Aus dem Institut für Medizinische Soziologie und Rehabilitationswissenschaft

der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Assessment, training and debriefing of team performance in medical ad hoc teams – a simulation-based approach

zur Erlangung des akademischen Grades

Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät

Charité – Universitätsmedizin Berlin

von

Julia Christina Freytag

aus St. Wendel

Datum der Promotion: 04.06.2021

Table of contents

List of tables II
List of figures II
List of abbreviations III
1 Summary of publications 1
1.1 Abstract 1
1.2 Zusammenfassung 3
1.3 Introduction 5
1.3.1 Medical error and teamwork 5
1.3.2 Simulation-based training 6
1.3.3 Debriefing 8
1.3.4 Measuring teamwork quality 9
1.3.5 Monitoring of teamwork 9
1.4 Methods 10
1.4.1 Study 1: Effectiveness of debriefing 10
1.4.2 Study 2: Rating of teamwork using TEAM 14
1.4.3 Study 3: Monitoring of teamwork 15
1.5 Results 17
1.5.1 Study 1: Effectiveness of debriefing 17
1.5.2 Study 2: Rating of teamwork using TEAM 18
1.5.3 Study 3: Monitoring of teamwork 21
1.6 Discussion 24
1.7 References 30
2 Affidavit 35
3 Share declaration 36
4 Peer-reviewed articles 38
5 Curriculum vitae 65
6 Publications and scientific contributions 67
6.1 Articles 67
6.2 Talks 67
6.3 Posters 68
6.4 Workshops 69
7 Acknowledgements 71

List of tables

Table 1 20
Variance components and percentage of variance for TEAM scores

Table 2 21
Descriptive statistics

Table 3 22
Correlations of external observations, team characteristics and performance

List of figures

Figure 1... 12
Study flow chart

List of abbreviations

ANCOVA	Analysis of covariance
CRM	Crisis/Crew/Cockpit resource management
ECC	Extracorporeal circulation
ER	Emergency room
GAS	Gather, analyse, summarise
GMA	Gesellschaft für medizinische Ausbildung (German Association for Medical Education)
GRS	Global rating scale
ICC	Intraclass correlation coefficient
M	Mean
NTS	Non-technical skills
OSAD	Objective Structured Assessment of Debriefing
OSCE	Objective Structured Clinical Examination
PCA	Principal component analysis
R	Randomisation
SBT	Simulation-based training
SD	Standard deviation
TEAM	Team Emergency Assessment Measure
TeamTAG	Teamwork Techniques Analysis Grid
TRAPD	Translation, review, adjudication, pre-testing, documentation
TS	Technical skills

1 Summary of publications

1.1 Abstract

In healthcare, teamwork is required in many situations, for example when treating a critically ill patient in the emergency room. In these situations, the quality of teamwork and team members' non-technical skills are crucial. This doctoral thesis is an investigation of training, debriefing and assessment of teamwork in medical ad hoc teams. All studies of this thesis aim at improving the process of training and evaluating teamwork – through the study of different debriefing approaches (study 1), through the validation of measurement instruments (study 2) and through examining teams' ability to monitor their teamwork (study 3).

The first study compared two debriefing approaches, which can be used to analyse and discuss teamwork after simulation-based trainings or real-life emergencies. One of the approaches included the use of TeamTAG (Teamwork Techniques Analysis Grid), a cognitive aid developed to support the debriefing process. TeamTAG was described as a feasible tool by facilitators and ensured that certain learning objectives were discussed. No differences in teamwork quality could be found after debriefing with versus debriefing without a cognitive aid.

The second study focused on the assessment of non-technical skills using the Team Emergency Assessment Measure, TEAM. This widely used tool was translated into German and validated for use by expert and novice raters. The results showed good psychometric properties for TEAM for both rater groups and a moderate interrater agreement, with a tendency to more lenient ratings from novice raters. Both experts in teamwork and emergency medicine as well as novices can therefore serve as raters for non-technical skills.

The third study examined the ability of physicians and nurses to monitor their teams' teamwork. For this, team-monitoring judgements were compared to TEAM ratings by external observers and to objective measures. It was shown that TEAM ratings correlated significantly with team members' evaluations of their teamwork but not with team leaders' evaluations. Furthermore, results showed that the quality of teamwork was associated with objective measures such as the amount of time until the first request for defibrillation.

This doctoral thesis extends the possibilities of training and assessing teamwork by providing a new cognitive aid, useful for debriefings with clear learning objectives, and a German version of the TEAM instrument, which can be used by both novice and expert raters. Apart from this, it provides further insights into the connection of teams' monitoring skills of their teamwork with external assessments of teamwork and objective performance measures.

1.2 Zusammenfassung

Im Gesundheitswesen ist Teamarbeit häufig erforderlich, zum Beispiel bei der Behandlung von schwerverletzten Patient*innen. In solchen Situationen sind die Qualität der Teamarbeit und die nicht-technischen Fertigkeiten der Teammitglieder zentral. Diese Promotionsarbeit untersucht das Training, die Nachbesprechung („debriefing“) und die Bewertung der Teamarbeit von ad hoc-Teams. Alle Studien im Rahmen dieser Arbeit haben zum Ziel, den Trainings- und Evaluationsprozess von Teamarbeit zu verbessern – durch die Untersuchung verschiedener Debriefing-Methoden (Studie 1), die Validierung eines Messinstruments (Studie 2) und die Untersuchung der Fähigkeit von Teams, ihre Teamarbeit einzuschätzen (Studie 3).

In der ersten Studie wurden zwei Debriefing-Methoden verglichen, um Teamarbeit nach simulierten oder echten Notfallszenarien zu besprechen. Der erste Ansatz beinhaltet die Anwendung des TeamTAG (Teamwork Techniques Analysis Grid), eines Leitfadens zur Unterstützung des Debriefingprozesses. TeamTAG wurde von den Instruktor*innen als praktikables Hilfsmittel beschrieben und seine Anwendung stellte sicher, dass bestimmte Lernziele besprochen wurden. In der Qualität der Teamarbeit traten keine Unterschiede auf zwischen den Teams deren Debriefings mit bzw. ohne Leitfaden durchgeführt wurden.

Die zweite Studie fokussierte auf die Bewertung nicht-technischer Fertigkeiten mittels des Team Emergency Assessment Measure, TEAM. Dieses weitverbreitete Messinstrument wurde übersetzt und validiert zur Anwendung durch Personen mit hoher und geringer Expertise im Bereich Notfallmedizin bzw. Simulationstrainings. Es zeigten sich gute psychometrische Eigenschaften für beide Anwendergruppen und eine moderate Übereinstimmung zwischen den Gruppen, wobei die Anwender*innen mit geringer Expertise zu milderen Bewertungen neigten. Demnach können sowohl Personen mit hoher als auch mit geringer Expertise zur Bewertung nicht-technischer Fertigkeiten mittels TEAM eingesetzt werden.

In der dritten Studie wurde die Fähigkeit von Ärzt*innen und Pflegekräften untersucht, die Qualität der Teamarbeit ihrer Teams einzuschätzen. Dafür wurden ihre Einschätzungen mit Beurteilungen von externen Beobachter*innen, die den TEAM nutzten, sowie mit objektiven Leistungsindikatoren verglichen. Es zeigte sich, dass die Bewertungen mittels TEAM signifikant mit der Einschätzung durch die Teammitglieder korrelierten, nicht jedoch mit der Einschätzung durch die Teamleitungen. Darüber

hinaus zeigten die Ergebnisse, dass die Qualität der Teamarbeit auch mit objektiven Kriterien korreliert, wie mit der Zeit, bis die erste Defibrillation angefordert wurde.

Diese Promotionsarbeit erweitert die Möglichkeiten, Teamarbeit zu trainieren und zu bewerten, durch die Bereitstellung eines Debriefing-Leitfadens, der besonders für Debriefings mit klaren Lernzielen nützlich ist, und einer deutschen Version des TEAM, welcher von Personen mit hoher als auch niedriger Expertise angewendet werden kann. Weiterhin bietet diese Arbeit Einblicke in die Fähigkeit von Teams, ihre Teamarbeit einzuschätzen, und darin, wie diese Fähigkeit mit externen Bewertungen und mit objektiven Leistungsindikatoren verbunden ist.

1.3 Introduction

In our modern and interconnected world, working in teams has become the norm and is almost inevitable in many lines of work. Being a team player is a common requirement in job advertisements and most people see themselves as capable of working in a team. In emergency medicine or critical care, where it needs a joint effort of highly specialised individuals to treat critically ill patients, working in teams is very common as well. However, teamwork can also be a source of error and emergency medicine is a high-stakes environment where errors can have grave consequences [1,2]. Many of these errors are not caused by a lack of clinical knowledge or technical skills (TS) but by failures in non-technical skills (NTS), like communications failures [3,4], ineffective team coordination [5,6], or a lack of leadership [7,8]. This dissertation will take a closer look at how training and debriefing can improve teamwork in medical ad hoc teams and how teamwork can be assessed. In this chapter, theoretical concepts of teamwork, medical error, simulation-based training, debriefing, and measuring and monitoring of teamwork will be introduced and will lead to the main research questions.

1.3.1 Medical error and teamwork

A *team* can be defined as a set of at least two individuals assigned to different roles or responsibilities, who have a common goal, who work together and make decisions to reach this goal, and who depend on each other in this endeavour [9,10]. In emergency medicine, highly trained individuals from different specialities form a team for a short period of time and need to improvise and coordinate their actions to succeed [11]. Because of these specific characteristics, they are referred to as interdisciplinary action teams, medical emergency teams or simply ad hoc teams [9,12]. But whereas working in teams has been shown to have advantages compared to working alone, for instance in reducing diagnostic errors [13], it can also fail, for example when a team leader gives unclear orders ('can we give some adrenaline?') not involving who should do exactly what, which increases the chance that no action is taken.

A *medical error* is defined as 'an act of omission or commission in planning or execution that contributes or could contribute to an unintended result' [14, p.42]. If this act does lead to patient harm, it is referred to as a *preventable adverse event,* an 'unintended injury to patients caused by medical management (rather than the underlying condition of the patient) that results in measurable disability, prolonged hospitalization or both' [14, p.40].

The prevalence of medical errors gained attention with the seminal report 'To err is human' from the Institute of Medicine in 1999 [15], which brought together research findings about the immense impact of medical error on patient safety. More recent studies show that medical error is the third leading cause of death in the USA [2]. In Germany, it is estimated that 5-10% of patients in hospitals experience adverse events, half of which are considered preventable. Of all patients treated in German hospitals 0.1% die of preventable causes, which comes to approximately 20,000 per year [16]. These events are mostly not caused by a single action on the part of one person, but by a system failure, which causes a chain reaction [17].

The term *non-technical skills* is widely used in healthcare to describe 'the cognitive, social and personal resource skills that complement technical skills, and contribute to safe and efficient task performance' [1, p.1]. These skills include *situation awareness* (gathering and interpreting information, anticipating possible developments), *decision making* (assessing a situation, generating and selecting a course of action, reviewing the results), *communication* (sending and receiving information, identifying barriers to communication), *leadership* (using assertiveness, planning and prioritising, managing workload, maintaining standards), and *teamwork/team working* (coordination activities, exchanging information, supporting each other, solving conflicts) [1]. *Teamwork* is also used as a generic term for the process in which these non-technical skills are applied. It 'describes interactions among team members who combine collective resources to resolve task demands' [18, p.2]. Besides *taskwork* – the team members' 'individual interaction with tasks, tools, machines and systems' [18, p.2], for which technical skills are needed – teamwork is a part of the process called *team performance* [9]. Lastly, the term *team effectiveness* describes the evaluation of a team's performance based on certain criteria [19,20].

1.3.2 Simulation-based training

Considering that failures in NTS represent a threat to patient safety [3,7,21], training those skills in team trainings is necessary to improve teamwork. The first trainings focusing on NTS were developed based on principles from aviation, where *cockpit* or *crew resource management* (CRM) trainings were started in the 1980s as a result of research into aviation errors and accidents [22,23,24]. Later, whole frameworks were developed for training and assessment of, for example, anaesthesiologists' [25] and surgeons' [26] non-technical skills. In healthcare, CRM stands for *crew* or *crisis resource management* and is based on 15 principles, such as 'anticipate and plan ahead' or 'communicate effectively' [27]. CRM aims to prevent critical incidents and – in cases where these nevertheless occur – to support the team to use all their resources effectively and manage the incident.

A widely used format to train teamwork is that of *simulation-based trainings* (SBTs). *Simulation* is 'a technique […] to replace or amplify real experiences with guided experiences, often immersive in nature, that evoke or replicate substantial aspects of the real world in a fully interactive fashion' [28, p.i2]. It can provide a safe training space for knowledge and skill acquisition and training. The theoretical background of simulation is based on experiential learning, which describes the process of learning from concrete experiences through reflecting on them, conceptualising them and applying these new insights to the next specific experience [29]. Theoretical underpinnings are also provided by the theory of situated learning, where learning is seen as a 'situated' activity happening in the context of a community of practitioners, not as an isolated activity [30,31]. The learning surroundings should therefore reflect the environment in which the task will have to be later carried out, as simulation is (to a certain degree) able to do. In the most common format for training teamwork, teams of healthcare professionals from one or several disciplines participate in certain scenarios (treating a patient in the form of a simulated patient or a simulator/manikin) in a replicated environment of a simulation centre or directly at their workplace (in situ trainings). Many trainings focusing on teamwork are based on the concept of CRM.

Studies have shown that such SBTs can improve knowledge, skills, behaviours, team performance and patient outcome, reduce stress and anxiety among learners and are effective when compared to no intervention or to traditional training [32,33,34,35,36]. Furthermore, research has provided initial indications that technical performance and non-technical performance are connected, at least in stressful situations [37].

Team trainings are now widespread in many disciplines of post-graduate education, but remain scarce in undergraduate education [38,39], which may be one of the reasons why final year medical students and junior doctors feel unprepared for clinical practice, especially for emergency care [40,41]. For a long-lasting effect and the integration of trained behaviours into clinical practice, however, an early implementation of teamwork trainings in medical education and repetitive practice are key components.

What is also important is that, though good teamwork is clearly linked to effective overall team performance [18,42], the exact meaning of “good teamwork” largely depends on the situation a healthcare worker occupies and the role that person has [8,43]. Especially in emergency and acute care, where things can change rapidly, adaptation is a key principle. This means that communication or leadership styles also need to be adapted to specific situations [44].

1.3.3 Debriefing

As mentioned before, to use the learning opportunity the simulated encounter provides, it should be reflected upon. Therefore, SBTs include a *debriefing*, a guided reflection as part of the experiential learning cycle that ‘helps learners develop and integrate insights from direct experience into later action’ [45, p.1010]. Different forms of debriefing exist, based on when they take place (post-simulation versus intra-simulation), whether or not a facilitator is included (facilitator-guided versus learner-guided) and the framework that is used (e.g. three-phase models like GAS [46] or multiphase models like TeamGAINS [47]). Alongside these aspects, adjuncts such as debriefing scripts or video review are often used [48]. Very common are frameworks including three phases, with the GAS approach as a widely used example. GAS stands for *gather* – participants share their simulation experience and perspective, *analyse* – participants’ actions are analysed and the facilitator uses questions to gain more insight and facilitate reflection, and *summarise* – the final phase dedicated to summarising what has been learned [46].

Debriefing is considered a crucial part in learning from simulated or real-life events. Post-event debriefings, for example, are recommended by different international resuscitation guidelines [49,50]. Nevertheless, studies into the various kinds of debriefing have not provided clear evidence concerning which debriefing approaches or methods provide the best learning – and ultimately patient outcome [48]. One possibly helpful tool are debriefing scripts, which can serve as a cognitive aid for the facilitator [51] and offer guidance on how to structure debriefings and to determine which topics to discuss. In a study of an SBT with novice instructors, participants showed a bigger improvement of their knowledge and of team leader behaviour when instructors used a debriefing script [52]. To improve undergraduate medical education, there is a need to know if such findings can be transferred to this area. Study 1 of this thesis [53] therefore analyses the effect of using a debriefing script during an SBT for undergraduate medical students [54].

1.3.4 Measuring teamwork quality

With one of the learning outcomes being a higher quality of teamwork, it is very important to measure teamwork quality accurately and reliably. Additionally, observing and assessing teamwork is a crucial method for collecting evidence about the connection between teamwork and patient outcome and for addressing the question of which team behaviours may lead to beneficial outcomes in a specific situation. The Team Emergency Assessment Measure (TEAM) is a well-established tool for measuring teamwork quality in simulated and real-life emergencies [55,56,57]. To make it accessible in German-speaking contexts, it was translated into German in study 1, and validated in study 2 of this thesis [53,58].

As most studies using the TEAM instrument so far have employed expert raters (i.e. physicians/nurses with experience in emergency medicine or researchers with experience in medical simulation), it is not yet clear how raters with less experience use TEAM compared to experts. Novice raters, however, have the advantage of being more easily available, thus enabling a wider and more frequent application of teamwork assessment. We therefore compared novice and expert raters using TEAM when assessing teamwork in different simulated scenarios in study 2 [58].

1.3.5 Monitoring of teamwork

Good teamwork requires a lot of training – which can happen in SBTs – but also needs to be transferred to and continually trained in clinical practice. Facilitator-guided debriefings to discuss and improve teamwork can be conducted in “real life” as well, but resources are often scarce and trained observers/facilitators are usually not present in everyday clinical practice. To recognise when a debriefing or a teamwork training is necessary, it would be extremely helpful if healthcare practitioners were able to monitor and assess their teamwork in a reliable way. Team-led debriefings would then enable teams and their individual members to learn from critical incidents in clinical practice and not just from designated simulation trainings. This aspect is also included in the concept of *team reflexivity*, which describes a ‘team’s ability to reflect collectively on group objectives, strategies (e.g. decision making), processes (e.g. communication), and the outcomes of past and current performance and adapt accordingly' [59, p.1]. Several frameworks exist to guide teams in this reflective debriefing process [60,61]. For these reflections to be helpful, the teams' monitoring of their teamwork must be accurate. This is especially important for team leaders, as it is often the job of the physician team leader or the charge nurse to lead this debriefing and – before that – to decide whether it is necessary at all [61]. Research indicates that *individuals* are in principle able to self-monitor their current performance, whereas it is much harder for them to self-assess their skills across situations in the form of a more general self-evaluation [62,63]. Study 3 of this thesis examines team members' and team leaders' monitoring of their teamwork and analyses whether their monitoring judgements relate to external observers' ratings of their teamwork, so they could be used as a basis for team reflection when no facilitator is around to observe and debrief their performance [64]. As a second step, study 3 examines, whether team-monitoring judgements are related to objective performance measures (for instance, early defibrillation in resuscitations), which are connected to patient outcome.

In sum, the three studies of this dissertation aim to advance research into team training, assessment, and monitoring by exploring the following research questions:

- Which effect does debriefing with a debriefing script have on the learning outcome of an SBT and the feasibility of the debriefing process compared to debriefing without a script (study 1)?

- Do raters with little experience as opposed to subject matter experts provide comparable evaluations of teamwork using the German version of TEAM [55] (study 2)?

- How do team members' and leaders' judgements of their teamwork relate to external observations of their teamwork and to objective performance measures (study 3)?

1.4 Methods

1.4.1 Study 1: Effectiveness of debriefing

To assess whether debriefing with a debriefing script leads to better learning outcomes and is more feasible than debriefing without a script, a prospective superiority study comparing a control and an intervention group was designed. In the following, study 1 will be described briefly; a more detailed description can be found in the related article [53].

Setting and participants

The study was conducted during an SBT for medical students, in which an eight-hour nightshift in the emergency room (ER) was simulated. The study took place in January 2017. This “nightshift” consisted of six scenarios, representing common cases in emergency medicine [54]. 32 medical students in their final year took part in the simulation, after they had given their informed consent. The students were randomly assigned to seven teams of (four to) five students (that is, three teams in the control condition, four in the intervention). In every scenario, one student was declared team leader, two students were team members and the remaining two observed the scenario. Within each team, roles were alternated during the “nightshift”.

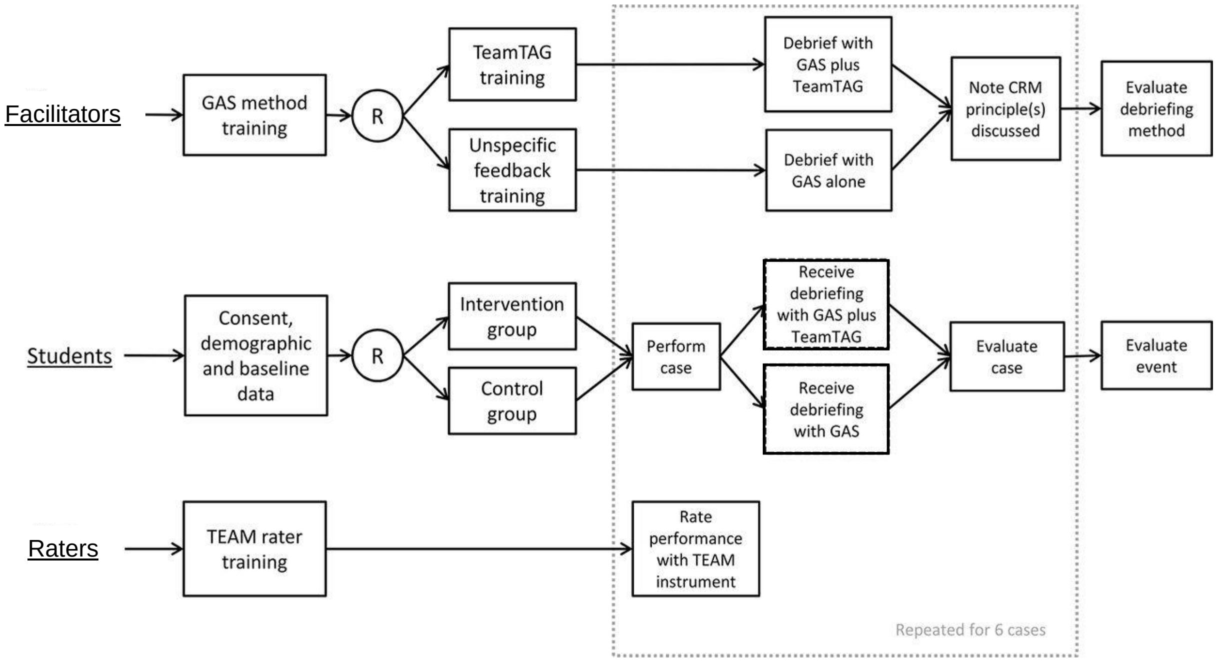
The ethics committee (EA2/172/16) and the institutional office for data protection (AZ 737/16) at Charité – Universitätsmedizin Berlin approved the study.

