistical power of the analyses and generalizability of our findings. Moreover, due to insufficient residual degrees of freedom, we were not able to perform MANOVAs for the third research question, resulting in a series of

independent t-tests and thus an increased risk of type I errors. Consequently, this research has a more exploratory character, which is why we interpreted the results with caution. In the future, we recommend to conduct similar studies with a bigger sample size. In addition to the low sample size, the observed frequency of the video coded behaviors was sometimes very low. This was due to the fact that we decided to zoom in on the beginning and end of the scenario instead of engaging in the time-consuming process of coding the whole scenario. Within these time windows of three to four minutes, some of the behaviors were hardly present and when we further zoomed in on the 30-second segments of high arousal, behaviors became even more infrequent. One could also question how many different behaviors one can display in only a 30-second segment. We therefore recommend for future studies to study multiple 30-second segments and longer time windows to allow for more fair comparisons.

Furthermore, it is known that stress is a complex phenomenon which is difficult to measure (Boucsein, 2012). In the present study we chose to include a physiological measure. As described above, these outcomes should be carefully interpreted, as eustress and distress produce similar physiological results. In addition, another limitation of our study is that in the context in which students were assessed it was not possible to obtain a baseline measure. Therefore, we could not compare teams on the team leader's level of arousal; only within person comparisons could be made where peaks in individual skin conductance data were identified in order to pinpoint moments of relative high arousal of a team leader. Therefore, future research is advised to measure skin conductance for longer periods and to obtain a baseline measurement to improve the quality of results. In addition, depending on the context of the study, it might be worthwhile for future research to explore the option of measuring the physiological data on the glabrous palmar or plantar surfaces, as this is more reliable and valid (but also more invasive) (Boucsein, 2012). Studies are available that provide insight into the differences between wristbands compared to palmar measures of skin conductance (e.g., van Lier et al., 2017).

Despite all experienced hurdles and limitations, our study strengthened our idea that wearable sensor technology has the potential to advance insights in team research. Sensor technology has the potential to provide objective and unobtrusive measures of complex behavioral and physiological processes as teams do not have to be interrupted while performing their task, which would disrupt their processes. In our study, the participants indicated that wearing the sociometric badges and E4 did not distract them from performing their tasks. Wearable sensor measures of interactions are especially promising when linked to psychological or team level constructs such as leadership emergence (Chaffin, Heidl, Hollenbeck, Howe, Voorhees, & Calatone, 2017) and in studying continuous streams of longitudinal data (Mathieu, Hollenbeck, Van Knippenberg, & Ilgen, 2017). Physiological measures of arousal can help to more objectively select stressful moments, which is highly relevant when studying action teams during crisis situations. However, applying these new methods is less straightforward and brings more challenges than is often suggested by manufacturers. First, in order to optimally use these methods, it is important to familiarize with this type of big data, which also adds computational complexity that requires specific expertise in data cleaning and dealing with noise (van Keulen, Kaminski, Matheia, & Katoen, 2018). Second, in order to apply sensors in a specific situation, many pre-studies are needed to test the reliability of the measures (cf. Chaffin, Heidl, Hollenbeck, Howe, Voorhees, & Calatone, 2017; de Laat, Endedijk, Ufkes, van Keulen, & de Vries, 2017). Ultimately, if one manages to extract meaningful data from the sensors, a final question is how to integrate these data with more traditional measures. As Fielding (2012) concludes in his analysis of how methods – including technological data - can be mixed, integrating data sources is an innovation in itself and should be treated as such. In other words, although time saving is often advocated as one of the benefits of using sensor technology, we still have a long road to go before this will become reality.

7   Conclusion

Effective team interaction is vital in medical situations and in medical learning and education. A combination of video-observational, sociometric and physiological data can enhance our understanding of the complex behavioral and interaction processes underlying effective team performance and provides alternative learning methods that can be used in the design of trainings and education of medical professionals. Technological advances together with the availability of more knowledge about the simultaneous application

of such methods are needed to use the full potential of wearable sensor technology in team research and overcome current teething troubles. Outcomes of this and future studies might enable future medical professionals to better understand what is required at stressful moments. Using these results in the training and during debriefing sessions can potentially optimize team interactions of future medical professionals and enhance the quality of medical care.