Debriefing

Prior to the SBT, the facilitators – experienced peer teachers – received a training on CRM principles, debriefing, and the GAS approach. After each scenario, multi-source feedback was provided, including a facilitator-guided team debriefing (of around ten minutes’ duration) focusing on teamwork. In the control condition, debriefings were conducted according to the GAS approach. For this, control group facilitators were instructed to use the GAS approach and freely choose one or two CRM principles, fitting the observed scenario, as topics for each debriefing. In the intervention teams, the GAS approach was used as well, but combined with a cognitive aid called *TeamTAG* (Teamwork Techniques Analysis Grid). For this, intervention group facilitators were instructed to use the GAS approach *and* the TeamTAG and to discuss all included CRM principles at least once (intervention group). The sequence of these events is depicted in the study flow chart (Figure 1).

Figure 1

Study flow chart



Note. CRM: crisis resource management; GAS: gather–analyse–summarise; R: randomisation; TEAM: Team Emergency Assessment Measure [55]; TeamTAG: Teamwork Techniques Analysis Grid. The grey dotted box indicates that this part of the study is repeated six times. This figure was published as part of the referenced study [53], and is used under CC BY-NC 4.0; it has been adapted for use in this dissertation.

TeamTAG

The TeamTAG has been developed by our research group. It is a debriefing script based on six CRM principles and aims to support the facilitator in observing and debriefing simulation trainings and in teaching basic teamwork skills to the participants by means of giving feedback. Two investigators from the research group selected the principles to fit the simulation setting, the participants' skill level, the facilitators' experience, and the principles' observability. TeamTAG fits on one A4 page and includes, apart from the principles, behavioural anchors for each principle and blank space for notes.

Data collection

Before the simulation started, all participants filled out a questionnaire on demographic data, their experience in emergency medicine and whether they had professional training as a paramedic or nurse to identify potential confounders. In their teams of five participants, they were instructed to discuss what teamwork principles they were familiar with (to assess their knowledge of NTS) and answered 15 multiple-choice questions about emergency medicine (to assess their knowledge of this field).

The quality of teamwork was assessed during each scenario by two raters using the TEAM tool [55], which was translated into German and validated in a pre-study. In this pre-study, inter-rater reliability was calculated between four raters who each rated two videos (one video showing intermediate quality teamwork, one high quality teamwork). Inter-rater reliability was excellent for both videos (video 1: intraclass correlation coefficient [ICC] = .99, $\text{mean}_{\text{TEAM}}$ [*M*] = 42.3, standard deviation [*SD*] = 1.3; video 2: ICC = .85, $M_{\text{TEAM}} = 22.5$, $SD = 3.1$) [53,65]. TEAM will be described in more detail as part of the second study. The raters were blinded to the condition the teams were in.

After every scenario and debriefing, the participants were asked to rate whether they found the debriefing to be helpful on a 7-point Likert scale from +3 (strongly agree) to −3 (strongly disagree). Facilitators tracked the topics that were discussed, and afterwards matched them to CRM principles to calculate the total number of principles discussed in each team.

At the end of the whole SBT, participants were asked to rate the importance of all 15 CRM principles. Furthermore, the facilitators rated the feasibility of the debriefing process in general and, in the intervention condition, of TeamTAG in particular. Both questionnaires used 7-point Likert scales as well.

TeamTAG has already been used in an earlier nightshift simulation (without a control group) [65,66] and was evaluated as feasible ($M = 1.9$, $SD = 0.9$) and helpful to observe and debrief the simulation ($M_{\text{observe}} = 2.3$, $SD = 0.8$; $M_{\text{debrief}} = 2.3$, $SD = 0.5$) by seven facilitators.

Data analysis

The data was analysed using the programs SPSS 24 (Armonk, NY: IBM Corp.) and R, version 3.4.4 [67]. Baseline data was analysed using qualitative and quantitative methods to check for potential confounders with parametric and non-parametric tests. The TEAM scores of the intervention and control condition teams, as well as the number of CRM principles discussed in the two conditions, were compared for statistical differences. Furthermore, participants' ratings of their satisfaction with the debriefing and the importance of the CRM principles were compared between the control and the intervention condition. P values of less than 0.05 were considered significant in this and the following studies. Facilitators' rating of the feasibility of the debriefing process was analysed descriptively.

1.4.2 Study 2: Rating of teamwork using TEAM

Study 2 assesses whether novices provide evaluations of teamwork that are comparable to the ones made by experts using the German version of TEAM, an instrument to rate the quality of teamwork in emergencies. Study 2 is described in detail in the related publication [58], and a summary follows below.

Setting and raters

The study took place in an SBT for medical students, which has already been described as part of study 1. In this SBT, seven teams of five medical students rotated through six emergency medical scenarios and their teamwork was rated by two observers using the TEAM tool in each of these scenarios, namely by one expert rater and one novice rater. The group of expert raters consisted of five physicians and one psychologist with broad experience in training and assessing teamwork skills in SBTs and/or in emergency medicine. The novice raters were six peer teachers from the skills lab. They were advanced medical students who had gained experience in emergency medicine through clinical electives or by working as paramedics. All raters received a rater training, including information about the TEAM tool, common rating errors and a frame-of-reference training using videotaped examples of teamwork [68].

The ethics committee (EA2/172/16) and the institutional office for data protection (AZ 737/16) at Charité – Universitätsmedizin Berlin approved the study.

TEAM

TEAM was developed in 2010 by Cooper and colleagues [55] as a tool to train and assess teamwork and has been used in a variety of environments (SBTs, real-life emergencies) and with different teams (medical/nursing students, multi-professional teams) since then [55,69,70]. Both the English original and the translated French version have been validated and show very good psychometric properties [56,71,72]. TEAM consists of eleven items in the three subscales *leadership*, *teamwork* (including communication, coordination and cooperation, team climate, adaptability, and situation awareness) and *task management*. All items can be rated on a scale from 0 (behaviour can hardly ever/never be observed) to 4 (behaviour can nearly always/always be observed). A twelfth item asks for a global rating of the team's non-technical performance (GRS = global rating scale) on a scale of 1 to 10. Furthermore, a sum score can be calculated by adding up the scores of the eleven items.

TEAM was translated into German by our research team following the TRAPD (translation, review, adjudication, pre-testing, and documentation) methodology [73]. As a preliminary validation, two videos were evaluated by four raters and showed excellent inter-rater reliability (see methods section: methods of study 1 on effectiveness of debriefing [53]).

Data analysis

The collected data were analysed using IBM SPSS 25.0 and R [67]. Descriptive measures were calculated for novice and expert raters separately. TEAM's reliability (Cronbach's α), the item-total-score correlation and the correlation of all items, including the sum score with the score of the GRS, were calculated for both rater groups. Furthermore, a principal component analysis (PCA) was conducted to analyse the underlying structure of the data – again separately for experts' and novices' ratings. To compare their ratings, inter-rater reliability between the two groups was calculated and Mann-Whitney *U* tests and a *t* test were calculated to compare the ratings of all eleven items, respectively of the sum score and the GRS.

As a last step, the sources of variance in the GRS ratings were analysed using a mixed effects model. The variance components for raters, rater status (novice or expert), the teams and the different scenarios were estimated, as well as their first-order interactions.

1.4.3 Study 3: Monitoring of teamwork

To examine how well medical emergency team leaders and members can monitor their teamwork, an experimental study was conducted during an SBT for the medical staff of a university-affiliated level-one emergency room. Study 3 is described in detail in the corresponding publication [64], and a summary follows.

Setting and participants

The SBT was part of the annual interprofessional training day, which was mandatory for all physicians and nurses at the particular hospital in Switzerland. Part of the training was a simulated resuscitation, which was happening in situ in the actual emergency room. Participants were randomly assigned to teams, stratified by profession, to make sure that each team consisted of at least one physician and two nurses. The most senior physician was assigned the role of team leader.

The study was deemed exempt from full ethical review by the ethics committee of the Canton of Bern, because it did not involve patients. All participants provided written informed consent.

Simulation scenario

The teams' task was to treat an elderly male patient who was found by paramedics lying on the floor in his cold flat, due to a fall ten hours previously. The patient went into cardiac arrest just prior to arriving at the hospital and the paramedics started cardiopulmonary resuscitation. At this point, the patient was handed over to the ER team. The patient, embodied by a simulator, showed ventricular fibrillation, and did not respond to treatment for the next 15 minutes. The scenario was concluded when the teams either called for an extracorporeal circulation (ECC) due to hypothermic arrest (patient's body temperature was 27.3°C) or diagnosed a different cause for the cardiac arrest. A facilitator-led debriefing focusing on teamwork followed the simulation.

Measurement instruments

Prior to the start of the SBT, participants filled out a questionnaire about demographic data (age, gender) and work experience (in years). Right after the simulation, every participant was asked to rate *how familiar* they were with cases like the one they had just worked on and to rate their team's teamwork, phrased as their *confidence in the quality of their teams' teamwork* on 5-point Likert scales, ranging from 1 (not confident at all) to 5 (very confident).

The simulations were recorded, and two researchers using the German version of TEAM rated the teamwork. As they showed a good inter-rater agreement after rating 20% of the videos (ICC = 0.87), the remaining videos were rated by one observer. Additionally, objective performance measures, like the time until the first defibrillation or the first administration of adrenalin was requested, were extracted from the recordings.

Data analysis

IBM SPSS 25.0 was used for data analysis. Descriptive measures were calculated using mean or median and standard deviation or range, as appropriate. Pearson's *r* was calculated as a correlation coefficient between the participants' evaluation of their teamwork and the observers' ratings of TEAM (sum score and GRS), as well as between participants' evaluation and the described objective measures. In addition, the TEAM sum score and GRS were correlated with objective measures. As a last step, observers' TEAM ratings were correlated with different team features, such as the team's size and the age of the team leader.

1.5 Results

1.5.1 Study 1: Effectiveness of debriefing

As the first publication is a study protocol, which was submitted before the study was conducted and was published in *BMJ Open* in June 2017 [53], it only contains preliminary results. The final results, which were partly presented at a scientific conference [74] and were submitted for publication in *GMS Journal for Medical Education* on May 29th 2020 (decision pending), are briefly described here.

Effects of debriefing with and without debriefing script

32 medical students took part in the study and were randomised to four intervention group teams (n = 19) and three control group teams (n = 13). The confounder analysis yielded no significant differences between control group and intervention group participants concerning their age, gender, previous experience, or knowledge of emergency medicine. Only the number of CRM principles which participants knew prior to the start of the SBT was higher in the control group ($M_{\text{CRM\_control}} = 5.7$, $SD = 0.6$; $M_{\text{CRM\_intervention}} = 4$, $SD = 0$), $t(2) = 5.00$, $p = .04$.

We analysed the potential effect of using TeamTAG to debrief teamwork using an ANCOVA (analysis of covariance), with the TEAM sum score of the last case as the dependent variable and the condition (intervention vs. control) as independent variable. The prior knowledge of CRM principles, TEAM sum score of the first case and the type of the last case were included as covariates. This analysis showed no effect of the condition on the TEAM score of the last case ($F(1,1) = 7.38$, $p = .23$).

Facilitators of control and intervention group teams differed in the topics they discussed in the debriefings. Intervention group facilitators discussed the topics included in the TeamTAG more consistently (median = 5; min = 4; max = 6) than control group facilitators (median = 3; min = 2; max = 5). Furthermore, intervention group facilitators tended to discuss different topics in each debriefing, whereas topics were repeated more often by control group facilitators. Overall, $M = 7.50$ CRM principles were discussed on average in the intervention condition ($SD = 1.29$) and $M = 6.33$ in the control condition ($SD = 3.06$), whereby $t(5) = -0.70$, $p = .51$.

The facilitators rated debriefing with as well as debriefing without the TeamTAG as feasible ($M_{intervention}$ = 2.25, *SD* = 0.96; $M_{control}$ = 2.00, *SD* = 1.00), $t(5) = -0.34$, $p = 0.75$. Intervention group facilitators reported that they had enough time to debrief their team ($M_{intervention}$ = 2.50, $SD_{intervention}$ = 0.587), whereas control group participants were undecided ($M_{control}$ = 0.33, $SD_{control}$ = 2.08), $t(2.23) = -1.75$, $p = 0.21$. The participants were satisfied with the debriefings throughout the whole SBT (*M* = 2.4-2.9), and no significant differences between participants of the intervention and the control condition were detectable (all $p \geq .06$). The same applied to participants' assessment of the importance of the CRM principles. All principles were thought to be (very) relevant (*M* = 1.9-2.8), and the central tendency of these ratings did not differ between the two conditions (all $p \geq .06$).

In sum, this study revealed no significant differences between the teams of intervention and control condition regarding the teamwork quality at the end of our nightshift simulation and the total number of CRM principles that were discussed in the debriefings. However, in intervention group teams more principles included in the TeamTAG were discussed and less principles were discussed repeatedly, compared to the control group. Facilitators of both conditions found the debriefing process to be feasible, the ones using TeamTAG reported less problems with time management. Participants were satisfied with the debriefings and rated all CRM principles to be relevant, no matter which condition they were in.

1.5.2 Study 2: Rating of teamwork using TEAM

The results described in this paragraph have been published in the *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* in February 2019 [58].

Descriptive statistics

In total, twelve raters (six novices and six experts) participated in the study. Each scenario of the SBT was rated by one expert rater and one novice rater using TEAM. As seven teams rotated through six scenarios, this yielded 84 observations of teamwork. The novice raters (aged 20-33, *M* = 24 years) were peer teachers with 1 to 2.5 years of experience in student-assisted learning, the expert raters (aged 26-37, *M* = 31.5 years) had 3.5 to 10 years of experience in teaching in clinical settings and/or SBTs and partly as well in emergency medicine. The participants were medical students in their final year of study; their age ranged from 22 to 46 years (*M* = 26.5 years, *SD* = 4), and 46.9% were female.

Measurement properties of German TEAM

The measurement properties for the German TEAM were calculated separately for the novice and the expert ratings. To analyse the TEAM's reliability, we calculated Cronbach's α for each of the six cases, which resulted in a mean Cronbach's α of $\alpha = 0.89$ ($SD = 0.06$) for experts and $\alpha = 0.85$ ($SD = 0.19$) for novices. The lowest α of expert ratings was $\alpha = 0.79$, while for novice raters it was $\alpha = 0.47$. Item-total-score correlations were calculated to check if the items were able to discriminate between high- and low-quality teamwork. All items were positively correlated with the sum score of TEAM with a mean of $M_{\text{corr(experts)}} = 0.71$ ($SD = 0.09$) and $M_{\text{corr(novices)}} = 0.69$ ($SD = 0.17$) across all cases. Likewise, all items were positively correlated with the GRS (mean correlation of $M_{\text{corr(experts)}} = 0.71$, $SD = 0.10$, $M_{\text{corr(novices)}} = 0.69$, $SD = 0.17$) as was the sum score with $r = 0.90$ for experts ($p < .001$) and $r = 0.85$ for novice raters ($p < .001$).

The next part of the analysis involved a PCA to examine whether the TEAM's items could be reduced to one general component. First, all prerequisites were checked (i.e. inter-item correlations at least 0.3, Kaiser–Meyer–Olkin criterion at least 0.6, significant Bartlett test of sphericity) and – since these preconditions were met – a PCA was conducted for expert and novice ratings. For both rater groups a dominant first component was found, explaining 59% (experts) and 65% (novices) of the observed variance.

Inter-rater agreement

The inter-rater agreement between novice and expert raters calculated for the sum scores was ICC = 0.66. The two rater groups also had a 75% agreement when comparing which teams' teamwork was rated among the two lowest and the two highest for every case. Furthermore, we compared the ratings on an item-level using *U*-tests. No statistically significant differences were found for seven of the eleven items ($p = .06$–$.86$) and for the sum scores ($M_{\text{novice}} = 30.4$, $SD_{\text{novice}} = 8.6$, $M_{\text{expert}} = 27.0$, $SD_{\text{expert}} = 8.4$; $t(82) = 1.8$, $p = .08$). On four of the eleven items, a statistically significant difference was revealed with novices rating teamwork behaviour higher than experts ($p = .04$–$.004$). This was also true for the GRS, where novices on average gave higher ratings ($M_{\text{novice}} = 7.1$, $SD_{\text{novice}} = 1.6$) as compared to experts ($M_{\text{expert}} = 6.1$, $SD_{\text{expert}} = 1.9$), $t(82) = 2.5$, $p = .02$.

Sources of variation of TEAM scores

A mixed effects model was used to estimate which sources contribute how much to the total variance of TEAM score on the GRS. Random effects were estimated for raters, cases, rater status (expert/novice) and team (the seven groups of participants), and for the first-order interactions between cases and teams and between cases and rater status. All components together explained 71.8% of the observed variance (Table 1). Rater status – being a novice or an expert rater – as the primary focus of this study, accounted for approximately 11% of variance of TEAM scores. Besides the cases, which accounted for 10.17% of variance, the interaction of case and team was the greatest source of variance. Explaining 43.21% of variance, it indicates the teams' teamwork quality differed a lot between cases.

Table 1

Variance components and percentage of variance for TEAM scores

Source of variance	Variance component	Percentage of variance
Rater[a]	0.048	1.32
Rater status[b]	0.397	11.05
Team	0.094	2.62
Case	0.366	10.17
Case × Team	1.553	43.21
Rater Status × Case	0.123	3.42
Residual	1.014	28.21

Note. TEAM: Teamwork Emergency Assessment Measure [55]. This table was published as part of the referenced study [58] and is used under CC BY 4.0; it has been adapted for use in this dissertation.

[a] Rater includes all 12 raters. [b] Rater status includes the categories 'novice rater' and 'expert rater'.

In sum, different measurement properties, like Cronbach's α and item-total-score correlations were calculated for novice and expert raters separately. Two PCAs showed that for both rater groups, the predominant amount of variance can be explained by one component. Looking at the TEAM's sum score, novice and expert raters had an inter-rater agreement of ICC = 0.66 and the mixed effects model showed that rater status explained 11% of the observed variance of TEAM scores. There was no significant difference between experts' and novices' ratings for seven of eleven

TEAM items and the sum score; for four items and the GRS novices' ratings were higher than experts' ratings.

1.5.3 Study 3: Monitoring of teamwork

The results of this study have been published in *BMC Medical Education* in June 2020 [64]. A summary of these results is provided in this section.

Descriptive statistics

Originally, 26 teams participated in the study. Four teams had to be excluded for various reasons (technical failure while recording, no consent, team included one of the researchers). Thus, data from 22 teams, consisting of 115 healthcare professionals (22 physician team leaders and 93 team members), was analysed.

Table 2

Descriptive statistics

	Mean	Standard deviation	Min	Max
Participants overall [N = 115]				
Age [years, N = 113]	36.7	10.3	20	63
Professional experience [years, N = 112]	13.1	10.3	0.4	39
Experience emergency medicine [years, N = 110]	6.3	7.5	0.2	36
Gender [N = 115]		77.4% female		
Team leader [N = 22]				
Age [years, N = 22]	36.4	8.3	29	61
Professional experience [years, N = 22]	9.0	8.6	2	37
Experience emergency medicine [years, N = 22]	3.4	5.8	1	27
Gender [N = 22]		77.3% female		
Scenarios and performance				
Total duration [min:sec]	8:15	1:05	5:45	10:15
Time to defibrillation [min:sec]	1:06	0:42	0:09	2:32
Time to adrenaline [min:sec]	3:31	1:34	0:50	6:47[a]
TEAM GRS score [points out of 10]	6.4	1.8	3	10
TEAM sum score [points out of 44]	31.8	5.5	21	42

Note. TEAM: Team Emergency Assessment Measure. This table was published as part of the referenced study [64] and is used under CC BY 4.0; it has been adapted for use in this dissertation.

[a] Two teams did not administer adrenaline.

The teams consisted of three to six members (median = 4), of whom one to three were physicians (median = 1) and two to four were nurses (median = 3). The participants were on average 36.7 years old, had worked for a mean of *M* = 13.1 years, and of those *M* = 6.3 years in emergency medicine (

Table ***2***).

On average, a simulation lasted 8min 15s (*SD* = 1:05 min) – at this point either the team called for an ECC (N = 9) or the instructor concluded the simulation (N = 13). The time until the first request for defibrillation ranged from 9s to 2min 32s; the range for the time until the first request for adrenaline ranged from 50s to 6min 47s (

Table ***2***). The observed teamwork quality varied noticeably among the teams. The TEAM sum scores ranged from 21 to 42, with a mean of 31.8 (*SD* = 5.5). The GRS showed similar results (*M* = 6.4; *SD* = 1.8; min = 3; max = 10; Table 2), as sum score and GRS were highly positively correlated (*r* = 0.943; Table 3).

Table 3

Correlations of external observations, team characteristics and performance

	external observations		objective performance measure		team characteristics		
	TEAM GRS Score	TEAM sum score	time to defibrillation	time to adrenaline	team size	number of physicians	number of nurses
TEAM GRS Score	1	0.943**	−.463**	−.217*	0.055	−0.072	0.116
TEAM sum score		1	−.451**	−.226*	0.061	−0.070	0.120
time to defibrillation			1	.486**	−0.048	0.193	−.214*
time to adrenaline				1	−0.093	0.202	−.284**
team size					1	.421**	.627**
number of physicians						1	−.443**
number of nurses							1

Note. TEAM: Team Emergency Assessment Measure [55]; GRS: global rating scale. This table was published as part of the referenced study [64] and is used under CC BY 4.0; it has been adapted for use in this dissertation.

*p < .05, two-tailed. ***p* < .01 two-tailed.

Relationship between team monitoring and TEAM scores

The relationship of team-monitoring judgements and TEAM scores was analysed for team members and team leaders separately. The observers' TEAM ratings were significantly correlated with the team members' judgements ($r_{TEAM\_members\text{-}monitor} = 0.573$, $p < 0.001$, $r_{GRS\_members\text{-}monitor} = 0.628$, $p < 0.001$), but not with the team leaders' judgements ($r_{TEAM\_leader\text{-}monitor} = 0.347$, $p = 0.145$, $r_{GRS\_leader\text{-}monitor} = 0.451$, $p = 0.052$) of their teamwork.

Relationship between team-monitoring and objective performance measures

The relationship of team-monitoring judgements and objective measures was analysed separately for team members and team leaders, as well. Team members' judgements were significantly correlated with the time until defibrillation was requested and administered ($r_{defi\text{-}req} = -0.459$, $p < 0.001$, $r_{defi\text{-}admin} = -0.295$, $p = 0.010$) and until adrenalin was requested ($r_{adrenalin\text{-}req} = -0.271$, $p = 0.025$). Team leaders' judgements were not significantly correlated with these measures ($r_{defi\ \text{-}req} = -0.447$, $p = 0.055$, $r_{defi\text{-}admin} = -0.290$, $p = 0.229$, $_{adrenalin\text{-}req} = 0.147$, $p = 0.572$).

Relationship between TEAM scores and objective performance measures

As depicted in Table 3, both TEAM scores (GRS and sum score) were significantly correlated with the length of time until the request for first the defibrillation ($r_{GRS} = -0.463$, $p < 0.001$; $r_{sum} = -0.451$, $p < 0.001$) and the request for the first administration of adrenaline ($r_{GRS} = -0.217$, $p = 0.035$; $r_{sum} = -0.226$, $p = 0.028$), with a higher TEAM score indicating less time until the requests for defibrillation/adrenaline.

Relationship between TEAM scores and team characteristics

Team size, number of physicians and number of nurses showed no significant relation to the TEAM scores (Table 3). Physicians' age was negatively correlated with both TEAM scores ($r_{sum\_leader\text{-}age} = -0.461$, $p = 0.047$, $r_{GRS\_leader\text{-}age} = -0.473$, $p = 0.041$).

In sum, the results show that team members' team-monitoring judgments are significantly correlated with observers' TEAM ratings and with different objective performance measures, while the same could not be shown for team leaders' judgements. TEAM scores were also correlated with objective performance measures but showed no connection to the team size or the number of physicians/nurses per team.

1.6 Discussion

This dissertation is dedicated to the topic of evaluating teamwork; in particular, it examined different aspects of training, debriefing, and assessing teamwork in medical emergencies. It involved the development of TeamTAG, a debriefing script, psychometric testing of the German-language version of the TEAM tool when used by expert and novice raters, and the comparison of TEAM scores with team members' and leaders' team-monitoring judgements and with objective performance measures.

Effectiveness of debriefing with TeamTAG

In the first study, debriefing with TeamTAG, a debriefing script, was compared to debriefing without a script to determine the effect of these two approaches on the learning outcome of an SBT and the feasibility of the debriefing process. TeamTAG is a cognitive aid for facilitators to structure the process of observing and debriefing teamwork, to trigger reflections and to reduce their mental workload. In our study, it was used by novice raters, who described it as feasible and better suitable for managing the debriefing time compared to raters who did not use TeamTAG. Our results showed that debriefing scripts such as TeamTAG can be useful to ensure that certain learning objectives will be discussed during an SBT. This might be particularly important for courses at the beginner level, for example, basic courses on CRM for students of healthcare professions. On the downside, scripts can lead to a less flexible debriefing, as they predefine the behaviours to observe and discuss. For participants with more experience, a more "open" and learner-centred approach may be more helpful [48,75]. It is important to note, though, that such an open and learner-centred approach is more complex than a scripted debrief and therefore calls for more skilled and experienced facilitators [76]. Therefore, when deciding on a debriefing approach, the experience level of both participants and facilitators, as well as the learning objectives, need to be considered.

Apart from the feasibility and the different foci of debriefings with and without TeamTAG, we could not detect any differences regarding the learning outcome between the two debriefing approaches. This contradicts former findings where using a debriefing script led to a greater improvement in participants' knowledge and team leader performance [52]. In the setting of our nightshift SBT, the changing team roles as well as the experience of fatigue might have hindered a more effective learning outcome and reduced the possibility of detecting differences between the debriefing approaches. Furthermore, the debriefing script in the previously mentioned study included more detail and more time was provided to conduct the debriefing, which might have supported the novice facilitators better and allowed for more reflection and learning to happen [52]. Thus, to conclude, while the positive effect of debriefing in general is clear [77], there might not be one particular debriefing approach, which is ‘the best’ [48]. As mentioned before, it needs to fit the context, the learning goals as well as the facilitators’ experience and preferences [78].

For further exploration of the benefits of TeamTAG, it should be incorporated into an SBT with a fixed team structure, more participants, and a clearer aim on solely training non-technical skills. A direct comparison to other debriefing scripts used by novice instructors would increase the knowledge about the specific needs of this group of instructors, for example, regarding the optimal level of detail of the script. In addition to an assessment of teamwork quality at the beginning and at the end of the SBT, a second post-test of teamwork quality several weeks after the training should be included in future studies, to allow for more insights into how long potential training effects last.

Rating of teamwork using TEAM by novice and expert raters

To train and debrief teamwork optimally, it is important to have reliable and valid assessment tools for teamwork at hand. Hence, this thesis also focused on assessing NTS using the German TEAM tool and especially on the question of whether novices can provide assessments, comparable to those of experts, and could therefore be employed as raters as well. TEAM is an assessment and training tool, consisting of 11 items and a global rating, to evaluate three categories – leadership, teamwork, and task management. The psychometric properties of the German TEAM showed good results, for both novice and expert raters, as well as when compared to the English original and the French translation [58]. The internal consistency was high and the PCA confirmed one underlying component for both novice and expert raters. The inter-rater agreement of novices and experts can be considered moderate to good. The results of the variance component estimation endorsed the use of novices and experts as potential raters as well, since rater status had only a small influence on the variance of TEAM scores.

Novices are also used in other areas to rate performance, such as in Objective Structured Clinical Examinations (OSCE). A variety of studies shows that novices, after being trained to use the rating tool, can provide reliable ratings of TS and NTS [79,80,81]. Ratings are especially consistent when looking at global ratings, with novices' ratings tending to be slightly higher than experts', whereas novices' checklist scores (for particular behaviours) are more variable and can be higher than, lower than or similar to experts' scores [79,82]. These related findings were (partly) reflected by our results, where novices' and experts' ratings were similar for seven of eleven TEAM items and the sum score, whereas for four TEAM items and the GRS, novices rated more leniently than experts. This could be explained by a lower standard, which novice raters used, possibly because of their lack of experience in (difficult) emergency cases. Our study differed from previous studies, in that it a) *compared* novices and experts to analyse *teamwork* and/or b) used the terms 'novice' and 'expert' to described their *experience regarding teamwork trainings and emergency medicine*, not their experience regarding the rating tool [83].

Another interesting aspect of study 2 is the result of the mixed effect model showing the interaction of teams and cases as the biggest source of variance, which implies that the teams' teamwork depended very much on the case and its specific demands. These findings can be connected to results from research about content specificity, showing that, for instance, clinical competence cannot be solely explained by one general ability like problem-solving, but depends heavily on the subject matter as well [84,85]. They also endorsed previous findings, suggesting that teamwork strategies need to fit the specific situation and that different leadership styles are beneficial in different situations [43]. Finally, the study showed that TEAM scores are case-specific and should not be compared across cases. Further research is therefore needed to set benchmarks for TEAM scores in different cases [71] and to develop a metric to characterise similar cases.

Relationship of external observations of teamwork, objective performance measures and teams' monitoring of teamwork

The lack of studies comparing observed teamwork (e.g. with the help of TEAM) and objective performance measures was one of the aspects addressed in study 3. Moreover, this study examined the relationship of TEAM ratings, performance measures and team members' and leaders' judgments of their teamwork to find out, whether they are able to accurately monitor their teamwork. In the SBT that we conducted, teams of physicians and nurses had to deal with a hypothermic cardiac arrest, a rather rare cause of cardiac arrest, which calls for a different treatment algorithm than usual and was therefore selected for the simulation training. The marked variety in team performance can be attributed to this rather difficult case. The results showed significant correlations between TEAM scores and objective measures: teams receiving higher ratings for teamwork were also quicker to request the first defibrillation and the first administration of adrenaline. These results represent a further validation of TEAM. Moreover, they confirm findings of other authors that teamwork (non-technical performance) and taskwork (technical performance) are connected [6,37,42]. This is, of course, highly relevant for clinical practice and emphasises the need to train teamwork. Since all these studies investigated the relationship of NTS and TS in resuscitation scenarios, future studies should explore whether similar results can be found for other settings where teamwork is necessary.

Team members' monitoring of their teamwork was moderately to highly correlated to the TEAM ratings and the objective measures. These findings underscore the hypothesis that individuals are capable of monitoring their performance while acting and – as in this case – shortly after as well. This is reflected by the literature on self-monitoring of one's own actions [63,86] and extends this concept to assessing tasks that have just been finished. Our results also show that team members were able to monitor the *teamwork* accurately and not just tasks solely completed by themselves. Team leaders' team-monitoring, on the contrary, was not connected to the observers' TEAM ratings or objective measures. One explanation for this finding could be that leading a team, which is a complex task per se, was even more demanding for the team leaders in our SBT due to the unfamiliar resuscitation scenario. As a result of this stressful situation, they might not have had the resources for the metacognitive task of monitoring the teamwork [87,88]. In addition, a possible diffusion of responsibilities among the team members might have led to more tasks and more stress for the team leader and consequently, even more difficulties in monitoring overall teamwork quality.

Limitations

There are certain limitations concerning the three described studies. Study 1 took place in a nightshift SBT, which allowed for only a small participant number (32 medical students in 7 teams). Furthermore – for educational reasons – participants frequently changed their roles between team leader, team member and observers during the night. This changing team structure can weaken the effect of the training and – as mentioned before – could have hindered the detection of different learning outcomes between the two debriefing approaches used in study 1.

Another limiting factor is the missing evaluation of the debriefing quality in study 1 with an instrument like the Objective Structured Assessment of Debriefing (OSAD) [89]. As described, all facilitators were trained in conducting a debriefing, but their actual adherence to the debriefing guidelines was not assessed due to limited personal resources – except for the debriefings' topics, which the facilitators reported themselves and the participants' high satisfaction with the debriefing, which can at least be seen as a form of indirect quality control. Nonetheless, further research into debriefing, different approaches and their effects should include an objective and standardised quality control.

In study 2, ratings were conducted by six expert and six novice raters, and the category 'expert' included both people with expertise in teamwork and NTS in terms of research and facilitating SBTs and people with expertise in emergency medicine from working in clinical practice. A study comparing emergency medical experts (who are not teamwork experts) and teamwork experts (who are not emergency medical experts) and novices in both areas is presently being conducted by our research group to fine-tune the findings regarding who can use TEAM as a tool to provide accurate ratings of teamwork.

Study 3 included 115 participants in 22 teams, therefore 'only' 22 team leaders. This makes it possible that a significant correlation between their team monitoring and the TEAM ratings could not be detected because of the sample size. Furthermore, this study was conducted in a real-world environment and included teams of different sizes and composition, to imitate real emergencies and teamwork of ad hoc teams as closely as possible. Nonetheless, as we did not control for these differences across the teams, these may also have affected the results. Our results showed that indeed teams with younger team leaders received higher TEAM scores, but no other team characteristic was associated with the observers' ratings. This correlation might have been the effect of more recent resuscitation training of the younger physicians and/or more familiarity with simulation trainings.

All three studies were single centre studies. As the institutional safety culture has an important influence on healthcare workers, similar studies need to be carried out at other hospitals and medical faculties to explore the generalisability of our results.

Conclusion

In conclusion, this dissertation provides several new components to the research and practice of training, debriefing, and assessing teamwork in healthcare. The use of debriefing scripts was examined and discussed, showing the benefits for novice facilitators. Still, more research is needed to determine the optimal debriefing approaches for novice as well as expert facilitators and different debriefing settings. In this process, debriefing scripts – and ways to support facilitators in general – are a key aspect to focus on to ensure debriefing quality.

This dissertation also provides a new assessment tool for NTS in medical emergencies for German-speaking countries and showed that both novices and experts can rate teamwork quality accurately. This finding should be replicated in real-life emergencies in the ER and it should be tested, how assessments (from novices) can be best incorporated into clinical practice as a tool for constant quality control and improvement.

Lastly, this thesis focused on teams' ability to monitor their own teamwork accurately. Team members were able to do so, and thus can provide valuable insights during team debriefing. These findings strongly suggest that team members should be included in the decision as whether to conduct a debriefing after an emergency. Team leaders' ability to monitor teamwork needs to be further investigated, as it is an important part of their leadership role.

1.7 References

1. Flin R, O'Connor P, Crichton M. Safety at the Sharp End. A Guide to Non-Technical Skills. Aldershot: Ashgate; 2008.
2. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. BMJ. 2016;353:i2139.
3. Lingard L, Espin S, Whyte S, Regehr G, Baker GR, Reznick R, Bohnen J, Orser B, Doran D, Grober E. Communication failures in the operating room: an observational classification of recurrent types and effects. Qual Saf Health Care. 2004;13(5):330-4.
4. Gawande AA, Zinner MJ, Studdert DM, Brennan TA. Analysis of errors reported by surgeons at three teaching hospitals. Surgery. 2003;133(6):614-21.
5. Wiegmann DA, ElBardissi AW, Dearani JA, Daly RC, Sundt TM. Disruptions in surgical flow and their relationship to surgical errors: An exploratory investigation. Surgery. 2007;142(5):658-65.
6. Hunziker S, Johansson AC, Tschan F, Semmer NK, Rock L, Howell MD, Marsch S. Teamwork and leadership in cardiopulmonary resuscitation. J Am Coll Cardiol. 2011;57(24):2381-8.
7. The Joint Commission. Most Commonly Reviewed Sentinel Event Types. 2019.
8. Tschan F, Semmer NK, Gautschi D, Hunziker P, Spychiger M, Marsch SU. Leading to recovery: group performance and coordinative activities in medical emergency driven groups. Hum Perform. 2006;19(3):277-304.
9. Fernandez R, Kozlowski SW, Shapiro MJ, Salas E. Toward a definition of teamwork in emergency medicine. Acad Emerg Med. 2008;15(11):1104-12.
10. Baker DP, Gustafson S, Beaubien J, Salas E, Barach P. Medical Teamwork and Patient Safety: The Evidence-based Relation. Literature Review. Rockville, MD: Agency for Healthcare Research and Quality; 2005. Available from: http://archive.ahrq.gov/research/findings/final-reports/medteam/index.html.
11. Sundstrom E, De Meuse KP, Futrell D. Work teams: Applications and effectiveness. American Psychologist. 1990;45(2):120-33.
12. PSNet Patient Safety Network. Glossary. Available from: https://psnet.ahrq.gov/glossary#.
13. Hautz WE, Kämmer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. Jama. 2015;313(3):303-4.
14. Grober ED, Bohnen JMA. Defining medical error. Canadian journal of surgery Journal canadien de chirurgie. 2005;48(1):39-44.
15. Kohn LT, Corrigan JM, Donaldson MS, (Institute of Medicine). To err is human. Washington, D.C.: National Academy Press; 2000.
16. Schrappe M. APS-Weißbuch Patientensicherheit. Berlin: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft; 2018.
17. Fung L, Boet S, Bould MD, Qosa H, Perrier L, Tricco A, Tavares W, Reeves S. Impact of crisis resource management simulation-based training for interprofessional and interdisciplinary teams: A systematic review. J Interprof Care. 2015;29(5):433-44.
18. Schmutz JB, Meier LL, Manser T. How effective is teamwork really? The relationship between teamwork and performance in healthcare teams: a systematic review and meta-analysis. BMJ Open. 2019;9(9):e028280.
19. Salas E, Cooke NJ, Rosen MA. On Teams, Teamwork, and Team Performance: Discoveries and Developments. Human Factors. 2008;50(3):540-7.
20. Marks MA, Mathieu JE, Zaccaro SJ. A Temporally Based Framework and Taxonomy of Team Processes. The Academy of Management Review. 2001;26(3):356-76.
21. Herzberg S, Hansen M, Schoonover A, Skarica B, McNulty J, Harrod T, Snowden JM, Lambert W, Guise JM. Association between measured teamwork and medical errors: an observational study of prehospital care in the USA. BMJ Open. 2019;9(10):e025314.

22. Helmreich RL, Schaefer H-G. Team performance in the operating room. In: Bogner MS, editor. Human error in medicine. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.; 1994. p. 225-54.
23. Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH. Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. Aviation, space, and environmental medicine. 1992;63(9):763.
24. Gordon S, Mendenhall P, O'Connor BB. Beyond the Checklist: What Else Health Care Can Learn from Aviation Teamwork and Safety. Ithaca, NY: Cornell University Press; 2012.
25. Fletcher G, Flin R, McGeorge P, Glavin RJ, Maran NJ, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. British Journal of Anaesthesia. 2003;90(5):580-8.
26. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. Medical Education. 2006;40(11):1098-104.
27. Rall M, Gaba DM. Human Performance and Patient Safety. In: Miller R, editor. Miller´s Anesthesia. 6 ed. Philadelphia: Elsevier Churchill Livingstone; 2005. p. 3021-72.
28. Gaba DM. The future vision of simulation in health care. Quality and Safety in Health Care. 2004;13(suppl 1):i2-i10.
29. Kolb DA. Experiential Learning: Experience as the Source of Learning and Development. 2 ed. Upper Saddle River, NJ: Pearson Education; 2014.
30. Kneebone R. Evaluating Clinical Simulations for Learning Procedural Skills: A Theory-Based Approach. Academic Medicine. 2005;80(6).
31. Lave J, Wenger E. Situated learning: Legitimate peripheral participation: Cambridge University Press; 1991.
32. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. Jama. 2011;306(9):978-88.
33. Ilgen JS, Sherbino J, Cook DA. Technology-enhanced simulation in emergency medicine: a systematic review and meta-analysis. Acad Emerg Med. 2013;20(2):117-27.
34. O'Dea A, O'Connor P, Keogh I. A meta-analysis of the effectiveness of crew resource management training in acute care domains. Postgrad Med J. 2014;90(1070):699-708.
35. Buljac-Samardzic M, Doekhie KD, van Wijngaarden JDH. Interventions to improve team effectiveness within health care: a systematic review of the past decade. Hum Resour Health. 2020;18(1):2.
36. Hayes CW, Rhee A, Detsky ME, Leblanc VR, Wax RS. Residents feel unprepared and unsupervised as leaders of cardiac arrest teams in teaching hospitals: a survey of internal medicine residents. Crit Care Med. 2007;35(7):1668-72.
37. Krage R, Zwaan L, Tjon Soei Len L, Kolenbrander MW, van Groeningen D, Loer SA, Wagner C, Schober P. Relationship between non-technical skills and technical performance during cardiopulmonary resuscitation: does stress have an influence? Emerg Med J. 2017;34(11):728-33.
38. Kiesewetter J, Gutmann J, Drossard S, Gurrea Salas D, Prodinger W, Mc Dermott F, Urban B, Staender S, Baschnegger H, Hoffmann G, Hubsch G, Scholz C, Meier A, Wegscheider M, Hoffmann N, Ohlenbusch-Harke T, Keil S, Schirlo C, Kuhne-Eversmann L, Heitzmann N, Busemann A, Koechel A, Manser T, Welbergen L, Kiesewetter I. The Learning Objective Catalogue for Patient Safety in Undergraduate Medical Education - A Position Statement of the Committee for Patient Safety and Error Management of the German Association for Medical Education. GMS J Med Educ. 2016;33(1):Doc10.
39. Ross AJ, Kodate N, Anderson JE, Thomas L, Jaye P. Review of simulation studies in anaesthesia journals, 2001-2010: mapping and content analysis. British Journal of Anaesthesia. 2012;109(1):99-109.
40. Ochsmann EB, Zier U, Drexler H, Schmid K. Well prepared for work? Junior doctors' self-assessment after medical education. BMC Med Educ. 2011;11:99.
41. Tallentire VR, Smith SE, Wylde K, Cameron HS. Are medical graduates ready to face the challenges of Foundation training? Postgrad Med J. 2011;87(1031):590.