Keypoints

- More effective teams show greater conversational imbalance than less effective teams, but not during moments of high arousal.

- Both team leaders and followers show some changes in the content of the team interaction during moments of high arousal.

- More knowledge about the simultaneous application of wearable sensor technologies is needed to use the full potential of these methods in team research.

- Information from sensor technology can in the future be used during debriefing sessions to improve medical training and simulations

References

Akinola, M. (2010). Measuring the pulse of an organization: Integrating physiological measures into the organizational scholar's toolbox. *Research in Organizational Behavior, 30*, 203-223. doi:10.1016/j.riob.2010.09.003

Andersen, P. O., Jensen, M. K., Lippert, A., & Østergaard, D. (2010). Identifying non-technical skills and barriers for improvement of teamwork in cardiac arrest teams. *Resuscitation, 81*(6), 695-702. doi:https://doi.org/10.1016/j.resuscitation.2010.01.024

Arnstein, F. (1997). Catalogue of human error. *British Journal of Anaesthesia, 79*(5), 645-656. doi:10.1093/bja/79.5.645

Atwal, A., & Caldwell, K. (2005). Do all health and social care professionals interact equally: a study of interactions in multidisciplinary teams in the United Kingdom. *Scandinavian Journal of Caring Sciences, 19*(3), 268-273. doi:10.1111/j.1471-6712.2005.00338.x

Bach, D. R., Flandin, G., Friston, K. J., Dolan, R. J. (2009). Time-series analysis for rapid event-related skin conductance responses. *Journal of neuroscience methods, 184*(2), 224-234. doi:10.1016/j.jneumeth.2009.08.005.

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods, 190*(1), 80-91. doi:10.1016/j.jneumeth.2010.04.028

Berntson, G. G., Cacioppo, J. T. (2000). From homeostasis to allodynamic regulation. In Cacioppo, J. T., Tassinary, L. G., Berntson, G. (Eds.), *Handbook of psychophysiology* (2nd ed., pp. 459–481). New York: Cambridge University Press.

Boucheix, J.-M. (2017). The interplay between methodologies, tasks and visualisation formats in the study of visual expertise. *Frontline Learning Research, 5*(3), 155-166. doi:10.14786/flr.v5i3.311

Boucsein, W. (2012). *Electrodermal activity* (2nd. ed.). New York, NY: Springer Science & Business Media.

Brindley, P. G., & Reynolds, S. F. (2011). Improving verbal communication in critical care medicine. *Journal of Critical Care, 26*, 155-159. Doi:10.1016/j.jcrc.2011.03.004

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist, 45*(1), 16-28. doi:10.1037/0003-066X.45.1.16

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, *20*(1), 3-31.

Chen, H. E., & Miller, S. R. (2017). Can Wearable Sensors Be Used to Capture Engineering Design Team Interactions? An Investigation Into the Reliability of Sociometric Badges. *ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. 7*. doi:10.1115/DETC2017-68183.

Christopoulos, G. I., Uy, M. A., & Yap, W. J. (2016). The body and the brain: measuring skin conductance responses to understand the emotional experience. *Organizational Research Methods*, 1094428116681073. doi:10.1177/1094428116681073

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. doi:10.1177/001316446002000104

Cohen, R. A. (2011). Yerkes–Dodson Law. In *Encyclopedia of clinical neuropsychology* (pp. 2737-2738). Springer, New York, NY.

Cooke, M., Irby, D. M., & O'Brien, B. C. (2010). *Educating physicians: a call for reform of medical school and residency*. San Francisco, CA: Jossey-Bass.