42. Peltonen V, Peltonen LM, Salantera S, Hoppu S, Elomaa J, Pappila T, Hevonoja E, Hurme S, Perkonoja K, Elomaa T, Tommila M. An observational study of technical and non-technical skills in advanced life support in the clinical setting. Resuscitation. 2020;153:162-8.
43. Hargestam M, Lindkvist M, Jacobsson M, Brulin C, Hultin M. Trauma teams and time to early management during in situ trauma team training. BMJ Open. 2016;6(1):e009911.
44. Kolbe M, Burtscher M, Manser T, Künzle B, Grote G. The Role of Coordination in Preventing Harm in Healthcare Groups: Research Examples from Anaesthesia and an Integrated Model of Coordination for Action Teams in Health Care. In: Boos M, Kolbe M, Kappeler PM, Ellwart T, editors. Coordination in Human and Primate Groups. Heidelberg: Springer; 2011. p. 75-92.
45. Rudolph JW, Simon R, Raemer DB, Eppich WJ. Debriefing as formative assessment: closing performance gaps in medical education. Acad Emerg Med. 2008;15(11):1010-6.
46. Phrampus PE, O'Donnell JM. Debriefing Using a Structured and Supported Approach. In: Levine AL, DeMaria SJ, Schwartz AD, Sim AJ, editors. The Comprehensive Textbook of Healthcare Simulation. New York: Springer; 2013. p. 73–84.
47. Kolbe M, Weiss M, Grote G, Knauth A, Dambach M, Spahn DR, Grande B. TeamGAINS: a tool for structured debriefings for simulation-based team trainings. BMJ Qual Saf. 2013;22(7):541-53.
48. Sawyer T, Eppich W, Brett-Fleegler M, Grant V, Cheng A. More than one way to debrief: a critical review of healthcare simulation debriefing methods. Simul Health. 2016;11(3):209-17.
49. Greif R, Lockey AS, Conaghan P, Lippert A, De Vries W, Monsieurs KG. European Resuscitation Council Guidelines for Resuscitation 2015: Section 10. Education and implementation of resuscitation. Resuscitation. 2015;95:288-301.
50. Morrison Laurie J, Neumar Robert W, Zimmerman Janice L, Link Mark S, Newby LK, McMullan Paul W, Hoek Terry V, Halverson Colleen C, Doering L, Peberdy Mary A, Edelson Dana P. Strategies for Improving Survival After In-Hospital Cardiac Arrest in the United States: 2013 Consensus Recommendations. Circulation. 2013;127(14):1538-63.
51. Stanford Medicine. What are Cognitive aids? 2020. Available from: http://med.stanford.edu/cogaids/whatarecogaids.html.
52. Cheng A, Hunt EA, Donoghue A, Nelson-McMillan K, Nishisaki A, Leflore J, Eppich W, Moyer M, Brett-Fleegler M, Kleinman M, Anderson J, Adler M, Braga M, Kost S, Stryjewski G, Min S, Podraza J, Lopreiato J, Hamilton MF, Stone K, Reid J, Hopkins J, Manos J, Duff J, Richard M, Nadkarni VM, Investigators E. Examining pediatric resuscitation education using simulation and scripted debriefing: a multicenter randomized trial. JAMA Pediatr. 2013;167(6):528-36.
53. Freytag J, Stroben F, Hautz WE, Eisenmann D, Kammer JE. Improving patient safety through better teamwork: how effective are different methods of simulation debriefing? Protocol for a pragmatic, prospective and randomised study. BMJ Open. 2017;7(6):e015977.
54. Stroben F, Schröder T, Dannenberg KA, Thomas A, Exadaktylos A, Hautz WE. A simulated night shift in the emergency room increases students' self-efficacy independent of role taking over during simulation. BMC Med Educ. 2016;16:177.
55. Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, Nestel D. Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). Resuscitation. 2010;81(4):446-52.
56. Cooper S, Cant R, Connell C, Sims L, Porter J, Symmons M, Nestel D, Liaw S. Measuring teamwork performance: validity testing of the Team Emergency Assessment Measure (TEAM) with clinical resuscitation teams. Rescucitation. 2016;101:97-101.
57. Couto TB, Kerrey BT, Taylor RG, FitzGerald M, Geis GL. Teamwork skills in actual, in situ, and in-center pediatric emergencies: performance levels across settings and perceptions of comparative educational impact. Simul Healthc. 2015;10(2):76-84.
58. Freytag J, Stroben F, Hautz WE, Schauber SK, Kämmer JE. Rating the quality of teamwork—a comparison of novice and expert ratings using the Team Emergency

Assessment Measure (TEAM) in simulated emergencies. Scand J Trauma Resusc Emerg Med. 2019;27(1):12.
59. Schmutz JB, Eppich WJ. When I say ... team reflexivity. Med Educ. 2019;53(6):545-6.
60. Boet S, Bould M, Sharma B, Revees S, Naik V, Triby E, Grantcharov T. Within-team debriefing versus instructor-led debriefing for simulation-based education: a randomized controlled trial. Ann Surg. 2013;258(1):53-8.
61. Mullan PC, Wuestner E, Kerr TD, Christopher DP, Patel B. Implementation of an in situ qualitative debriefing tool for resuscitations. Resuscitation. 2013;84(7):946-51.
62. Eva KW, Regehr G. Exploring the divergence between self-assessment and self-monitoring. Advances in Health Sciences Education. 2011;16(3):311-29.
63. McConnell MM, Regehr G, Wood TJ, Eva KW. Self-monitoring and its relationship to medical knowledge. Advances in Health Sciences Education. 2012;17(3):311-23.
64. Hautz SC, Oberholzer DL, Freytag J, Exadaktylos A, Kammer JE, Sauter TC, Hautz WE. An observational study of self-monitoring in ad hoc health care teams. BMC Med Educ. 2020;20(1):201.
65. Stroben F. Sicherheit und Selbstwirksamkeit junger Ärzte – Beitrag des Medizinstudiums. Berlin: Charité Universitätsmedizin Berlin; 2019.
66. Freytag J, Stroben F. Implementing Crisis Resource Management (CRM) principles into undergraduate education using a peer-teaching approach. Poster presented at: AMEE 2016 Proceedings of the AMEE annual conference; 2016 Aug 27-31; Barcelona, Spain. Dundee, UK: AMEE; 2016.
67. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.
68. Eppich W, Nannicelli AP, Seivert NP, Sohn MW, Rozenfeld R, Woods DM, Holl JL. A rater training protocol to assess team performance. J Contin Educ Health Prof. 2015;35(2):83-90.
69. Bogossian F, Cooper S, Cant R, Beauchamp A, Porter J, Kain V, Bucknall T, Phillips NM. Undergraduate nursing students' performance in recognising and responding to sudden patient deterioration in high psychological fidelity simulated environments: an Australian multi-centre study. Nurse Educ Today. 2014;34(5):691-6.
70. Cant RP, Porter JE, Cooper SJ, Roberts K, Wilson I, Gartside C. Improving the non-technical skills of hospital medical emergency teams: the Team Emergency Assessment Measure (TEAM). Emerg Med Australas. 2016;28(6):641-6.
71. Maignan M, Koch F-X, Chaix J, Phellouzat P, Binauld G, Collomb Muret R, Cooper SJ, Labarère J, Danel V, Viglino D, Debaty G. Team Emergency Assessment Measure (TEAM) for the assessment of non-technical skills during resuscitation: validation of the French version. Resuscitation. 2016;101:115-20.
72. Valentine MA, Nembhard IM, Edmondson AC. Measuring teamwork in health care settings: a review of survey instruments. Med Care. 2015;53(4):e16-30.
73. Dorer B. Round 6 translation guidelines. Mannheim: European Social Survey, GESIS; 2012.
74. Freytag J, Stroben F, Hautz WE, Eisenmann D, Kämmer JE. How to debrief non-technical skills? – Comparison of two debriefing methods during a large emergency simulation for undergraduates. Paper presented at: 13th International Skills Lab Symposium; 2018 Mar 23-24; Maribor, Slovenia.
75. Cheng A, Morse KJ, Rudolph J, Arab AA, Runnacles J, Eppich W. Learner-Centered Debriefing for Health Care Simulation Education: Lessons for Faculty Development. Simul Healthc. 2016;11(1):32-40.
76. Cheng A, Eppich W, Kolbe M, Meguerdichian M, Bajaj K, Grant V. A Conceptual Framework for the Development of Debriefing Skills: A Journey of Discovery, Growth, and Maturity. Simul Healthc. 2020;15(1):55-60.
77. Tannenbaum SI, Cerasoli CP. Do Team and Individual Debriefs Enhance Performance? A Meta-Analysis. Human Factors. 2013;55(1):231-45.
78. Krogh K, Bearman M, Nestel D. “Thinking on your feet”—a qualitative study of debriefing practice. Advances in Simulation. 2016;1:12.

79. Read EK, Bell C, Rhind S, Hecker KG. The use of global rating scales for OSCEs in veterinary medicine. PLoS One. 2015;10(3):e0121000.
80. Sawyer JM, Anton NE, Korndorffer JR, Jr., DuCoin CG, Stefanidis D. Time crunch: increasing the efficiency of assessment of technical surgical skill via brief video clips. Surgery. 2018;163(4):933-7.
81. Pradarelli JC, George E, Kavanagh J, Sonnay Y, Khoon TH, Havens JM. Training Novice Raters to Assess Nontechnical Skills of Operating Room Teams. Journal of Surgical Education. 2020.
82. Khan R, Payne MWC, Chahine S. Peer assessment in the objective structured clinical examination: A scoping review. Med Teach. 2017;39(7):745-56.
83. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A, Vincent CA. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. Ann Surg. 2009;249(6):1047-51.
84. Eva KW. On the generality of specificity. Med Educ. 2003;37(7):587-8.
85. Wimmers PF, Splinter TAW, Hancock GR, Schmidt HG. Clinical Competence: General Ability or Case-specific? Advances in Health Sciences Education. 2007;12(3):299-314.
86. Kämmer JE, Hautz WE, März M. Self-monitoring accuracy does not increase throughout undergraduate medical education. Medical Education. 2020;54(4):320-7.
87. Driskell JE, Salas E, Johnston J. Does stress lead to a loss of team perspective? Group Dynamics: Theory, Research, and Practice. 1999;3(4):291-302.
88. Norris EM, Lockey AS. Human factors in resuscitation teaching. Resuscitation. 2012;83(4):423-7.
89. Runnacles J, Thomas L, Sevdalis N, Kneebone R, Arora S. Development of a tool to improve performance debriefing and learning: the paediatric Objective Structured Assessment of Debriefing (OSAD) tool. Postgrad Med J. 2014;90(1069):613-21.

2 Affidavit

‘I, Julia Freytag, certify under penalty of perjury by my own signature that I have submitted the dissertation on the topic ‘Assessment, training and debriefing of team performance in medical ad hoc teams – a simulation-based approach’ [‘Assessment, Training und Debriefing der Leistung von medizinischen ad hoc Teams’]. I wrote this thesis independently and without assistance from third parties, I used no other aids than the listed sources and resources.

All parts, based literally or in spirit on publications or presentations of other authors, are, as such, in proper citations indicated. The section on methodology (in particular practical work, laboratory requirements, statistical processing) and results (in particular images, graphics and tables) are my responsibility.

My contribution in the selected publications for this dissertation corresponds to those that are specified in the following joint declaration with my supervisor. All the publications of this dissertation comply with the ICMJE recommendations (International Committee of Medical Journal Editors; www.icmje.og) on authorship. Furthermore, I declare that I agree to comply with the Statute for Ensuring Good Scientific Practice of Charité – Universitätsmedizin Berlin.

I certify that I did not submit this dissertation in identical or similar form to another faculty.

I know and I am aware of the importance of this affidavit and the criminal consequences of a false affidavit (section 156,161 of the Criminal Code).’

Date Signature

3 Share declaration

I, Julia Freytag, had the following share in the following publications:

Publication 1

Freytag J*, Stroben F*, Hautz WE, Eisenmann D, Kämmer JE. Improving patient safety through better teamwork: How effective are different methods of simulation debriefing? Protocol for a pragmatic, prospective and randomized study. BMJ Open 2017, 7:e015977. doi: 10.1136/ bmjopen-2017-015977

IF = 2.562

**Shared authorship*

Contribution: Conceptualization and design of study with FS. Translation of the 'Teamwork Emergency Assessment Measure' (TEAM) together with FS, resulting in the German TEAM, which can be found in Supplementary Material of this publication. Planning and carrying out of pre-study with FS, including the development of the TeamTAG (Supplementary Material). Data analysis and interpretation with FS. Preparation of Figure 1. Drafting and writing the manuscript with FS and JEK. Revision work based on the comments of the reviewers.

Publication 2

Freytag J, Stroben F, Hautz WE, Schauber, SK, Kämmer JE. Rating the quality of teamwork—a comparison of novice and expert ratings using the Team Emergency Assessment Measure (TEAM) in simulated emergencies. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 2019, 27(12). doi: 10.1186/s13049-019-0591-9

IF = 2.036

Contribution: Conceptualization and design of study with FS, WEH and JEK. Literature research (resulting in Additional File 1). Responsibility for data analysis and interpretation, resulting in Table 1 and 2, Figure 1 and Additional Files 3 and 4. Drafting and writing the manuscript with FS. Drafting of the cover letter accompanying the manuscript. Drafting of manuscript revision including a point-by-point response to reviewer feedback. Revision work based on the comments of the reviewers. Composition of final revised manuscript together with all co-authors.

Publication 3

Hautz C., Oberholzer DL, Freytag J, Exadaktylos A, Kämmer JE, Sauter TC, Hautz WE. An observational study of self-monitoring in ad hoc health care teams. BMC Med Educ 2020, 20, 201. doi: 10.1186/s12909-020-02115-3

IF = 1.870

Contribution: Data collection: Responsibility for rating all video recordings using the TEAM instrument. Data analysis together with WEH and JEK and interpretation with all authors. Preparation of Table 1 and 2. Reviewing the manuscript and providing revisions. Supporting the revision work based on the comments of the reviewers, including a point-by-point response to reviewer feedback.

Signature of the doctoral candidate

4 Peer-reviewed articles

Publication 1, pp. 39 - 47

Freytag J*, Stroben F*, Hautz WE, Eisenmann D, Kämmer JE. Improving patient safety through better teamwork: How effective are different methods of simulation debriefing? Protocol for a pragmatic, prospective and randomized study. BMJ Open 2017, 7:e015977. doi: 10.1136/bmjopen-2017-015977

IF = 2.562

**Shared authorship*

Publication 2, pp. 48-55

Freytag J, Stroben F, Hautz WE, Schauber, SK, Kämmer JE. Rating the quality of teamwork—a comparison of novice and expert ratings using the Team Emergency Assessment Measure (TEAM) in simulated emergencies. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 2019, 27(12). doi: 10.1186/s13049-019-0591-9

IF = 2.036

Publication 3, pp. 56-64

Hautz C., Oberholzer DL, Freytag J, Exadaktylos A, Kämmer JE, Sauter TC, Hautz WE. An observational study of self-monitoring in ad hoc health care teams. BMC Med Educ 2020, 20, 201. doi: 10.1186/s12909-020-02115-3

IF = 1.870

BMJ Open

Improving patient safety through better teamwork: how effective are different methods of simulation debriefing? Protocol for a pragmatic, prospective and randomised study

Julia Freytag,[1] Fabian Stroben,[2,3] Wolf E Hautz,[3] Dorothea Eisenmann,[2,4] Juliane E Kämmer[5,6]

To cite: Freytag J, Stroben F, Hautz WE, *et al*. Improving patient safety through better teamwork: how effective are different methods of simulation debriefing? Protocol for a pragmatic, prospective and randomised study. *BMJ Open* 2017;**7**:e015977. doi:10.1136/bmjopen-2017-015977

► Prepublication history and additional material for this paper are available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-015977).

JF and FS contributed equally.

Received 16 January 2017
Revised 2 May 2017
Accepted 18 May 2017

For numbered affiliations see end of article.

Correspondence to
Fabian Stroben; fabian.stroben@charite.de

ABSTRACT

Introduction Medical errors have an incidence of 9% and may lead to worse patient outcome. Teamwork training has the capacity to significantly reduce medical errors and therefore improve patient outcome. One common framework for teamwork training is crisis resource management, adapted from aviation and usually trained in simulation settings. Debriefing after simulation is thought to be crucial to learning teamwork-related concepts and behaviours but it remains unclear how best to debrief these aspects. Furthermore, teamwork-training sessions and studies examining education effects on undergraduates are rare. The study aims to evaluate the effects of two teamwork-focused debriefings on team performance after an extensive medical student teamwork training.

Methods and analyses A prospective experimental study has been designed to compare a well-established three-phase debriefing method (gather–analyse–summarise; the *GAS method*) to a newly developed and more structured debriefing approach that extends the GAS method with *TeamTAG* (teamwork techniques analysis grid). TeamTAG is a cognitive aid listing preselected teamwork principles and descriptions of behavioural anchors that serve as observable patterns of teamwork and is supposed to help structure teamwork-focused debriefing. Both debriefing methods will be tested during an emergency room teamwork-training simulation comprising six emergency medicine cases faced by 35 final-year medical students in teams of five. Teams will be randomised into the two debriefing conditions. Team performance during simulation and the number of principles discussed during debriefing will be evaluated. Learning opportunities, helpfulness and feasibility will be rated by participants and instructors. Analyses will include descriptive, inferential and explorative statistics.

Ethics and dissemination The study protocol was approved by the institutional office for data protection and the ethics committee of Charité Medical School Berlin and registered under EA2/172/16. All students will participate voluntarily and will sign an informed consent after receiving written and oral information about the study. Results will be published.

Strengths and limitations of this study

- The study design builds on established principles of teaching and assessing teamwork.
- The study will be one of the first to explore the effects of teamwork-focused debriefing on team performance with undergraduate medical students.
- The study will be embedded in a well-established simulation setting with proven efficacy.
- The study will be a pragmatic, randomised comparison of two debriefing methods.
- Only a single centre will be studied.
- Feedback quality will not be externally evaluated.

INTRODUCTION

Medical errors and adverse events occur with an incidence of about 9% and can seriously harm patients.[1 2] Error rates in emergency settings are even reported to be twice as high.[3–5] Most medical errors originate from human factors and teamwork[6] or medication errors[7] and about half of all medical errors are considered preventable.[1 7]

Empirical evidence[6 8–11] suggests that improving teamwork may be key to reducing medical error. Yet, although teamwork and patient safety are prominent objectives in many national outcome frameworks,[12–14] these topics are insufficiently represented in undergraduate education and are rarely assessed, even though validated teamwork assessment tools exist.[15 16] Consequently, about 60% of junior doctors in Germany reported feeling inadequately prepared for clinical practice[17] and almost half of the residents in a Canadian survey reported feeling overwhelmed when leading a resuscitation team.[18]

In addition, common interventions targeting the quality of teamwork and human factors, such as simulation training and crisis resource management (CRM) training, have produced a variety of effects.[19 20] In both simulation and CRM training, debriefing is considered crucial to enhancing learning[21] but little is known about how best to debrief. In fact, the widely differing effects of simulation may very well result from differences in debriefing. A feasible and beneficial debriefing method, particularly for undergraduates, could lead to more effective simulation sessions and thus ease the transition into clinical practice for junior doctors. This could ultimately lead to a reduction of medical errors and thus improved patient outcome. In this study we will compare the effects of two different debriefing methods on team performance and the acquisition of teamwork skills during teamwork simulations for medical students.

Training and debriefing

The concept of CRM was originally derived from safety training in aviation and has been adapted to the healthcare sector, another high-stakes environment.[22] The idea of CRM is to guide individuals and teams in emergency situations (crises), encouraging them to use all available resources to manage the situation effectively and prevent critical incidents from occurring in the first place. CRM training has been shown to be a potent tool to improve teamwork and—as a consequence—patient safety.[23–25] In our study, elements of CRM set the framework for teamwork training and debriefing during an emergency room simulation.

Simulation debriefing is defined as a bidirectional and interactive discussion after a simulation in which participants reflect on their actions and analyse their performance.[21] Feedback is a central process element of debriefing that is often used as a conversational technique especially in participants with little experience in debriefing.[26] Feedback is defined as the delivery of information to improve reasoning or behaviour compared with defined performance standards,[26 27] and it is critical in improving learning.[21] How best to integrate feedback into debriefing, what specific aspects to address and how to structure debriefing to foster learning are, however, still unknown.[21 28] The goal of this study is thus to evaluate the potential benefit of preselecting certain aspects to be discussed during debriefing and of structuring debriefing with the help of a cognitive aid. To this end, we will compare a well-established debriefing method to a more structured and feedback-focused method to evaluate their effects on teamwork, learning opportunities, feasibility and helpfulness for participants (and instructors). We will focus on two debriefing methods, the *gather–analyse–summarise* (GAS) method and the GAS method plus a cognitive aid:

1. The GAS method: This debriefing method consists of three parts: gathering, analysing and summarising.[29 30] The GAS method is one of many similar three-step debriefing structures[26] and has been used, for example, in simulation courses run by the American Heart Association.[30] During the first phase (gather), participants are given the opportunity to report their thoughts on the simulated situation. They are encouraged to exchange their views on what actually happened to establish a shared mental model of the situation. This model can afterwards be used to discuss the simulation in a learner-centred way (analyse). During this process, questions tailored towards specific learning objectives are used to facilitate participants' reflection on and analysis of their actions and induce learning. Finally, the debriefing is summed up and critically reviewed by the team and its instructor (summarise).[26 29] Topics discussed during the debriefing using this method are mostly self-selected by the team and instructor, which makes this method highly flexible. A possible drawback with regard to teamwork (or any other specific learning objective) is that its potential to enhance the quality of teamwork is influenced by the instructor's level of experience.[26] A typical question to start the debriefing with the gather step might be 'How do you feel now?' followed in the analysis step by 'What worked well?' or 'Do you see any opportunities for improvement?' The summarise step might be initiated by 'What we learned from this session….'
2. The GAS method plus a cognitive aid: This newly developed debriefing method uses the GAS structure detailed above and additionally provides the instructors with a cognitive aid to structure the debriefing in more detail. It further provides a selection of important aspects to address during debriefing. Cognitive aids are 'structured pieces of information designed to enhance cognition and adherence to…best practices.'[31] Cognitive aids have been shown to be beneficial in different areas of medicine.[32–34] Moreover, cognitive aids are useful for debriefing: Instructors' use of a cognitive aid may improve participants' acquisition of behavioural and cognitive outcomes after simulation—especially so with novice instructors.[35] In practice, such aids are often a pocket card, script or poster.

We will use a specific cognitive aid called 'TeamTAG' (teamwork techniques analysis grid) to foster observation and feedback relevant to teamwork. TeamTAG is a guideline for structuring the feedback process during debriefing and remembering what to address during the analysis step of the GAS method. The TeamTAG lists teamwork-relevant CRM principles together with descriptions of behavioural anchors that serve as directly observable patterns of teamwork and provides space for notes (see online supplementary information). The TeamTAG can be printed on a single sheet of paper (A4) and filled in during observation of the simulation. After the simulation, instructors have the flexibility to set priorities for debriefing based on their observations and structured notes. The debriefing itself will follow the same structure as under the GAS method. However, the TeamTAG might, for example, remind instructors that team leaders 'allocate roles & tasks' or are responsible for 'monitoring progress' (according to the CRM principle 'exercise leadership and followership'). These aspects might be specifically addressed by group instructors to improve group reflection during the analysis step.

Hypotheses

First, we assume that the GAS method plus TeamTAG will be a more effective debriefing tool than the common GAS method alone and will lead to the discussion of more teamwork-relevant principles. Debriefing using the GAS method plus TeamTAG should thus result in more learning opportunities for teams and ultimately in improved team performance. This hypothesis is based on the fact that the TeamTAG is concise and guides observation and feedback with practical examples. Using these examples during observation may help focus the observers' attention[36] and result in the team discussing more teamwork-relevant CRM principles. In undergraduate education, instructors are often novices and vary considerably regarding how experienced they are in debriefing. Because novices were shown to benefit more from structured debriefing scripts than more experienced instructors,[35] we consider our environment (see the Methods and analysis section) ideal for detecting differences between the two debriefing methods if they exist.

Hypothesis 1a: Participants who receive debriefing based on the GAS method plus TeamTAG will show a greater improvement in team performance than those who discuss the simulation according to the common GAS method alone.

Hypothesis 1b: Participants who receive debriefing based on the GAS method plus TeamTAG will report discussing a higher number of CRM principles than participants who are debriefed with the GAS method alone.

Second, we expect that teams receiving debriefing based on the GAS method plus TeamTAG will perceive teamwork skills as more important after the simulation event, which should increase their sensitivity to a culture of safety and the likelihood of changing their behaviour.[37 38] Moreover, perceiving the content of the debriefing as more important should lead to higher overall satisfaction with and perception of helpfulness of the debriefing.

Hypothesis 2a: Participants who receive debriefing based on the GAS method plus TeamTAG will report a higher level of perceived importance of teamwork principles than those who are debriefed according to the common GAS method.

Hypothesis 2b: Participants who receive debriefing based on the GAS method plus TeamTAG will report higher satisfaction with and helpfulness of the debriefing they received than those who are debriefed according to the GAS method alone.

Third, we will focus on the satisfaction of the instructors as a measure of feasibility and efficiency. We expect higher satisfaction when they use the GAS method plus TeamTAG as it might facilitate more structured feedback and it provides a better opportunity for instructors to address the learning objectives of their participants.

Hypothesis 3: Instructors who use the GAS method plus TeamTAG will report higher levels of feasibility and efficiency of their debriefing than instructors who use the GAS method alone.

METHODS AND ANALYSIS

This investigation is designed as a prospective experimental superiority study with intervention and control groups receiving debriefing during a simulation training based on either the GAS method plus TeamTAG or the GAS method alone, respectively. The study will be executed during an emergency department (ED) simulation at Charité Medical School, Berlin, Germany, on 14 January 2017. The ED simulation has been implemented at the local skills laboratory since 2013 on a peer-led basis. The main goal of this extensive, 8-hour night-shift simulation training is to give students the opportunity to experience being the person in charge of a patient's healthcare. This event takes place once a year, with about 35 students in their final year of medical studies participating voluntarily. Participants are recruited via newsletter and advertising posters. The students act in randomly assigned teams of five and self-select into different roles (team leader, team member, observer), which they switch during the night. Simulated patients and high-fidelity simulators are used to create realistic case simulations; simulated radiological and laboratory services are provided. One of the main goals of the event is to improve students' confidence in working with medical emergencies in an ED over the course of the night.[39] The simulation was awarded a project prize by the German Association for Medical Education in 2016.

Each student team has to work on six simulated cases. Each case is staffed with a case instructor who is responsible for the simulation and provides technical help. Each student team is accompanied by a group instructor who guides the participants during the night. After every case, multisource feedback is provided by simulated patients, observing participants and case instructors. As part of our study, in 2017 participants will additionally receive a teamwork-based debriefing by the group instructors after every case in one of two conditions (GAS method vs GAS method plus TeamTAG). Additionally, the quality of teamwork will be rated by trained raters throughout the night.

As group instructors we will choose experienced peer teachers who are advanced in their healthcare studies (medicine, nursing) and have completed emergency room courses/electives during their studies. Peer teachers at Charité Medical School Berlin frequently give courses in clinical skills training and simulator-based emergency medicine trainings for other medical students. All group instructors undergo extensive feedback training during their studies and are furthermore trained in working with and debriefing groups.

Development of the TeamTAG as cognitive aid

As a basis for this study, the TeamTAG guideline was developed with the goal of having a feasible and time-efficient feedback instrument that supports teaching basic teamwork skills to participants. Two investigators (JF and FS) developed the TeamTAG guidelines that present six common CRM principles,[22 40] each accompanied by the description of behavioural anchors. The six principles are (1) anticipate and plan ahead, (2) set priorities dynamically, (3) call for help early, (4) exercise leadership and followership, (5) communicate effectively and (6) re-evaluate repeatedly. The TeamTAG can be found in the online supplementary material. The CRM principles and their behavioural anchors were chosen to fit the following criteria: (A) simulation setting, (B) presumed skills of participants, (C) experience of instructors and (D) observability. The tool was reviewed and adjusted by an experienced group of anaesthesiologists, emergency medicine physicians, simulation instructors and peer tutors, all experienced in medical education and simulation-based learning. In a prestudy, feasibility for instructors was examined (see the Preliminary results section) but not compared with an approach without the TeamTAG.

Team performance measurement

To measure team performance, we will use the Team Emergency Assessment Measure (TEAM).[15] TEAM is an assessment tool that has been applied to both clinical and simulation environments.[15 16 41] It consists of 11 items belonging to the three subscales leadership, teamwork and task management. Example items are 'the team leader maintained a global perspective' and 'the team prioritized tasks', measured on a 5-point Likert scale of 0 (*never*) to 4 (*always*). Additionally, it includes an overall rating of team performance (range: 1 (*very poor performance*) to 10 (*very good performance*)).

As there was no German version of the TEAM, the English version was translated into German using elements of the TRAPD (translation, review, adjudication, pretest, documentation) methodology.[42] Two investigators (JF and FS) independently translated the TEAM into German in parallel, reviewed the results and consented to one version, which was translated back by a native English speaker. This new version was compared with the original TEAM and agreed to by both investigators and the native speaker. All steps of the translation were documented.

After the TEAM was translated, we developed a rater training. The training involves three aspects that are important in preparation for accurately assessing a certain behaviour or skill[1 43]: a *rater error training* in which information is provided on typical rating errors to raise awareness and prevent them,[2] a *performance dimension training* to teach raters about the targeted dimensions, including definitions and videotaped examples, and[3] a *frame-of-reference training*, in which videotaped examples showing teamwork of different levels of quality are assessed and discussed. All raters who will be responsible for TEAM ratings in this study (case instructors and additional raters) will receive this rater training and additional written material on teamwork and how to use the TEAM.

Group instructors debriefing training

Before data collection, all group instructors will receive a teamwork-related training and additional written material with information about how to provide feedback and conduct debriefings and about human factors in general and CRM in particular, which is intended to serve as a framework for discussing all teamwork aspects during debriefing. The training will include videos showing good and bad examples of teamwork and will be followed by discussions about opportunities for debriefing in these specific situations (adapted from *frame of reference training*[43]). After this training, which will be the same for all group instructors, the instructors will be randomly assigned, stratified by level of academic education and additional professional training (eg, nurse or paramedic), to the two conditions. The two groups will receive separate instruction from the investigators: The intervention group instructors will be told to discuss their groups' performance with the help of the TeamTAG and to focus on each CRM principle of the TeamTAG at least once during the first five cases (ie, one or two principles per case) so that by case 6 all CRM principles will have been debriefed and team performance during case 6 can be compared between conditions. Furthermore, they will be instructed to re-evaluate their previous focus of debriefing after each case if behaviour does not change sufficiently from their perspective. The order of chosen topics can be varied by the instructors and should be adjusted to observed difficulties in teamwork during the simulation. The control group instructors will be advised to give feedback regarding whatever teamwork-related aspect they deem important during the first five cases and also to re-evaluate the teamwork if needed. Instructors will stay with their groups during the whole simulation event to guarantee coordinated, consistent and longitudinal feedback.

Data collection

Upon arrival, every student participant will create an individual anonymised study code, which will be entered on every form and questionnaire and will allow us to link all measurements during the course of the night. Students will also track their role (leader, member, observer) after every case to allow subgroup analyses in relation to these roles. Figure 1 depicts the data collection procedure during the night-shift simulation.

Before starting the simulation, all 35 participants will be asked to fill in a first questionnaire that assesses possible confounders such as demographic data, professional training as a nurse or paramedic, or any training in teamwork/human factors. Next, students will be randomly assigned to seven groups via a computer-generated algorithm by the principal investigator. Four groups will serve as intervention groups and the remaining three

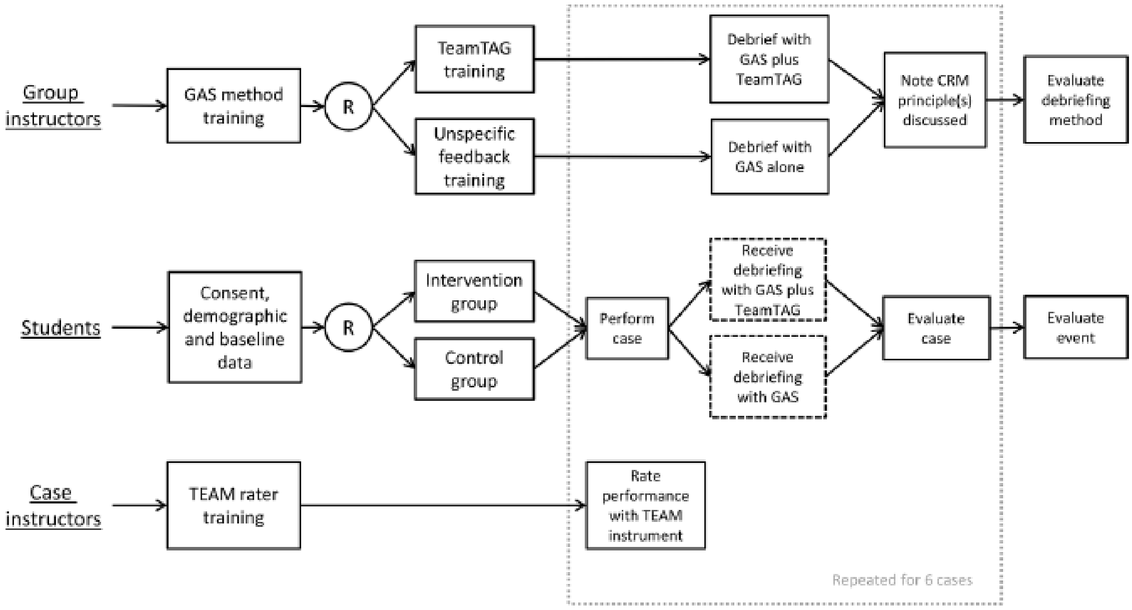


Figure 1 Study flow chart. CRM, crisis resource management; GAS, gather–analyse–summarise; R, randomisation; TEAM, Team Emergency Assessment Measure; TeamTAG, teamwork techniques analysis grid.

as controls; participants will not know to which condition they are assigned. After randomisation, all groups will gather separately and will be asked to discuss already known principles of teamwork and 15 multiple-choice questions concerning emergency medicine. A recent study showed that the results of such discussions are linked to team performance.[44]

During the simulation, all groups will face six simulations where teamwork will be measured and teamwork-related feedback provided. All cases depict common emergency situations where the participation of an emergency team in the emergency room is needed. Table 1 gives a brief overview of the diagnoses of the six cases and challenges for teamwork.

During every case, team performance will be measured using the TEAM,[16] which will be filled in by the case instructors and an additional rater. The two TEAM raters will be blind to the debriefing condition the group is assigned to.

After every case (duration about 30 min), debriefing will start (duration about 20 min) with checklist-based feedback from the simulated patients (focus: communication skills, empathy) and the case instructors and peer observers (focus: factual knowledge, diagnostic skills). As the last part of the debriefing process, the teamwork-related debriefing will be conducted by the group instructor using the GAS method with or without the support of the TeamTAG depending on the experimental condition. The strict timing, which will be centrally coordinated, will be necessary for a smooth transition of groups between cases and to ensure that the total length of the simulation does not exceed 8 hours.

After the debriefing process, all group members will be asked to evaluate the case and rate how helpful the debriefing was. Group instructors in both conditions will track the main topics of their teamwork debriefing in a debriefing protocol as free text. After the simulation, the content of these debriefing protocols will be clustered

Table 1 Teamwork-relevant cases presented in the emergency department simulation

Case	Diagnosis	Challenges for teamwork
1	Exacerbated COPD	Conflict management, control of emotions due to challenging patient
2	Ischaemic stroke of middle cerebral artery	Task management, communication with colleagues Manage aphasic patient
3	STEMI and non-sustained ventricular tachycardia	Patient deterioration (cardiac arrhythmia) during care
4	Ventricular fibrillation following STEMI	Team leadership, structured ACLS
5	Haemodynamically unstable ruptured spleen	Set priorities in evaluation and management, structured ATLS
6	Head laceration with ethanol intoxication	Manage agitated patient

ACLS, advanced cardiac life support; ATLS, advanced trauma life support; COPD, chronic obstructive pulmonary disease; STEMI, ST-elevation myocardial infarction.

independently (JF and FS) and matched with CRM principles.

Right after the last case of the night, all participants will fill in a final evaluation, which will ask them to list all the CRM principles on which they received feedback during the night. Participants will also evaluate the importance of each principle for their future work as physicians and provide a general evaluation of the night. Every group tutor will rate the feasibility, efficiency and difficulty of providing feedback.

Collected data

1. Baseline characteristics: The data collected on the first questionnaire and the results of group and teamwork discussions will be used to compare the baseline between the two conditions. Discussion results will be analysed qualitatively to identify differences in knowledge and in the personal definition of good teamwork at the beginning of the night. Furthermore, the TEAM scores during the first simulation case will serve as the baseline team performance.
2. Hypothesis 1 measurement (team performance, number of CRM principles discussed): Team performance will be evaluated using the 11 items of the translated TEAM. Similar to previous studies,[15 16 41 45 46] we will analyse ratings on the item level (range: 0–4), the sum score (range: 0–44) and the overall rating per case (range: 1–10). The number of CRM principles discussed will be derived from two sources, namely, the debriefing protocols of the group instructors and participants.
3. Hypothesis 2 measurement (importance, satisfaction, helpfulness): Estimated relevance of the CRM principles learnt and overall satisfaction with the simulation will be evaluated on 7-point Likert scales at the end of the night. Helpfulness of the debriefing from the different providers (simulated patient, peer, case tutor and group tutor) will be rated by participants after every case on a 7-point Likert scale.
4. Hypothesis 3 measurement (instructor ratings): Debriefing evaluation of the group instructors (feasibility, efficiency and difficulty of providing feedback) will be measured with 7-point Likert scales and as free-text answers at the end of the night.
5. Other measures: The general evaluation form will ask participants to rate pleasure, quality of instruction during the night, difficulty of cases and possibility of applying knowledge on 7-point Likert scales.

All 7-point Likert scales will be coded from +3 (*strongly agree*) to −3 (*strongly disagree*). All data collection forms will be available upon request.

Analyses

Data will be analysed in SPSS 24 and R using descriptive, inferential and explorative statistics. We conducted a calculation of power for our primary research question (team performance). Recent studies, reporting mainly data for well-trained and experienced teams, showed TEAM sum scores up to 40.[45 46] Only one study provided data for less experienced teams with a TEAM sum score of 21.[45] On the basis of these results and data from a prestudy (see the TeamTAG section in the Preliminary results section), we expect a TEAM sum score of about 20 for an untrained team and a score of around 40 for teams that receive a training related to teamwork skills and/or have a lot of experience in this area. These scores indicate a potential increase due to training of up to 20 points on the TEAM sum score. As a relevant training effect for a single training event such as ours, we estimate a gain in the TEAM sum score of 11 points (ie, one point per item). Using the SD from the last published study on the TEAM[46] (*SD*=4.4) and $\alpha<0.05$, we have determined that about six teams are needed to detect a significant difference between the conditions with a power of 80%. Missing data will be handled using pairwise deletion.

1. Baseline characteristics: Discussion results of the intervention and control groups will be compared using qualitative methods and confounder analysis (demographics, prior training) with parametric and non-parametric tests for testing equivalence. The TEAM scores (single items, sum score, overall score) from the first simulation case will be compared between conditions using multilevel analyses to take the hierarchical structure of data into account.
2. Analyses for hypothesis 1: The TEAM scores (single items, sum score, overall score) of the intervention and control groups during the sixth simulation case will be compared using multilevel analyses. The development of team performance over the six cases will be analysed using descriptive statistics and plotting 'training curves' for each team. The total number of CRM principles discussed in the control and intervention groups will be compared using a multilevel model.
3. Analyses for hypothesis 2: The participants' ratings of the feedback's helpfulness, the importance of CRM principles and satisfaction with the debriefing will be compared between the control and intervention groups using multilevel models.
4. Analyses for hypothesis 3: Group instructors' evaluations of the instrument will be examined descriptively.
5. Other measures: The general evaluation will be examined in a descriptive way.

Methodological limitations

Group instructors will not be observed while debriefing due to our limited labour force. Therefore, we cannot be sure the quality of the debriefing will be comparable among the seven participating groups. Further studies could use debriefing assessment tools such as the Observational Structured Assessment of Debriefing tool,[47] which might help distinguish between effects of overall debriefing quality and our approach. In our study, we will try to address this limitation with extensive group instructor training to ensure an equal qualification level regarding debriefing and with a randomisation of instructors to conditions. Furthermore, participants will be asked to state the debriefing topic and to rate the quality of debriefing after every simulation case, which will be reported in later publications.

The time for debriefing after every case will be relatively short due to the design of our 8-hour simulation, where all groups will rotate through six cases to give participants a broad overview of emergency medicine and application areas of CRM. To use this limited time most productively, we have added additional specifications for debriefing (eg, focus on one or two principles per debriefing session, as described in the Methods and analysis section) because some instructors stated in a prestudy that the time allowed for debriefing was not sufficient. Future studies could investigate whether results of this study hold if all CRM principles are being discussed and thus repeated after every case/more often during the night and if time for debriefing is longer. Until now, there has been no strong evidence for the superiority of a longer debriefing.[21]

The study will focus only on short-term effects of two different debriefing approaches. Further research should investigate long-term effects on performance or changes in behaviour during clinical practice. A last limitation of this study is that it is a single-centre study and so results might be limited to local circumstances.

Data sharing statement

Data analysis will be conducted by the investigator's team (data management team). As the study is not a clinical trial, a data-monitoring team is not needed. The anonymised full data set will be published together with the journal publication or using the Dryad Data Repository (Durham, NC, USA) as required by the journal's guidelines. Data will furthermore be stored in the local data repository at Charité Medical School Berlin according to the local guidelines for good scientific practice.

PRELIMINARY RESULTS

Validation of the German TEAM

The German TEAM can be found in the online supplementary information. As a preliminary validation, inter-rater correlation was checked between three investigators (JF, FS and DE) and an external expert on two videotaped resuscitations. Both resuscitations were simulation based and had similar factual content; however, the first simulation showed good teamwork and the second intermediate teamwork performance. The videotaped simulations were used for group instructors' debriefing training and for validity testing of the German TEAM.

Intraclass correlation coefficients were .99 for the first resuscitation (mean TEAM score=42.3, *SD*=1.3) and .85 for the second (mean TEAM score=22.5, *SD*=3.1), which indicates excellent inter-rater agreement. For this reason, we consider the German TEAM a valid instrument for assessing team performance in our study.

TeamTAG

A first version of TeamTAG was used in a prestudy, conducted during the previous simulated night shift in 2016. In this prestudy, all instructors (n=7) used TeamTAG as part of their debriefing (similar to the GAS method plus TeamTAG). They were asked to rate the feasibility and helpfulness of the TeamTAG (7-point Likert scale; −3 to +3), as well as whether time for debriefing was sufficient (7-point Likert scale; −3 (*strongly insufficient*) to +3 (*strongly sufficient*)). Furthermore, they could comment on specific aspect of the guideline they liked or disliked (free-text answers). All participants were asked how useful the instructors' feedback was (7-point Likert scale; −3 to +3).

Instructors rated the guideline as a feasible tool (M=1.9, SD=0.9) and stated that it helped them in both observing and giving feedback to the participants of the simulation ($M_{observe}$=2.3, SD=0.8; $M_{feedback}$=2.3, SD=0.5). They had a heterogeneous view of the adequacy of time available for debriefing (M=−0.3, SD=1.1) The participants declared having found the feedback to be useful (M=1.7, SD=1.0).

ETHICS AND DISSEMINATION

The study protocol was designed according to the Declaration of Helsinki, the local guidelines for good scientific practice at Charité Medical School Berlin and the ICMJE (International Committee of Medical Journal Editors) recommendations. The study protocol was approved by the institutional office for data protection (AZ 737/16) and the ethics committee at Charité Medical School Berlin (EA2/172/16).

All participants and instructors will provide informed consent. Because the simulation is already a well-known event at Charité Medical School Berlin and receives official teaching funds, participants who refuse to take part in our study must have a chance to participate nevertheless. In this case, students will not provide the informed consent prior to randomisation; instead, an independent 'no-study' group will then be created, which will be identical to the control group but without any teamwork debriefing. We do not expect any harm for students who undergo the intervention.

Publication

Results of the study will be presented during national and international scientific meetings. The authors aim to publish all results in a peer-reviewed journal. Part of the protocol has been previously presented at the Research in Medical Education (RIME) conference in Duesseldorf, Germany, in March 2017 and was awarded the RIME Award: Best Research Protocol 2017.[48]

Author affiliations
[1]Simulated Patients Program, Charité Medical School Berlin, Berlin, Germany
[2]Lernzentrum (Skills Lab), Charité Medical School Berlin, Berlin, Germany
[3]Department of Emergency Medicine, Inselspital, University of Bern, Bern, Switzerland
[4]Department of Anesthesiology and Operative Intensive Care Medicine CCM & CVK, Charité Medical School Berlin, Berlin, Germany

[5]Progress Test Medizin, Charité Medical School Berlin, Berlin, Germany
[6]Max Planck Institute for Human Development, Center for Adaptive Rationality, Berlin, Germany

Acknowledgements The authors would like to acknowledge Hanno Heuzeroth and David Steinbart for support in conducting the prestudy and Simon Cooper for support in using and translating the TEAM. Furthermore, the authors thank all instructors and physicians, especially Tobias Deselaers, for organising the simulation and their willingness to participate as raters and instructors, and such. In addition, we would like to acknowledge Jane Runnacles (London), Suzanne Bentley (New York) and Christopher Timmis (Wolverhampton) for their constructive critique on an earlier version of the manuscript. The authors thank Anita Todd for editing the manuscript.