Cooper, S., & Wakelam, A. (1999). Leadership of resuscitation teams: 'Lighthouse Leadership'. *Resuscitation, 42*(1), 27-45. doi:10.1016/S0300-9572(99)00080-5

Curhan, J. R., & Pentland, A. (2007). Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology, 92*(3), 802-811. doi:10.1037/0021-9010.92.3.802

Davis, W. A., Jones, S., Crowell-Kuhnberg, A. M., O'Keeffe, D., Boyle, K. M., Klainer, S. B., … Yule, S. (2017). Operative team communication during simulated emergencies: Too busy to respond? *Surgery, 161*(5), 1348-1356. doi:http://doi.org/10.1016/j.surg.2016.09.027

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 159-181). New York: Cambridge University Press.

de Laat, S., Endedijk, M. D., Ufkes, E. G., van Keulen, M., & de Vries, R. (2017, 24 November). *Real-time measures of social interaction as predictors for team effectiveness*. Paper presented at the WAOP conference 2017, Nijmegen, the Netherlands.

Dias, R. D., & Neto, A. S. (2016). Stress levels during emergency care: A comparison between reality and simulated scenarios. *Journal of Critical Care, 33*, 8-13. doi:10.1016/j.jcrc.2016.02.010

DiMicco, J. M., Hollenbach, K. J., Pandolfo, A., & Bender, W. (2007). The impact of increased awareness while face-to-face. *Human-Computer Interaction, 22*(1), 47-96

Doumouras, A. G., Keshet, I., Nathens, A. B., Ahmed, N., & Hicks, C. M. (2012). A crisis of faith? A review of simulation in teaching team-based, crisis management skills to surgical trainees. *Journal of Surgical Education, 69*(3), 274-281. doi:10.1016/j.jsurg.2011.11.004

Edmondson, A. C. (2003). Speaking up in the operating room: How team leaders promote learning in interdisciplinary action teams. *Journal of Management Studies, 40*(6), 1419-1452. doi:10.1111/1467-6486.00386

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors, 41*(2), 312-325. doi:10.1518/001872099779591196

Fanning, R. M., & Gaba, D. M. (2007). The role of debriefing in simulation-based learning. *Simulation in Healthcare, 2*(2), 115-125. doi:10.1097/SIH.0b013e3180315539

Fernandez Castelao, E., Russo, S. G., Riethmüller, M., & Boos, M. (2013). Effects of team coordination during cardiopulmonary resuscitation: A systematic review of the literature. *Journal of Critical Care, 28*(4), 504-521. doi:10.1016/j.jcrc.2013.01.005

Fielding, N. G. (2012). Triangulation and Mixed Methods Designs. *Journal of Mixed Methods Research, 6*(2), 124-136. doi:10.1177/1558689812437101

Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. In M. Schulte-Mecklenbeck, A. Kuehberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research: a critical review and user's guide* (pp. 163-184). New York: Psychology Press.

Fischer, F., & Järvelä, S. (2014). Methodological advances in research on learning and instruction. *Frontline Learning Research, 2*(4), 1-6.

Garbarino, M., Lai, M., Tognetti, S., Picard, R. W., & Bender, D. (2014). Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on* (pp. 39–42). Athens, Greece: IEEE.

Gibson, C. B., Cooper, C. D., & Conger, J. A. (2009). Do you see what we see? The complex effects of perceptual distance between leaders and teams. *Journal of Applied Psychology, 94*(1), 62-76. doi:10.1108/dlo.2009.08123ead.009

Goldman, S. R. (2014). Perspectives on learning: Methodologies for exploring learning processes and outcomes. *Frontline Learning Research,* 2(4), 46-55.

Grenvik, A., Schaefer, J. J., DeVita, M. A., & Rogers, P. (2004). New aspects on critical care medicine training. *Current Opinion in Critical Care, 10*(4), 233-237. doi:10.1097/01.ccx.0000132654.52131.32

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods* for *studying daily life* (pp. 43-61). New York, NY: Guilford Publications.

Hannah, S. T., Uhl-Bien, M., Avolio, B. J., & Cavarretta, F. L. (2009). A framework for examining leadership in extreme contexts. *The Leadership Quarterly*, *20*(6), 897-919.