Contributors JF and FS translated the TEAM, designed the study and will be responsible for conduction. DE, WEH and JEK contributed to the study design. JEK supervised the study design and will supervise conduction. JF and FS are responsible for data analyses. JF, FS and JEK wrote the manuscript. JF and FS conducted the prestudy. DE is responsible for funding and local administration at CharitÃ© Medical School Berlin and heads the steering committee. All authors carefully read the manuscript, made critical and substantial revisions, and gave their approval for publication.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The investigators JF and FS are funded by the German Federal Ministry of Research and Education (BMBF). The sponsor did not interfere with the conception and conduction of the study, data analysis or production of the manuscript.

Competing interests WEH received financial compensation for educational consultancy from the AO Foundation, Zurich, Switzerland. All other authors report no competing interests.

Ethics approval Ethics committee of Charité Medical School Berlin.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. de Vries EN, Ramrattan MA, Smorenburg SM, *et al*. The incidence and nature of in-hospital adverse events: a systematic review. *Qual Saf Health Care* 2008;17:216–23.
2. Kohn LT, Corrigan JM, Donaldson MS, eds. *To err is human: building a Safer Health System*. Washington: National Academies Press, 2000.
3. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med* 2008;121(5 Suppl):S2–S23.
4. Schwartz A, Elstein AS. *et al*Clinical reasoning in medicine. In: Higgs J, Jones M, Loftus S, Christensen N, . eds. *Clinical reasoning in the health professions*. 3rd ed. Oxford (UK, 2008:223–34.
5. Hautz SC, Schuler L, Kämmer JE, *et al*. Factors predicting a change in diagnosis in patients hospitalised through the emergency room: a prospective observational study. *BMJ Open* 2016;6:e011585.
6. Risser DT, Rice MM, Salisbury ML, *et al*. The Potential for Improved Teamwork to Reduce Medical Errors in the Emergency Department. *Ann Emerg Med* 1999;34:373–83.
7. Leape LL. Error in medicine. *JAMA* 1994;272:1851–7.
8. Hautz WE, Kämmer JE, Exadaktylos A, *et al*. How thinking about groups is different from groupthink. *Med Educ* 2017;51:229.
9. Hautz WE, Kämmer JE, Schauber SK, *et al*. Diagnostic performance by medical students working individually or in teams. *JAMA* 2015;313:303–4.
10. Knaus WA, Draper EA, Wagner DP, *et al*. An evaluation of outcome from intensive care in major medical centers. *Ann Intern Med* 1986;104:410–8.
11. Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiol Scand* 2009;53:143–51.
12. Frank J, Snell L, Sherbino J, eds. *CanMEDS 2015 Physician Competency Framework*. Ottawa: Royal College of Physicians and Surgeons of Canada, 2015.
13. Kaba A, Wishart I, Fraser K, *et al*. Are we at risk of groupthink in our approach to teamwork interventions in health care? *Med Educ* 2016;50:400–8.
14. Kiesewetter J, Gutmann J, Drossard S, *et al*. The Learning Objective Catalogue for Patient Safety in Undergraduate Medical Education–A Position Statement of the Committee for Patient Safety and Error Management of the German Association for Medical Education. *GMS J Med Educ* 2016;33:1–14.
15. Cooper S, Cant R, Connell C, *et al*. Measuring teamwork performance: Validity testing of the Team Emergency Assessment Measure (TEAM) with clinical resuscitation teams. *Resuscitation* 2016;101:97–101.
16. Cooper S, Cant R, Porter J, *et al*. Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). *Resuscitation* 2010;81:446–52.
17. Ochsmann EB, Zier U, Drexler H, *et al*. Well prepared for work? Junior doctors' self-assessment after medical education. *BMC Med Educ* 2011;11:99.
18. Hayes CW, Rhee A, Detsky ME, *et al*. Residents feel unprepared and unsupervised as leaders of cardiac arrest teams in teaching hospitals: a survey of internal medicine residents. *Crit Care Med* 2007;35:1668–72.
19. Brydges R, Hatala R, Zendejas B, *et al*. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med* 2015;90:246–56.
20. Morgan L, Hadi M, Pickering S, *et al*. The effect of teamwork training on team performance and clinical outcome in elective orthopaedic surgery: a controlled interrupted time series study. *BMJ Open* 2015;5:e006216.
21. Cheng A, Eppich W, Grant V, *et al*. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. *Med Educ* 2014;48:657–66.
22. Gaba D, Fish K, Howard S. *Crisis Management in Anesthesiology*. New York: Churchill Livingstone, 1994.
23. Morey JC, Simon R, Jay GD, *et al*. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. *Health Serv Res* 2002;37:1553–81.
24. DeVita MA, Schaefer J, Lutz J, *et al*. Improving medical emergency team (MET) performance using a novel curriculum and a computerized human patient simulator. *Qual Saf Health Care* 2005;14:326–31.
25. Theilen U, Fraser L, Jones P, *et al*. Regular in-situ simulation training of paediatric Medical Emergency Team leads to sustained improvements in hospital response to deteriorating patients, improved outcomes in intensive care and financial savings. *Resuscitation* 2017;115:61–7.
26. Sawyer T, Eppich W, Brett-Fleegler M, *et al*. More Than One Way to Debrief: A Critical Review of Healthcare Simulation Debriefing Methods. *Simul Healthc* 2016;11:209–17.
27. van de Ridder JM, Stokking KM, McGaghie WC, *et al*. What is feedback in clinical education? *Med Educ* 2008;42:189–97.
28. Cook DA, Hamstra SJ, Brydges R, *et al*. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. *Med Teach* 2013;35:e867–e898.
29. Phrampus PE, O'Donnell JM. *et al*Debriefing Using a Structured and Supported Approach. In: Levine AL, DeMaria SJ, Schwartz AD, . eds. *The Comprehensive Textbook of Healthcare Simulation*. New York: springer, 2013:p.73–84.
30. Cheng A, Rodgers DL, van der Jagt ?lise, *et al*. Evolution of the Pediatric Advanced Life Support course. *Pediatric Critical Care Medicine* 2012;13:589–95.
31. Stanford School of Medicine. What Are Cognitive Aids? Cognitive Aids in Medicine . 2016 http://cogaids.stanford.edu/whatarecogaids.html (cited 2016 Dec 14).
32. Harrison TK, Manser T, Howard SK, *et al*. Use of cognitive aids in a simulated anesthetic crisis. *Anesth Analg* 2006;103:551–6.
33. Marshall S. The use of cognitive aids during emergencies in anesthesia: a review of the literature. *Anesth Analg* 2013;117:1162–71.
34. Haugen AS, Søfteland E, Almeland SK, *et al*. Effect of the World Health Organization checklist on patient outcomes: a stepped wedge cluster randomized controlled trial. *Ann Surg* 2015;261:821–8.
35. Cheng A, Hunt EA, Donoghue A, *et al*. Examining pediatric resuscitation education using simulation and scripted debriefing: a multicenter randomized trial. *JAMA Pediatr* 2013;167:528–36.

36. Stegmann K, Pilz F, Siebeck M, *et al*. Vicarious learning during simulations: is it more effective than hands-on training? *Med Educ* 2012;46:1001–8.
37. Weinman J. *Psychology as applied to medicine*. 2nd ed. Bristol, England: Wright, 1987.
38. Jenkins V, Fallowfield L. Can communication skills training alter physicians' beliefs and behavior in clinics? *J Clin Oncol* 2002;20:765–9.
39. Stroben F, Schröder T, Dannenberg KA, *et al*. A simulated night shift in the emergency room increases students' self-efficacy independent of role taking over during simulation. *BMC Med Educ* 2016;16:177.
40. Rall M, Oberfrank S. 'Human factors 'und 'crisis resource management'–Erhöhung der Patientensicherheit. *Unfallchirurg* 2013;116:892–9.
41. Couto TB, Kerrey BT, Taylor RG, *et al*. Teamwork Skills in Actual, In Situ, and In-Center Pediatric Emergencies. *Simulation in Healthcare* 2015;10:76–84.
42. Dorer B. ESS Round 7 translation guidelines [Internet]. *European Social Survey* 2014. http://www.europeansocialsurvey.org/docs/methodology/ESS_R7_Translation_Guidelines_FINAL.pdf (cited 2017 Jan 5).
43. Feldman M, Lazzara EH, Vanderbilt AA, *et al*. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof* 2012;32:279–86.
44. Park HR, Kim CJ, Park JW, *et al*. Effects of team-based learning on perceived teamwork and academic performance in a health assessment subject. *Collegian* 2015;22:299–305.
45. Maignan M, Koch FX, Chaix J, *et al*. Team Emergency Assessment Measure (TEAM) for the assessment of non-technical skills during resuscitation: Validation of the French version. *Resuscitation* 2016;101:115–20.
46. Cant RP, Porter JE, Cooper SJ, *et al*. Improving the non-technical skills of hospital medical emergency teams: The Team Emergency Assessment Measure (TEAM™). *Emerg Med Australas* 2016;28:641–6.
47. Arora S, Ahmed M, Paige J, *et al*. Objective structured assessment of debriefing: bringing science to the art of debriefing in surgery. *Ann Surg* 2012;256:982–8.
48. Freytag J, Stroben F, Kämmer JE, *et al*. Teamwork-related debriefings in undergraduate emergency-medicine simulations : evaluation of the 'TeamTAG' using a prospective- randomized pilot-study design. *Paper Presented at 5th International Conference for Research in Medical Education (RIME 2017)*. Düsseldorf:German Medical Science GMS Publishing House, 2017

Freytag et al. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine (2019) 27:12
https://doi.org/10.1186/s13049-019-0591-9

Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine

ORIGINAL RESEARCH Open Access

Rating the quality of teamwork—a comparison of novice and expert ratings using the Team Emergency Assessment Measure (TEAM) in simulated emergencies

Julia Freytag[1], Fabian Stroben[2,3], Wolf E. Hautz[4,5], Stefan K. Schauber[5] and Juliane E. Kämmer[3,6*]

Abstract

Background: Training in teamwork behaviour improves technical resuscitation performance. However, its effect on patient outcome is less clear, partly because teamwork behaviour is difficult to measure. Furthermore, it is unknown who should evaluate it. In clinical practice, experts are obliged to participate in resuscitation efforts and are thus unavailable to assess teamwork quality. Consequently, we sought to determine if raters with little clinical experience and experts provide comparable evaluations of teamwork behaviour.

Methods: Novice and expert raters judged teamwork behaviour during 6 emergency medicine simulations using the Teamwork Emergency Assessment Measure (TEAM). Ratings of both groups were analysed descriptively and compared with *U* and *t* tests. We used a mixed effects model to identify the proportion of variance in TEAM scores attributable to rater status and other sources.

Results: Twelve raters evaluated 7 teams rotating through 6 cases, for a total of 84 observations. We found no significant difference between expert and novice ratings for 7 of the 11 items of the TEAM or in the sums of all item scores. Novices rated teamwork behaviour higher on 4 items and overall. Rater status accounted for 11.1% of the total variance in scores.

Conclusions: Experts' and novices' ratings were similarly distributed, implying that raters with limited experience can provide reliable data on teamwork behaviour. Novices show a consistent, but slightly more lenient rating behaviour. Clinical studies and real-life teams may thus employ novices using a structured observational tool such as TEAM to inform their performance review and improvement.

Keywords: Teamwork, Non-technical skills, Expert rater, Novice rater, Assessment, Simulation, Resuscitation, Emergency

Background

Medical response to high-urgency situations such as cardiac arrest remains an area for improvement. Depending on their initial rhythm, only around 25% of patients with out-of-hospital cardiac arrest achieve a return of spontaneous circulation (ROSC) [1] and overall survival to discharge lies around 10% [1, 2]. Survival of patients with in-hospital cardiac arrest is higher but still only ranges between 18 and 44% [3, 4].

Besides technical skills such as providing an adequate compression rate [5], working effectively together in a team is connected to patient outcome in high-urgency patients; therefore, training in teamwork behaviour[1] has the potential to improve survival rates [6, 7]. For example, different studies have shown that training in communication and leadership skills in emergency response teams leads to improved ROSC and survival rates [8–10]. Findings from experimental investigations suggest that improved team communication and leadership result in a significant reduction of no-flow time and

* Correspondence: kaemmer@mpib-berlin.mpg.de
[3]AG Progress Test Medizin, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany
[6]Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
Full list of author information is available at the end of the article

better chest compressions in simulated resuscitations [6]. Further, working together in teams can improve diagnostic accuracy in emergency medicine [11, 12] as well as the quality of care compared to individual performance [13].

However, what exactly good teamwork behavior is depends on the task and the role of each team member. Generic rules such as "always practice closed-loop communication" are misleading. For example, one study demonstrated that closed-loop communication initiated by the team leader was associated with a shorter time until the correct diagnosis in an emergency trauma case was made, whereas the same communication pattern delayed the decision significantly if initiated by team members [14]. Also, directive leadership behaviour improved technical performance at the beginning of a resuscitation, whereas in later phases, structuring inquiry (e.g., "What do we know about the patient?") was associated with improved technical performance [6]. These findings show the need to collect more data on teamwork, investigate specific individual and team behaviours, and take differences in task requirements into account. For this, we need valid and reliable tools with known properties that are feasible to use in real-world settings.

In addition, evidence of improvements in patient outcomes as a result of teamwork interventions is limited to a few small studies, many conducted in simulated emergencies [6, 7, 9, 10, 14]. Fung and colleagues suggested that the lack of an objective measurement of team performance is one reason for this paucity of data [15]. While, for example, chest compression rate and depth can nowadays be tracked [16] and technical solutions help to document resuscitations more precisely [17], teamwork behaviour is not easy to measure, especially in real-life situations. Such information is not only relevant for research but also a necessity to inform debriefings after resuscitation [18]. Consequently, different tools have been developed to assess individuals non-technical skills as well as teamwork behaviour. Some of these tools are designed for a specific context, such as the anaesthetists' non-technical skills behavioural marker system (ANTS) [19] or the observational teamwork assessment for surgery (OTAS) [20–22], others are intended to be more generic and independent of context, such as the Ottawa Crisis Resource Management Global Rating Scale [23].

One tool that has been used in both, real-life emergency situations and simulated emergency trainings, is the Teamwork Emergency Assessment Measure (TEAM) [24–27]. The TEAM was designed for emergency teams and is particularly used to assess teamwork, leadership and task-management in high emergency situations such as resuscitation [24, 28]. Since its development in 2010, TEAM has been translated into French [29], Hebrew and Chinese (available via www.medicalemergencyteam.com) and was used in real-life resuscitations [27, 28] and simulated environments (in centre and in situ) [24, 29–32], observing teams of medical and nursing students [24, 31], nurses and physicians [25, 27, 30, 32] and comparing teams with different levels of expertise [29] (see Additional file 1). A recent review showed that it has good psychometric properties in contrast to most other tools for assessing teamwork [18]. In summary, the TEAM has been used in several clinical and simulation-based studies with comparable outcomes (see Additional file 1) and is the most appropriate and valid tool for evaluating teamwork in emergency teams.

While some of the tools meant to quantify non-technical skills and teamwork are intended to be used as self-assessments by practitioners and trainees alike (such as the Mayo High Performance Teamwork Scale [33]), all of the above were designed for raters external to the team they observe [19, 22–24]. Selecting raters to use such instruments is as important as having a suitable tool, yet empirical evidence is lacking concerning *who* should or can assess teamwork behaviour in real or simulated emergencies. During training, it is usually the task of expert raters to assess and debrief participants [34, 35]. Until now, most studies using TEAM have employed expert raters; in two cases TEAM was used as a self-rating instrument for experienced team members as logistical reasons did not allow to recruit external observers [25, 27]. In practice, it might be even harder to find raters with high clinical expertise to observe resuscitations because of their high workload. Such an approach would also lead to ethical problems—especially given that expert raters would have broad knowledge of teamwork and emergency medicine (making them expert in this area), but would be restricted to observing. A possible solution for this methodological, ethical, and organisational dilemma could be the use of less clinically experienced raters, such as residents [36, 37].

We therefore compared novices with expert raters, as these two groups represent the widest difference in clinically relevant qualifications. Both types of raters evaluated teamwork behaviour in an extensive emergency simulation using TEAM. Equivalent ratings from the two rater groups would justify ratings by less experienced raters such as residents also in the workplace.

Methods

Description and translation of TEAM

TEAM consists of 11 items measuring the teamwork behaviour of medical teams dealing with critical situations [24]. The tool consists of 3 subscales: leadership (2 items), teamwork (7 items), and task management (2 items); all items are rated on a Likert scale of 0 (*never/hardly ever*) to 4 (*always/nearly always*). A sum score with a possible range of 0 to 44 can be calculated. Furthermore, overall performance is rated on a global rating scale (GRS) of 1 to 10.

Although a French version exists that confirmed the excellent psychometric properties of the original English version [29], a German version of TEAM is currently lacking. Addressing this gap, our research team has translated TEAM into German using the TRAPD (translation, review, adjudication, pre-testing, and documentation) methodology [38]. A pre-study was conducted to check feasibility and inter-rater reliability and showed excellent results [39].

Data collection

The study was conducted at Charité Universitätsmedizin Berlin during an emergency medicine simulation for final year medical students [40]. During this simulation, the participants acted in teams of 5 and rotated through 6 cases (duration about 30 min each; see Additional file 2: Table S2 for details), in which they had to deal with common emergencies including 1 resuscitation. These cases were realized using simulated patients and high-fidelity simulation. For every case, 1 participant was declared team leader; leadership changed after every case.

Raters

Two groups of raters, one of novices and one of content experts, evaluated participants' teamwork behaviour throughout each case. For the novice raters, we recruited tutors from the local skills lab. They were advanced medical students with emergency medicine experience through clinical electives and/or work experience as paramedics. Expert raters were physicians and psychologists with broad experience in emergency medicine and/or expertise in rating and teaching teamwork during simulation-based education.

Before using TEAM to rate the teams' performances, all raters participated in a rater training [39], which included an introduction to TEAM as a rating instrument, information about common rating errors, and a frame-of-reference training, where videotaped examples of teamwork were rated and discussed [41]. Novice and expert raters received the same training (same length, content etc.) Due to organisational reasons they were trained on two separate occasions. Neither the experts nor the novices had any previous experience with the TEAM as a rating instrument.

Data analysis

Data were analysed using SPSS 24 (Armonk, NY: IBM Corp.) and R, version 3.4.4 [42]. Different descriptive measures were computed separately for the ratings given by novice and expert raters. To analyse the measurement properties of the German version of TEAM, we calculated its' reliability (Cronbach's α), the item-total-score correlation and the correlation of all items plus the sum score with the GRS. As a measure of construct validity, we conducted a principal component analysis (PCA). In a PCA, the objective is to analyse the structure of a data set and to combine a number of observed variables into one factor. We used PCA to check if the items of the German TEAM could be combined into one general component, as was shown for the original version [24, 25]. All results were compared to other studies using TEAM.

Inter-rater reliability between novice and expert raters was calculated (using the intraclass correlation coefficient, ICC) to explore the agreement between these 2 groups. Additionally, their ratings were compared using Mann–Whitney U tests (for the 11 single items) and t tests (for the sum score and GRS).

We used a mixed effects model to identify the sources of variance in TEAM's global rating scale [43]. Mixed effects models are an extension of the ordinary linear regression model that allow for estimating one or more variance components (i.e., random effects) in addition to the residual variance term. In this study, we estimated variance components for teams, raters, rater status (novice or expert), cases, and their first-order interactions.

Results

Participants

During our 8-h emergency simulation, 12 raters (6 novices, 6 experts) rated 7 teams rotating through 6 cases each, resulting in 84 observations in total. Each team consisted of 5 participants; their age ranged between 22 and 46 years (mean [M] = 26.5 years, standard deviation [SD] = 4); 46.9% of the participants were female. The team's performance was rated by pairs of independent observers, 1 expert and 1 novice rater. Both of them were present while the simulation took place and independently rated the teamwork right afterwards. The novice raters had between 1 and 2.5 years of experience in student-assisted learning; experts (5 physicians and 1 psychologist) had 3.5 to 10 years of experience in teaching, including facilitating medical simulations. Further information about the characteristics of the novice and expert raters can be found in Table 1.

Measurement properties of the German translated version of TEAM

We report the measurement properties of the German translated version of TEAM in terms of (1) reliability, (2) item-total-score correlation (i.e., discrimination) and (3) correlation of individual items and the TEAM sum score to the GRS. First, reliability of TEAM instrument—calculated separately for each case and independently for expert and novice raters—had a mean Cronbach's alpha of

Table 1 Characteristics of the novice and expert raters

	Novice raters	Expert raters
N	6	6
Age (Median)	20–33 (24)	26–37 (31.5)
Profession	medical students	5 medical doctors, 1 psychologist
Teaching experience	1–2.5 years (student-assisted learning)	3.5 to 10 years (clinical teaching, simulation-based education, faculty development)
Clinical expertise	Internships (up to 120 days)	1–10 years

0.89 ($SD = 0.06$) for experts and a mean Cronbach's alpha of 0.85 ($SD = 0.19$) for novices. For expert raters, the lowest alpha was .79; it was observed on the case 1 (discipline: surgery). The lowest alpha for novice raters was observed on case 5 (discipline: anaesthesia; alpha = .47). Second, items generally were positively correlated to the sum score of TEAM with a mean of $M_{corr(experts)} = 0.71$ ($SD = 0.09$) and $M_{corr(novices)} = 0.69$ ($SD = 0.17$) across cases for experts and novices, respectively. Third, the TEAM items and the GRS score showed a mean correlation of $M_{corr(experts)} = 0.71$ ($SD = 0.10$) for experts and $M_{corr(novices)} = 0.69$ ($SD = 0.17$) for novices. Finally, across stations, the TEAM sum score and the GRS were significantly correlated both for experts ($r = 0.90$, $p < .001$) and novices ($r = 0.85$, $p < .001$). All psychometric properties mentioned above are compared to the data of studies with the English and French versions of TEAM in Table 2.