Härgestam, M., Lindkvist, M., Brulin, C., Jacobsson, M., & Hultin, M. (2013). Communication in interdisciplinary teams: exploring closed-loop communication during in situ trauma team training. *BMJ Open, 3*(10).

Heaphy, E. D., & Dutton, J. E. (2008). Positive social interactions and the human body at work: Linking organizations and physiology. *The Academy of Management Review, 33*(1), 137-162.

Hoogeboom, A. M. G. M., & Wilderom, C. P. M. (2015). Effective leader behaviors in regularly held staff meetings: Surveyed vs. videotaped and video-coded observations. In J. A. Allen & N. Lehmann-Willenbrock & S. G. Rogelberg (Eds.), *The Cambridge Handbook of Meeting Science* (pp. 381-412). Cambridge Handbooks in Psychology. Cambridge University Press. http://dx.doi.org/10.1017/cbo9781107589735.017.

Humphrey, S. E., & Aime, F. (2014). Team microdynamics: Toward an organizing approach to teamwork. *The Academy of Management Annals, 8*(1), 443–503. doi:10.1080/19416520.2014.904140

Hunziker, S., Johansson, A. C., Tschan, F., Semmer, N. K., Rock, L., Howell, M. D., & Marsch, S. (2011). Teamwork and leadership in cardiopulmonary resuscitation. *Journal of the American College of Cardiology, 57*(24), 2381-2388. doi:10.1016/j.jacc.2011.03.017

Hunziker, S., Laschinger, L., Portmann-Schwarz, S., Semmer, N. K., Tschan, F., & Marsch, S. (2011). Perceived stress and team performance during a simulated resuscitation. *Intensive Care Medicine, 37*(9), 1473-1479. doi:10.1007/s00134-011-2277-2

Hunziker, S., Semmer, N. K., Tschan, F., Schuetz, P., Mueller, B., & Marsch, S. (2012). Dynamics and association of different acute stress markers with performance during a simulated resuscitation. *Resuscitation, 83*(5), 572-578. doi:10.1016/j.resuscitation.2011.11.013

Hunziker, S., Tschan, F., Semmer, N., Howell, M., & Marsch, S. (2010). Human factors in resuscitation: Lessons learned from simulator studies. *Journal of Emergencies, Trauma and Shock, 3*(4), 389-394. doi:10.4103/0974-2700.70764

Jacobsson, M., Hargestam, M., Hultin, M., & Brulin, C. (2012). Flexible knowledge repertoires: communication by leaders in trauma teams. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 20*(1), 44. Doi:10.1186/1757-7241-20-44

Kanki, B. G., Folk, V. G., & Irwin, C. M. (1991). Communication variations and aircrew performance. *The International Journal of Aviation Psychology, 1*(2), 149-162. doi:10.1207/s15327108ijap0102_5

Keitel, A., Ringleb, M., Schwartges, I., Weik, U., Picker, O., Stockhorst, U., et al. (2011). Endocrine and psychological stress responses in a simulated emergency situation. *Psychoneuroendocrino, 36*(1), 98–108.

Kim, T., McFee, E., Olguin, D. O., Waber, B., & Pentland, A. (2012). Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior, 33*(3), 412-427. doi:10.1002/job.1776

Klonek, F. E., Burba, M., Kauffeld, S., & Quera, V. (2016). Group interactions and time: Using sequential analysis to study group dynamics in project meetings. *Group Dynamics, 20*(3). 209-222.

Kneebone, R. L., Nestel, D., Vincent, C., & Darzi, A. (2007). Complexity, risk and simulation in learning procedural skills. *Medical Education*, 41, 808-814.