Combination of TEAM items into a general component

We conducted the PCA to examine to which degree the individual TEAM items could be combined into a general component. Prior to conducting the PCA, the adequacy of the observed correlation matrix was evaluated using three related statistical criteria. First, the range of inter-item correlations was $range_{r;expert} = 0.29–0.73$ and $range_{r;novices} = 0.42–0.75$. Second, the Kaiser–Meyer–Olkin (KMO) criterion summarizes in how far the obtained variables share unique variance and thus might be combined into a single factor. The KMO was 0.87 for both, expert and novice ratings and therefore exceeded the commonly recommended cut-off of 0.6. Third, the Bartlett test of sphericity which was statistically significant ($p < 0.001$) for both experts and novices, suggesting that the correlation matrix is different from an identity matrix (that is, a correlation matrix where only auto-correlations in the diagonal are of substantial magnitude).

Taken together, the items in the TEAM were sufficiently inter-related to conduct a PCA. The according PCA was, again, conducted independently for novice and expert raters. Results were largely comparable since for both, experts and novices, a dominant first component was found which explained 59 and 65% of the observed variance, respectively.

Agreement between expert- and novice-based ratings

We calculated the inter-rater reliability between novice and expert raters based on the sum scores of TEAM and found an intra-class correlation of ICC = 0.66 (considered *moderate* [44] to *good* [45]). This resemblance between the ratings is also reflected by the finding that expert and novice raters agreed by and large on the lowest and best performing groups for a given case. That is, ratings of experts and novices were consistent in 75% of cases when comparing which teams received the 2 highest and the 2 lowest scores for each case. Furthermore, ratings of experts and novices were compared on the item-level using *U*-tests. On 7 of 11 items, no statistically significant difference was found (items 1, 4, 5, 6, 7, 9, 11; $p = .06–.86$). However, on 4 of 11 items novices rated teamwork behaviour higher than experts on average (items 2, 3, 8, 10; $p = .04–.004$). Furthermore, across cases, we found no statistically significant difference between the TEAM sum scores for experts and novices ($M_{novice} = 30.4$, $SD_{novice} = 8.6$, $M_{expert} = 27.0$, $SD_{expert} = 8.4$; $t(82) = 1.8$, $p = .08$). Finally, for the GRS, we found that novices ($M_{novice} = 7.1$, $SD_{novice} = 1.6$) gave generally higher ratings as compared to experts ($M_{expert} = 6.1$, $SD_{expert} = 1.9$). This difference was statistically significant with $t(82) = 2.5$ and $p = .02$.

Further details on the differences and similarities in ratings between experts and novices are given in Fig. 1 which shows the distribution of standardised GRS and TEAM sum scores. Furthermore, the ranges and quartiles of all items (Additional file 3: Table S3) and the mean sum and

Table 2 Psychometric properties of the German, the French, and the original English version of TEAM [24, 25, 27, 29–32, 54]

Measurement	English TEAM	French TEAM	German TEAM expert rating	German TEAM novice rating
Cronbach's α	0.78–0.97	0.95	0.93	0.94
Inter-item correlation (Spearman's rho)	0.21–1	0.47–0.85	0.29–0.73	0.42–0.75
Item–total correlation	0.42–0.94	0.64–0.79	0.59–0.81	0.38–0.81
Inter-rater reliability[a] (ICC)	0.60–0.94	0.93	0.66	

Legend: *TEAM* Teamwork Emergency Assessment Measure, *ICC* intraclass correlation coefficient
[a]The French and German ICC represent the ICC of the sum score; the range for the ICC in studies with the original TEAM contains both ICC of sum scores and mean ICC of the 11 TEAM items

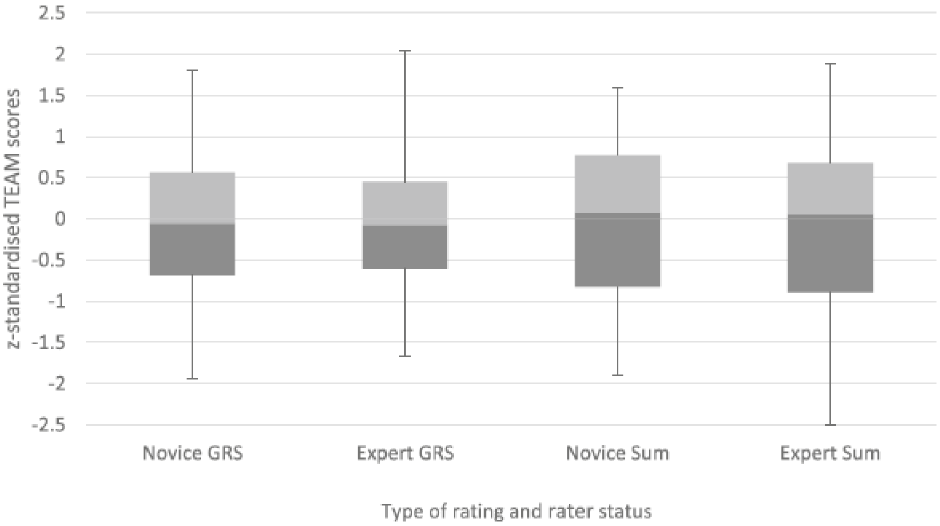


Fig. 1 Distribution of standardised global rating scale (GRS) and sum scores of novice and expert raters. *Legend:* Quartiles 1 and 2 are shown as dark grey boxes, quartiles 3 and 4 as light grey boxes; Whiskers show the minimum and maximum scores. TEAM = Teamwork Emergency Assessment Measure

mean GRS scores as percentages (Additional file 4: Table S4) are provided in the additional files.

Sources of variation of TEAM scores across stations

In order to explain the variations in the overall TEAM scores (GRS) across cases, we estimated variance components and their relative contributions to the total variance using a mixed effects model (Table 3). The model includes random effects for raters, cases, rater-status (i.e., expert/novice) and team (i.e., the particular group of participants). We furthermore estimated random effects for the first-order interactions between cases and teams (do teams perform consistently across cases?) and cases and rater status (do experts and novices differ in their evaluations dependent on particular cases?). In total, the model accounted for 71.8% of the observed variance. We found that rater status (expert vs. novice) accounted for 11.1% of the variance of scores while the cases explained 10.2% of the variance. Teams accounted for 2.6% percent of variation in the observed scores while the biggest source of variance was the interaction of cases and teams with 43.2%, indicating that differences in scores were related to teams performing inconsistently across the different cases.

Discussion

The aim of this study was to compare the rating behaviour of novices and experts using the previously established TEAM instrument. The idea to use novices to assess practical skills is not new, though we could find only one study that examined novices evaluating teamwork behaviour. Sevdalis and colleagues compared the ratings of an expert/expert pair to a novice/expert pair assessing surgical teamwork to analyse the construct validity of the OTAS tool and found relevant differences between expert and novice ratings on almost all items [46]. It is important to notice, though, that in this study the terms *expert* and *novice* referred to their experience in using the tool and both the two participating *experts* and the *novice* had backgrounds in psychology/human factors and were experienced in observing and rating behaviour. The present study, in contrast, defines experts and novices in terms of their content knowledge about teamwork and their practical experience. None of our raters had used TEAM before and they all received a rater training before the simulation.

When focussing on novice raters as raters who are new to or rather unexperienced in a certain area, the literature is generally in favour of novices (even students)

Table 3 Variance Components and Percentage of Variance for TEAM scores

Source of variance	Variance component	Percentage of variance
Rater[a]	0.048	1.32
Rater status[b]	0.397	11.05
Team	0.094	2.62
Case	0.366	10.17
Case × Team	1.553	43.21
Rater Status × Case	0.123	3.42
Residual	1.014	28.21

Legend: *TEAM* Teamwork Emergency Assessment Measure
[a]Rater includes all 12 raters. [b]Rater status includes the categories 'novice rater' and 'expert rater'

being able to assess their peers, although the similarity to expert ratings depends on what skill is assessed and how [47–49]. A recent review [50] on peer assessment in objective structured clinical examinations (OSCE) showed that students awarded consistently higher ratings to their peers than experts when using GRS. Our study shows similar results when comparing the GRS scores, as novices rated the team behaviour on average 1 point higher than experts did (scale: 1–10); on some single items, novices rated significantly higher than experts, whereas in the majority of cases, including the sum score of all 11 items, there was no difference. In this context it is important to notice the large positive correlation of the sum scores of experts and novices as well as their consistent ratings of the best and worst performances, which justify the use of novices as raters. Novice raters' tendency to give better ratings might be explained by a lower standard against which they compared their peers. Looking from the experts' point of view, it seems plausible that experts are more aware of potentially serious consequences of bad teamwork because of their work experience and therefore rated more strictly [51, 52]. The moderate ICC of 0.66 is connected to this discrepancy between experts and novices. The 2 rater groups seem to have had different baselines, although all raters underwent the same training and anchoring process. The results of the z standardization of GRS and TEAM sum scores endorse this theory of different baselines. When each rater group's scores were transformed to have a mean of 0 and a standard deviation of 1, their ratings showed very similar distribution patterns (similar range/interquartile range).

Unexpectedly, the teams themselves were only a very small source of the variance in performance scores (3%) and the interaction of team and case was by far the biggest source of variance (43%). In other words, a team's performance varied considerably between the different cases and there were no superb or incapable teams per se. Importantly, since team leadership changed across cases, the 2 components (team leader and case) are confounded and thus cannot be disentangled statistically. Therefore, it is not clear whether variation in performances across cases is attributable to team leadership or the specific task. Still, our results suggest that a team's performance depends to a considerable extent on the specifics of the situation. This finding has several implications. Firstly, it suggests that the recurrent finding of context specificity in clinical decision making of the individual is also relevant at the team level [53]. Secondly, this further emphasises the importance of a close investigation of what teamwork behaviour by whom is beneficial in exactly what situation—as opposed to generic rules meant to characterize 'good teamwork'. Future training should abandon statements such as 'practice closed-loop communication' in favour of advice such as 'During the first minutes of cardiopulmonary resuscitation (CPR), closed-loop communication initiated by the directive team leader is beneficial for CPR quality' [6, 14]. Thirdly, TEAM scores should not be compared across different cases. The absence of clear benchmarks and the uncertain connection of TEAM scores and objective criteria remain problems when rating teams [25, 27].

As a beneficial side effect of our study, we validated the German version of TEAM, which is now available for clinical use (Additional file 5: Figure S1). Psychometric properties were comparable to those of the English original [24, 25, 27, 30, 31, 54] and the French translation [29]. The internal consistency for both novice and expert ratings was very high, the inter-rater reliability can be considered moderate, and the PCA confirmed 1 underlying component.

This study has several limitations. Firstly, it was a single-centre study with a small sample size. Although our number of observations (84) is similar to or even higher than in other studies using TEAM, our results are based on the ratings of 6 novice and 6 expert raters and each scenario was only observed by 2 of those 12 raters. Secondly, this study took place in a simulation setting that included different cases and changing team structure. Thirdly, our raters only observed monoprofessional teams, consisting of final year medical students. As our study is one of the first to use TEAM outside of typical resuscitation scenarios, more research is needed to decide how suitable TEAM is for rating teamwork behaviour in situations other than CPR and how to set performance benchmarks.

Conclusions

Teamwork behaviour can be assessed with TEAM by novices just as well as by clinically experienced raters, though novices tend to rate slightly more lenient than experts do. Further research is needed on the comparability of TEAM scores across different cases. The German TEAM is a reliable and valid tool to assess teamwork performance that closes a gap in measuring teamwork behaviour in German-speaking countries.

Endnotes

[1]In this study, we use the term *teamwork behaviour* to highlight that we treat non-technical skills such as communication and leadership skills at the team level as a kind of 'collective' non-technical skill; we did not evaluate team members individually.

Additional files

Additional file 1: Overview of studies using TEAM including raters, ratees, and settings of these studies. (DOCX 24 kb)

Additional file 2: Cases and simulation settings (Discipline, diagnosis and mode of simulation of all 6 cases used in the study). (DOCX 17 kb)

Additional file 3: Range and quartiles of the 11 items of TEAM for novice and expert raters. (DOCX 21 kb)

Additional file 4: Means and standard deviations of sum and global rating scale scores and mean scores as percentages. (DOCX 14 kb)

Additional file 5: Team Emergency Assessment Measure (German translation). (DOCX 919 kb)

Additional file 6: Data Set (Team Emergency Assessment Measure scores of novices and experts rating 7 teams rotating through 6 cases each). (XLSX 15 kb)

Abbreviations

CPR: Cardiopulmonary resuscitation; GRS: Global rating scale; ICC: Intraclass correlation coefficient; KMO: Kaiser–Meyer–Olkin; M: Mean; OSCE: Objective structured clinical examination; PCA: Principal component analysis; ROSC: Return of spontaneous circulation; SD: Standard deviation; TEAM: Teamwork Emergency Assessment Measure; TRAPD: Translation, review, adjudication, pre-testing, documentation

Acknowledgements

The authors would like to acknowledge Simon Cooper for his help in translating TEAM and Hanno Heuzeroth, David Steinbart and Dorothea Eisenmann for their support in conducting the prestudy. Furthermore, we thank all participants and raters for participating in this study. Our manuscript was revised by Anita Todd, a language editor who was paid by the Max Planck Institute for Human Development, Berlin, the host institution of the last author.

Funding

JF and FS are funded by the German Federal Ministry of Research and Education (BMBF). The sponsor did not interfere with the conception and conduction of the study, data analysis, or production of the manuscript.

Availability of data and materials

All data generated and analysed during this study are included in Additional file 6.

Authors' contributions

JF and FS designed the study, analysed and interpreted the data, and drafted the manuscript. SS contributed to data analysis and interpretation and helped to revise the manuscript. WEH and JEK contributed to the design of the study, the interpretation of the findings and revised the manuscript. All authors have read and approved of the final version of this manuscript.

Ethics approval and consent to participate

The ethics committee (EA2/172/16) and the institutional office for data protection (AZ 737/16) at Charité Universitätsmedizin approved the study. All participants and raters consented orally and in written form.

Consent for publication

Not applicable, since manuscript does not include individual person's data.

Competing interests

WEH received financial compensation for educational consultancy from the AO Foundation, Zurich, Switzerland and research funding from Mundipharma Medical, Basel, Switzerland. All other authors report no competing interests.

Publisher's Note

Author details

[1]Simulated Patients Program, Office of the Vice Dean for Teaching and Learning, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany. [2]Lernzentrum, Office of the Vice Dean for Teaching and Learning, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany. [3]AG Progress Test Medizin, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany. [4]Department of Emergency Medicine, Inselspital, Bern University Hospital, University of Bern, Freiburgstrasse 4, 3010 Bern, Switzerland. [5]Centre for Health Sciences Education, University of Oslo, Gaustadalléen 30, 0373 Oslo, Norway. [6]Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany.

Received: 18 July 2018 Accepted: 14 November 2018
Published online: 08 February 2019

References

1. Gräsner J-T, Lefering R, Koster RW, Masterson S, Böttiger BW, Herlitz J, et al. EuReCa ONE-27 nations, ONE Europe, ONE registry: a prospective one month analysis of out-of-hospital cardiac arrest outcomes in 27 countries in Europe. Resuscitation. 2016;105:188–95.
2. Daya MR, Schmicker RH, Zive DM, Rea TD, Nichol G, Buick JE, et al. Out-of-hospital cardiac arrest survival improving over time: results from the resuscitation outcomes consortium (ROC). Resuscitation. 2015;91:108–15.
3. Nadkarni VM, Larkin G, Peberdy MA, Carey SM, Kaye W, Mancini ME, et al. First documented rhythm and clinical outcome from in-hospital cardiac arrest among children and adults. JAMA. 2006;295:50–7.
4. Nolan JP, Laver SR, Welch CA, Harrison DA, Gupta V, Rowan K. Outcome following admission to UK intensive care units after cardiac arrest: a secondary analysis of the ICNARC case mix Programme database. Anaesthesia. 2007;62:1207–16.
5. Talikowska M, Tohira H, Finn J. Cardiopulmonary resuscitation quality and patient survival outcome in cardiac arrest: a systematic review and meta-analysis. Resuscitation. 2015;96:66–77.
6. Tschan F, Semmer NK, Gautschi D, Hunziker P, Spychiger M, Marsch SU. Leading to recovery: group performance and coordinative activities in medical emergency driven groups. Hum Perform. 2006;19:277–304.
7. Westli HK, Johnsen BH, Eid J, Rasten I, Brattebo G. Teamwork skills, shared mental models, and performance in simulated trauma teams: an independent group design. Scand J Trauma Resusc Emerg Med. 2010;18:47.
8. Ornato JP, Peberdy MA, Reid RD, Feeser VR, Dhindsa HS, NRCPR Investigators. Impact of resuscitation system errors on survival from in-hospital cardiac arrest. Resuscitation. 2012;83:63–9.
9. Hunziker S, Johansson AC, Tschan F, Semmer NK, Rock L, Howell MD, et al. Teamwork and leadership in cardiopulmonary resuscitation. J Am Coll Cardiol. 2011;57:2381–8.
10. Ford K, Menchine M, Burner E, Arora S, Inaba K, Demetriades D, et al. Leadership and teamwork in trauma and resuscitation. West J Emerg Med. 2016;17:549–56.
11. Hautz WE, Kämmer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. JAMA. 2015;313:303–4.
12. Wolf M, Krause J, Carney PA, Bogart A, Kurvers RH. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. PLoS One. 2015;10:e0134269.
13. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An evaluation of outcome from intensive care in major medical centers. Ann Intern Med. 1986;104:410–8.
14. Hargestam M, Lindkvist M, Jacobsson M, Brulin C, Hultin M. Trauma teams and time to early management during in situ trauma team training. BMJ Open. 2016;6:e009911.
15. Fung L, Boet S, Bould MD, Qosa H, Perrier L, Tricco A, et al. Impact of crisis resource management simulation-based training for interprofessional and interdisciplinary teams: a systematic review. J Interprof Care. 2015;29:433–44.
16. Bobrow BJ, Vadeboncoeur TF, Stolz U, Silver AE, Tobin JM, Crawford SA, et al. The influence of scenario-based training and real-time audiovisual feedback on

out-of-hospital cardiopulmonary resuscitation quality and survival from out-of-hospital cardiac arrest. Ann Emerg Med. 2013;62:47–56 e1.

17. Grundgeiger T, Albert M, Reinhardt D, Happel O, Steinisch A, Wurmb T. Real-time tablet-based resuscitation documentation by the team leader: evaluating documentation quality and clinical performance. Scand J Trauma, Resusc Emerg Med. 2016;24:51.
18. Valentine MA, Nembhard IM, Edmondson AC. Measuring teamwork in health care settings: a review of survey instruments. Med Care. 2015;53:e16–30.
19. Fletcher G, Flin R, McGeorge P, Glavin RJ, Maran NJ, Patey R. Anaesthetists' non-technical skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth. 2003;90:580–8.
20. Undre S, Healey AN, Darzi A, Vincent CA. Observational assessment of surgical teamwork: a feasibility study. World J Surg. 2006;30:1774–83.
21. Undre S, Sevdalis N, Healey AN, Darzi A, Vincent CA. Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. World J Surg. 2007;31:1373–81.
22. Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. Qual Saf Health Care. 2004;13:i33–40.
23. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa critical care medicine, high-Fidelity simulation, and crisis resource management I study. Crit Care Med. 2006;34:2167–74.
24. Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, et al. Rating medical emergency teamwork performance: development of the TEAM emergency assessment measure (TEAM). Resuscitation. 2010;81:446–52.
25. Cooper S, Cant R, Connell C, Sims L, Porter J, Symmons M, et al. Measuring teamwork performance: validity testing of the TEAM emergency assessment measure (TEAM) with clinical resuscitation teams. Rescucitation. 2016;101:97–101.
26. Couto TB, Kerrey BT, Taylor RG, FitzGerald M, Geis GL. Teamwork skills in actual, in situ, and in-center pediatric emergencies: performance levels across settings and perceptions of comparative educational impact. Simul Healthc. 2015;10:76–84.
27. Cant RP, Porter JE, Cooper SJ, Roberts K, Wilson I, Gartside C. Improving the non-technical skills of hospital medical emergency teams: the TEAM emergency assessment measure (TEAM). Emerg Med Australas. 2016;28:641–6.
28. Cooper S. Teamwork: what should we measure and how should we measure it? Int Emerg Nurs. 2017;32:1–2.
29. Maignan M, Koch F-X, Chaix J, Phellouzat P, Binauld G, Collomb Muret R, et al. TEAM emergency assessment measure (TEAM) for the assessment of non-technical skills during resuscitation: validation of the French version. Resuscitation. 2016;101:115–20.
30. McKay A, Walker ST, Brett SJ, Vincent C, Sevdalis N. Team performance in resuscitation teams: comparison and critique of two recently developed scoring tools. Resuscitation. 2012;83:1478–83.
31. Bogossian F, Cooper S, Cant R, Beauchamp A, Porter J, Kain V, et al. Undergraduate nursing students' performance in recognising and responding to sudden patient deterioration in high psychological fidelity simulated environments: an Australian multi-Centre study. Nurse Educ Today. 2014;34:691–6.
32. Cooper S, Cant R, Porter J, Missen K, Sparkes L, McConnell-Henry T, et al. Managing patient deterioration: assessing teamwork and individual performance. Emerg Med J. 2012;30:377–81.
33. Malec JF, Torsher LC, Dunn WF, Wiegmann DA, Arnold JJ, Brown DA, et al. The Mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. Simul Healthc. 2007;2:4–10.
34. Cheng A, Eppich W, Grant V, Sherbino J, Zendejas B, Cook DA. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. Med Educ. 2014;48:657–66.
35. Sawyer T, Eppich W, Brett-Fleegler M, Grant V, Cheng A. More than one way to debrief: a critical review of healthcare simulation debriefing methods. Simul Healthc. 2016;11:209–17.
36. Hughes TC, Jiwaji Z, Lally K, Lloyd-Lavery A, Lota A, Dale A, et al. Advanced Cardiac Resuscitation Evaluation (ACRE): a randomised single-blind controlled trial of peer-led vs. expert-led advanced resuscitation training. Scand J Trauma Resusc Emerg Med. 2010;18:3.
37. Harvey PR, Higenbottam CV, Owen A, Hulme J, Bion JF. Peer-led training and assessment in basic life support for healthcare students: synthesis of literature review and fifteen years practical experience. Resuscitation. 2012;83:894–9.
38. Dorer B. Round 6 translation guidelines. Mannheim: European Social Survey, GESIS; 2012.
39. Freytag J, Stroben F, Hautz WE, Eisenmann D, Kammer JE. Improving patient safety through better teamwork: how effective are different methods of simulation debriefing? Protocol for a pragmatic, prospective and randomised study. BMJ Open. 2017;7:e015977.
40. Stroben F, Schröder T, Dannenberg KA, Thomas A, Exadaktylos A, Hautz WE. A simulated night shift in the emergency room increases students' self-efficacy independent of role taking over during simulation. BMC Med Educ. 2016;16:177.
41. Eppich W, Nannicelli AP, Seivert NP, Sohn MW, Rozenfeld R, Woods DM, et al. A rater training protocol to assess team performance. J Contin Educ Heal Prof. 2015;35:83–90.
42. CoreTeam R. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
43. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48.
44. Koo TK, Li MY. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155–63.
45. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6:284–90.
46. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A, Vincent CA. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. Ann Surg. 2009;249:1047–51.
47. Falchikov N, Goldfinch J. Student peer assessment in higher education: a meta-analysis comparing peer and teacher Marks. Rev Educ Res. 2000;70:287–322.
48. Falchikov N. Improving assessment through student involvement: practical solutions for aiding learning in higher and further education. New York: Routledge; 2005.
49. Topping K. Peer assessment between students in colleges and universities. Rev Educ Res. 1998;68:249–76.
50. Khan R, Payne MWC, Chahine S. Peer assessment in the objective structured clinical examination: a scoping review. Med Teach. 2017;39:745–56.
51. Chenot JF, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, et al. Can student tutors act as examiners in an objective structured clinical examination? Med Educ. 2007;41:1032–8.
52. Iblher P, Zupanic M, Karsten J, Brauer K. May student examiners be reasonable substitute examiners for faculty in an undergraduate OSCE on medical emergencies? Med Teach. 2015;37:374–8.
53. Eva KW. On the generality of specificity. Med Educ. 2003;37:587–8.
54. Cooper S, Beauchamp A, Bogossian F, Bucknall T, Cant R, Devries B, et al. Managing patient deterioration: a protocol for enhancing undergraduate nursing students' competence through web-based simulation and feedback techniques. BMC Nurs. 2012;11:18.