Kolbe, M., Grote, G., Waller, M. J., Wacker, J., Grande, B., Burtscher, M. J., & Spahn, D. R. (2014). Monitoring and talking to the room: Autochthonous coordination patterns in team interaction and performance. *Journal of Applied Psychology, 99*(6), 1254-1267. doi:10.1037/a0037877

Koudenburg, N., Postmes, T., & Gordijn, E. H. (2017). Beyond content of conversation: The role of conversational form in the emergence and regulation of social structure. *Personality and Social Psychology Review, 21*(1), 50-71. doi:10.1177/1088868315626022

Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest, 7*(3), 77-124. doi: 10.1111/j.1529-1006.2006.00030.x

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Emotion, motivation, and anxiety: brain mechanisms and psychophysiology. *Biological Psychiatry, 44*(12), 1248-1263. doi:10.1016/S0006-3223(98)00275-3

Larsen, R. J., Diener, E., & Lucas, R. E. (2002). Emotion models, measures, and individual differences. In R. G. Lord, R. J. Klimoski, & R. Kanfer (Eds.), *Emotions in the workplace* (pp. 64–106). San Francisco: Jossey-Bass.

Lazarus, R. S., & Folkman, S. (1984). Coping and adaptation. In W. D. Gentry (Ed.), *The handbook of behavioral medicine* (pp. 282-325). New York: Guilford.

Lei, Z., Waller, M. J., Hagen, J., & Kaplan, S. (2016). Team adaptiveness in dynamic contexts: Contextualizing the roles of interaction patterns and in-process planning. *Group & Organization Management, 41*(4), 491-525. doi:10.1177/1059601115615246

Marci, C. D, Ham, J., Moran, E., & Orr, S. P. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of nervous and mental disease, 195*(2), 103-111. doi:10.1097/01.nmd.0000253731.71025.fc

Marks, M. A., Zaccaro, S. J., & Mathieu, J. E. (2000). Performance implications of leader briefings and team-interaction training for team adaptation to novel environments. *Journal of Applied Psychology, 85*(6), 971-986. doi:10.1037/0021-9010.85.6.971

Marsch, S. C., Müller, C., Marquardt, K., Conrad, G., Tschan, F., & Hunziker, P. R. (2004). Human factors affect the quality of cardiopulmonary resuscitation in simulated cardiac arrests. *Resuscitation, 60*(1), 51-56.

Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the Journal of Applied Psychology. *Journal of applied psychology*, *102*(3), 452.

McGaghie, W. C., Issenberg, S. B., Cohen, M. E. R., Barsuk, J. H., & Wayne, D. B. (2011). Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Academic Medicine: Journal of the Association of American Medical Colleges, 86*(6), 706-711. doi:10.1097/ACM.0b013e318217e119

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009. *Medical Education, 44*(1), 50-63. doi:10.1111/j.1365-2923.2009.03547.x

Molenaar, I. (2014). Advances in temporal analysis in learning and instruction. *Frontline Learning Research, 2*(4), 15-24.

Noldus, L. P., Trienes, R. J., Hendriksen, A. H., Jansen, H., & Jansen, R. G. (2000). The Observer Video-Pro: new software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments, & Computers, 32*(1), 197-206. doi:10.3758/bf03200802

Olguín, D. O., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(1), 43-55. doi:10.1109/TSMCB.2008.2006638

Pentland, A. (2012). The new science of building great teams. *Harvard Business Review, 90*(4), 60-69.

Pugliese, A., Nicholson, G., & Bezemer, P. J. (2015). An observational analysis of the impact of board dynamics and directors' participation on perceived board effectiveness. *British Journal of Management, 26*(1), 1-25. doi:10.1111/1467-8551.12074

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sahu, S., & Lata, I. (2010). Simulation in resuscitation teaching and training, an evidence based practice review. *Journal of Emergencies, Trauma and Shock, 3*(4), 378-384. doi:10.4103/0974-2700.70758

Sandroni, C., Fenici, P., Cavallaro. F., Bocci, M. G., Scapigliati, A., & Antonelli, M. (2005). Haemodynamic effects of mental stress during cardiac arrest simulation testing on advanced life support courses. *Resuscitation, 66*(1), 39-44.

Satish, U. & Streufert, S. (2002). Value of a cognitive simulation in medicine: towards optimizing decision making performance of healthcare personnel. *Quality and Safety in Health Care, 11*(2): 163-167.

Schmutz, J., Hoffmann, F., Heimberg, E., & Manser, T. (2015). Effective coordination in medical emergency teams: The moderating role of task type. *European Journal of Work and Organizational Psychology, 24*(5), 761-776. Doi:10.1080/1359432X.2015.1018184

Spiers, J. A. (2004). Tech tips: Using video management/analysis technology in qualitative research. *International Journal of Qualitative Methods, 3*(1), 57-61. doi:10.1177/160940690400300106

Stachowski, A. A., Kaplan, S. A., & Waller, M. J. (2009). The benefits of flexible team interaction during crises. *Journal of Applied Psychology, 94*(6), 1536-1543. doi:10.1037/a0016903

Stemmler, G. 2004. Physiological processes during emotion. In P. Philippot, & R. S. Feldman, (Eds.), *Regulation of emotion* (pp. 33–70). Mahwah, NJ: Erlbaum.

Strijbos, J.-W., Martens, R. L., Prins, F. J., & Jochems, W. M. G. (2006). Content analysis: What are they talking about? *Computer & Education, 46*, 29-48.

Tannenbaum, S.I., Cerasoli, C. P. (2013).Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors*, *55*(1): 231–245.

Tschan, F., Jenni, N., Semmer, N. K., Hunziker, S., Marsch, S. U., & Kolbe, M. (2014). Leadership in different resuscitation situations. *Trends in Anaesthesia and Critical Care, 4*(1), 32-36.

Tschan, F., Semmer, N. K., Gautschi, D., Hunziker, S., Spychiger, M., & Marsch, S. U. (2006). Leader to recovery: Group performance and coordinative activities in medical emergency driven groups. *Human Performance, 19*(3), 277-304.

van der Haar, S., Koeslag-Kreunen, M., Euwe, E., & Segers, M. (2017). Team leader structuring for team effectiveness and team learning in command-and-control rooms. *Small Group Research, 48*(2), 215-248.

Vangrieken, K., Boon, A., Dochy, F., & Kyndt, E. (2017). Group, team, or something in between? Conceptualising and measuring team entitativity. *Frontline Learning Research, 5*(4), 1-41. doi:10.14786/flr.v5i4.297

van Keulen, M., Kaminski, M., Matheja, C., Katoen, J.-P. (2018). Rule-based Conditioning of Probabilistic Data. In: Proceedings of the 12th International Conference on Scalable Uncertainty Management (SUM 2018), 3-5 October 2018, Milan, Italy. Springer.

van Lier H.G. et al. (2017) Design Decisions for a Real Time, Alcohol Craving Study Using Physio- and Psychological Measures. In: de Vries P., Oinas-Kukkonen H., Siemons L., Beerlage-de Jong N., van Gemert-Pijnen L. (eds) Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors. PERSUASIVE 2017. Lecture Notes in Computer Science, vol. 10171. Springer, Cham.

Weis, P. P., & Herbert, C. (2017). Bodily reactions to emotional words referring to own versus other people's emotions. *Frontiers in Psychology, 8*, 1277. doi:10.3389/fpsyg.2017.01277

Wetzel, C.M., Black, S.A., Hanna, G.B., Athanasiou, T., Kneebone, R.L., Nestel, D., et al. (2010). The effects of stress and coping on surgical performance during simulations. *Ann Surg., 251*(1), 171–6.

Wiesenfeld, A. R., Whitman, P. B., & Malatesta, C. Z. (1984). Individual differences among adult women in sensitivity to infants: evidence in support of an empathy concept. *Journal of Personality and Social Psychology, 46*(1), 118-124. doi:10.1037/0022-3514.46.1.118

Zaccaro, S. J., Rittman, A. L., & Marks, M. A. (2001). Team leadership. *The Leadership Quarterly, 12*(4), 451-483. doi:http://doi.org/10.1016/S1048-9843(01)00093-5

Zijlstra, F. R., Waller, M. J., & Phillips, S. I. (2012). Setting the tone: Early interaction patterns in swift-starting teams as a predictor of effectiveness. *European Journal of Work and Organizational Psychology, 21*(5), 749-777. doi:10.1080/1359432X.2012.690399