Hautz et al. BMC Medical Education (2020) 20:201
https://doi.org/10.1186/s12909-020-02115-3

BMC Medical Education

RESEARCH ARTICLE Open Access

An observational study of self-monitoring in ad hoc health care teams

Stefanie C. Hautz[1], Daniel L. Oberholzer[1], Julia Freytag[2], Aristomenis Exadaktylos[1], Juliane E. Kämmer[1,3], Thomas C. Sauter[1,4] and Wolf E. Hautz[1*]

Abstract

Background: Working in ad hoc teams in a health care environment is frequent but a challenging and complex undertaking. One way for teams to refine their teamwork could be through post-resuscitation reflection and debriefing. However, this would require that teams have insight into the quality of their teamwork. This study investigates (1) the accuracy of the self-monitoring of ad hoc resuscitation teams and their leaders relative to external observations of their teamwork and (2) the relationship of team self-monitoring and external observations to objective performance measures.

Methods: We conducted a quantitative observational study of real-world ad hoc interprofessional teams responding to a simulated cardiac arrest in an emergency room. Teams consisting of residents, consultants, and nurses were confronted with an unexpected, simulated, standardized cardiac arrest situation. Their teamwork was videotaped to allow for subsequent external evaluation on the team emergency assessment measure (TEAM) checklist. In addition, objective performance measures such as time to defibrillation were collected. All participants completed a demographic questionnaire prior to the simulation and a questionnaire tapping their perceptions of teamwork directly after it.

Results: 22 teams consisting of 115 health care professionals showed highly variable performance. All performance measures intercorrelated significantly, with the exception of team leaders' evaluations of teamwork, which were not related to any other measures. Neither team size nor cumulative experience were correlated with any measures, but teams led by younger leaders performed better than those led by older ones.

Conclusion: Team members seem to have better insight into their team's teamwork than team leaders. As a practical consequence, the decision to debrief and the debriefing itself after a resuscitation should be informed by team members, not just leaders.

Keywords: Debriefing, Self-monitoring, Ad hoc teams, Resuscitation, Emergency medicine, Postgraduate education, Interprofessional education

Background

Health care is an inherently social and interdisciplinary endeavour [1, 2]. Take, for example, emergency departments, where teams of physicians, nurses, and other health care professionals are routinely involved in diagnosing and treating patients [1, 3–5]. Collaboration is particularly common in high-urgency situations such as trauma calls or resuscitations that are typically handled by ad hoc teams (also known as health care action teams or interdisciplinary action teams): interdisciplinary, interprofessional groups of specialized individuals who work together in a highly dynamic, complex situation and under time pressure to accomplish critical tasks [6].

* Correspondence: wolf.hautz@insel.ch
[1]Department of Emergency Medicine, Inselspital University Hospital Bern, Freiburgstrasse 16c, 3010 Bern, Switzerland
Full list of author information is available at the end of the article

Especially when time is of the essence, the performance of such teams critically depends on the quality of their teamwork [7, 8], a term that summarizes "*human factors and non-technical skills (leadership, [collaboration], situation awareness and decision making)*" [9]. For example, Tiel and colleagues demonstrated that patient mortality is lower when trauma teams are trained in teamwork, including leadership skills [10]. One such key skill is coordination [11] or, as Marks et al. put it, "orchestrating the sequence and the timing of interdependent actions" [12]. Teamwork in emergencies furthermore should be contextually adaptive and account for team member experience and patient acuity [13].

Obviously, developing such teamwork requires training and experience of team members. Plenty of publications describe simulations to address this need (e.g., [14–19]). A few common training aims, often referred to as crew (or crisis) resource management principles (CRM), are widely adopted [20]. This conception implicitly assumes that training CRM in one ad-hoc team transfers to other ad-hoc teams, in which trainees may find themselves in the next day. However, it remains unclear if that is indeed the case and whether and how the skills acquired in such simulations translate to clinical practice [21, 22]. Fortunately, clinical practice itself also provides many opportunities for real-life ad hoc teams to (further) refine their teamwork.

Several tools have been developed to guide feedback to resuscitation teams by scaffolding outside observations (e.g., [9, 23–26]). Post-resuscitation debriefing, for example, is recommended by resuscitation guidelines internationally [27, 28] to facilitate reflection upon teamwork as it has been associated with important patient outcomes such as return of spontaneous circulation and shorter no-flow time in cardiopulmonary resuscitation. The use of these tools requires, however, the presence of a trained observer, which poses both logistical and ethical challenges [29]. Because ad hoc teams, particularly in emergency care, can be needed at any time of any day, it is rarely feasible to ensure an observer's availability. In addition, it may be ethically unacceptable to limit the outside expert's role to observation if the team's performance is less than optimal. Yet, prompt intervention would disturb both the team and the collection of information for later debriefing.

One potential way to inform debriefing that circumvents the logistical and ethical challenges is self-reflection of the team. Several frameworks for post-resuscitation debriefing exploit that possibility in that they ask the team to reflect on "what went well" and "what could have gone better", for example [30]. Almost all emergency rooms (ERs) establish the role of a physician team leader in their resuscitation teams. Because most post-resuscitation debriefing frameworks call upon the physician team leader or the charge nurse to decide whether a debrief is required at all (e.g. [30]), the accuracy of their self-evaluation is of critical importance.

But are teams and their leaders accurate judges of their own teamwork? The literature on self-monitoring suggests so [31–33]: Although it has repeatedly been demonstrated that individuals' *self-assessments* are low in accuracy [34, 35], people are much better able to *self-monitor* their current performance [31–33]. The key conceptual difference between self-assessment and self-monitoring is the timing of introspection: self-assessment refers to a summative overall self-evaluation detached from a single event (e.g., "How good a team player am I?"), whereas self-monitoring is a moment-to-moment assessment of one's performance in a given situation (e.g., "How is our teamwork in this particular case?") [35]. Self-monitoring is what prompts people to "slow down when they should" [33] or to look up a word when they are unsure of its meaning [35]. Self-reflection of team members before debriefing within a team directly after attending to a patient is conceptually closer to self-monitoring than to self-assessment, because the reflection refers to a single event in close temporal proximity, not an overall and aggregate judgement.

Based on this conception of self-monitoring, we hypothesize that the ability of both teams and their physician team leaders to self-monitor their teamwork directly after attending to a patient will be comparable to that of an external observer. Such an ability would justify relying on team members' self-monitoring to inform debriefing in real-life teams, thus easing and facilitating training on the job.

Furthermore, we investigated whether the ratings of teamwork provided by the team, their leader and external observers are related to objective measures of performance that directly affect patient outcomes [36]. It is arguably only clinically meaningful and defensible to assess how well teams and their physician team leaders self-monitor their teamwork relative to an external observer if either the team's self-monitoring judgements or the external observations are related to direct patient outcomes or measures of proven importance.

This quantitative observational study therefore addressed the following research questions:

1) How accurately do ad hoc resuscitation teams and their leaders monitor their teamwork relative to external observers' rating of that teamwork?
2) How do teams' self-monitoring judgements and external observations of their teamwork relate to objective performance measures?

Methods

We conducted an observational study of real-world interprofessional ad hoc teams responding to a simulated cardiac arrest in the emergency room. Teams consisting of residents, consultants, and nurses were confronted with an unexpected, simulated, standardized cardiac arrest situation. Their performance was videotaped to allow for subsequent external evaluation.

Participants, teams, and setting

The study was conducted in a single university-affiliated level-one emergency room attending to more than 50.000 patients annually [37]. All medical staff in this emergency room are required to attend an interprofessional training day once annually. Staff were assigned to one of 10 training days in their work schedule; participation in these trainings is mandatory. On each training day, participants were randomly assigned to one of four teams. Randomization was stratified by profession, so that each team included at least one physician and two nurses. Depending on training day attendance, some teams were larger than others, including up to three physicians and up to four nurses, reflecting team variability in clinical practice. The most senior physician in each team was appointed *physician team leader* upon team formation, reflecting clinical practice in the emergency room under investigation. We did not provide further instructions to physician team leaders because all participating physicians regularly fulfil this role in clinical practice, and we did not expect any differences between that role in the simulation and clinical practice.

Simulation and scenario

The 2019 interprofessional training day was designed as a rotation through different skill stations, where participants trained techniques such as intraosseous access, patient positioning, and paediatric advanced life support. Further, all teams rotated to a simulated shift handover, which was unexpectedly interrupted by a resuscitation call. Teams attended to this call as they would in reality; the simulation was designed as an in situ simulation and took place in the actual resuscitation bay of the emergency room.

In this resuscitation bay, teams were confronted with an elderly male patient on a stretcher (represented by a Leardal SimMan Essential patient simulator) under ongoing cardiopulmonary resuscitation (CPR) by paramedics. The paramedics reported to have initially encountered a conscious but helpless elderly patient in his flat. There, the patient was found lying on the floor after tripping over the sill of his balcony door 10 h previously when coming back inside from his balcony. He had fallen from body height and had been unable to alert help for an extended period up until a neighbour heard his calls for help. The paramedics found a hemodynamically stable, cold, and conscious but disoriented patient. They established a venous access and transferred the patient to the hospital uneventfully. Upon transfer from the ambulance onto the hospital's stretcher just minutes prior, the patient had gone into cardiac arrest. The paramedics started CPR and alerted the ER staff.

The paramedics repeatedly informed the attending ER team that the patient's flat was cold due to the balcony door being open and that the patient was cold as well. Paramedics had just initiated basic life support. Neither medication nor defibrillation had yet been provided. The paramedics then handed over resuscitation to the ERs team but remained available for further questions while they reorganized their material.

The patient simulator was programmed to exhibit a ventricular fibrillation and not to respond to resuscitative measures for the next 15 min. The team worked in its familiar emergency room environment with its usual equipment. Whenever a team member providing chest compressions was replaced, a study aid quietly informed them that their hands felt very cold due to the patient's cold chest. This was necessary because the simulator used could not be programmed to feel cold. Whenever a team decided to take the patient's temperature, the thermometer read 27.3 °C independent of the location of temperature measurement. All phone calls made by the team were answered by the simulation instructors, who responded with a standard response indicating that they would be available within a few minutes.

The simulation was wound up as soon as teams called for extracorporeal circulation (ECC) due to hypothermic arrest or ended by the instructor if teams had clearly made a different diagnosis of why the patient was in arrest. The whole scenario was intended to last less than 15 min. It was followed by a structured, instructor-led debriefing that included reviewing a video recording of the simulation. The debriefing was led by a trained and experienced simulation instructor and revolved around teamwork, team leadership and non-technical skills.

Ethics

The ethics committee of the Canton of Bern deemed the study to be exempt from full ethical review (Req-2017-00968) because it did not involve patients. All participants provided written informed consent for their data to be used in the study. As an incentive, all participants were entered in a lottery for the chance to win one of two tablet computers.

Measurement instruments

Participant questionnaires

All participants in the study independently answered two custom questionnaires. The first, which was administered prior to the simulation scenario, collected demographic information, data on participants' professional experience, and information on their degree of acquaintance with the other team members. The data were collected as numbers (age, experience), through checkboxes (gender), or on 5-point Likert scales (acquaintance). The second questionnaire was completed individually directly after the simulation scenario and before debriefing. It assessed participants' familiarity with the task ("Please indicate how familiar you are with cases like the one you just encountered.") and their self-monitoring of their team's teamwork ("Please indicate your confidence in the quality of your team's teamwork.") on 5-point Likert scales. It is generally assumed that the resulting confidence scores reflect the metacognitive feeling of fluency (see e.g., [38]). Previous research has shown a close relation of such confidence scores to other indicators of fluency, such as response time on diagnostic tasks [39] and the likelihood to change an initial answer in multiple choice tests [40].

Video recordings and choice of rating tool

All simulations were video recorded and recordings were evaluated by independent external expert raters, one psychologist (JF) and one emergency physician (WEH), both with extensive simulation experience. Their ratings were recorded using the Team Emergency Assessment Measure (TEAM) checklist [9]. In contrast to many other tools assessing teamwork (for a review, see [41]), TEAM has good psychometric properties such as a high interrater-reliability and high inter-item correlations, has been validated in several translated versions, and used in a variety of simulated and real-life resuscitation scenarios (for a review see [29]). In this study, we employed the German translation of the instrument [29].

The TEAM checklist is a previously published instrument that consists of 11 items assessing teamwork of medical emergency teams on 3 subscales: leadership (2 items), teamwork (7 items), and task management (2 items). All items are rated on a Likert scale ranging from 0 (*never/hardly ever*) to 4 (*always/nearly always*) and are then summarized into a sum score ranging between 0 and 44. In addition, teamwork is rated on a global rating scale (GRS) ranging from 1 (*worst possible performance*) to 10 (*best possible performance*). The GRS was worded as "Please provide an overall evaluation of the non-medical performance of this team." Because a previous factor analysis [36] identified one underlying component accounting for 81% of the observed variation, with factor loadings >0.6 and Eigenvalues >1, it is defensible and common practice to summarize items of the TEAM into a sum score in the range between 0 and 44. The two expert raters each assessed 20% of the videos independently and in duplicate. As inter-rater agreement was good to excellent (intraclass correlation coefficient ICC = 0.87), the remaining videos were rated by a single rater.

We further extracted objective performance measures from the videos. In patients under resuscitation, such measures include early defibrillation and timing and number of doses of adrenaline administered [42], and implementation of extracorporeal circulation (ECC) in hypothermic cardiac arrest [43] and will be referred to as *performance* hereafter.

Data management and analysis

Questionnaire data were recorded on paper and entered into an excel sheet by an administrative assistant. 20% of the data entered were randomly selected and cross-checked by one of the authors (SCH). Videos were recorded using the AV system permanently installed in the resuscitation bay and stored on a secure network storage drive institutionally approved for storage of health related personal information. Video ratings and the extracted performance data such as time to first defibrillation were directly entered into a spreadsheet by expert raters. All data collected were then imported into an SPSS file. SPSS version 25 (IBM cooperation) was used for data analysis. Data are described using mean or median and standard deviation or range, as appropriate. We calculated Pearson's correlations between the rater-based measures of teamwork (TEAM sum score and GRS) and participants' self-evaluation of teamwork. We also correlated the rater-based measures of teamwork with team characteristics such as size, cumulative experience, and age of the team leader. Last, we assessed the relationship between TEAM sum scores or GRS and other behavioural data, such as time to first defibrillation. *P* values of less than 0.05 are considered significant.

Results

A total of 26 teams participated in the study. Two teams were excluded from the analysis because the videos were not recorded due to a technical failure. One team was excluded because two of its participants did not consent to study participation, and one team because the team contained a physician involved in this study. The analysis was thus based on 22 teams consisting of 115 health care professionals (22 physician team leaders and 93 team members). Participants were on average 36.7 years old and had a mean of 13.1 years' professional experience, 6.3 years of those in emergency medicine; 77.4% were female (Table 1). Team leaders were on average 36.4 years old and had a mean of 9 years' professional experience, 3.4 years of those in emergency medicine; 77.3% were female. Each team contained at least one physician (range 1–3, median 1) and two nurses (range 2–4, median 3); no team was smaller than 3 persons (range 3–6, median 4).

The objectively observable performance of the teams was highly variable: On average, it took 8 min and 15 s

Table 1 Descriptive statistics

	Mean	Standard deviation	Min	Max
Participants overall				
Age [years]	36.7	10.3	20	63
Professional experience [years]	12.7	9.7	0.4	39
Experience emergency medicine [years]	6.3	7.5	0.2	36
Gender	77.4% female			
Team leader				
Age [years]	36.4	8.3	29	61
Professional experience [years]	9.0	8.6	2	37
Experience emergency medicine [years]	3.4	5.8	1	27
Gender	77.3% female			
Scenarios and Performance				
Total duration [min:sec]	8:15	1:05	5:45	10:15
Time to defibrillation [min:sec]	1:06	0:42	0:09	2:32
Time to adrenaline [min:sec]	3:31	1:34	0:50	6:47[a]
TEAM GRS score [points out of 10]	6.4	1.8	3	10
TEAM sum score [points out of 44]	31.8	5.5	21	42

[a] Two teams did not administer adrenaline at all.

(SD = 1:05 min) before teams called for ECC due to hypothermic arrest ($n = 9$) or the instructor ended the simulation ($n = 13$). Although all teams correctly recognized an arrest requiring early defibrillation, only half ($n = 11$) limited the number of attempted defibrillations, as is best practice in cases of hypothermic arrest. Time to defibrillation (mean = 1:06 min) and time to first adrenaline (mean = 3:31 min) was also highly variable across teams (Table 1). Two teams failed to administer adrenaline at all.

Teamwork quality as reflected by the external observations also varied greatly (Table 1). The TEAM sum score and the GRS were strongly correlated ($r_{TEAM\text{-}GRS} = 0.92$, $p < 0.001$). Those teams with the highest scores on both scales were also the quickest to defibrillate and to provide adrenaline to the patient, providing further evidence for the concurrent validity of the TEAM measurement instrument (Table 2).

Team members' self-monitoring judgements of their team's teamwork were strongly correlated with both of the measures based on external observations ($r_{TEAM\_members\text{-}self} = 0.573$, $p < 0.001$, $r_{GRS\_members\text{-}self} = 0.628$, $p < 0.001$; Table 3). Interestingly, the same did not apply to the team leaders' self-monitoring of their team's teamwork ($r_{TEAM\_leader\text{-}self} = 0.347$, $p = 0.145$, $r_{GRS\_leader\text{-}self} = 0.451$, $p = 0.052$; Table 4).

Neither team size, nor number of physicians per team, nor cumulative or average experience within a team was associated with the external observations. However, teams led by younger physicians achieved better teamwork scores from external observers than those led by older physicians ($r_{TEAM\_leader\text{-}age} = -0.461$, $p = 0.047$, $r_{GRS\_leader\text{-}age} = -0.473$, $p = 0.041$; Table 4).

Table 2 Correlations of external observations, team characteristics and performance

	external observations		team characteristics			performance	
	TEAM GRS Score	TEAM sum score	team size	number of physicians	number of nurses	time to defibrillation	time to adrenaline
TEAM GRS Score	1	0.943**	0.055	−0.072	0.116	−.463**	−.217*
TEAM sum score	.943**	1	0.061	−0.070	0.120	−.451**	−.226*
team size	0.055	0.061	1	.421**	.627**	−0.048	−0.093
number of physicians	−0.072	−0.070	.421**	1	−.443**	0.193	0.202
number of nurses	0.116	0.120	.627**	−.443**	1	−.214*	−.284**
time to defibrillation	−.463**	−.451**	−0.048	0.193	−.214*	1	.486**
time to adrenaline	−.217*	−.226*	− 0.093	0.202	−.284**	.486**	1

*. Significant on a 0.05 level (two-sided); **. Significant on a 0.01 level (two-sided)

Table 3 Correlations between self-monitoring, member characteristics and external observations

	Self-monitoring	member characteristics		external observations	
		Age	ER experience	TEAM GRS Score	TEAM sum score
Self-monitoring	1	0.095	0.046	0.628**	0.573**
Age	0.095	1	.763**	−0.175	−0.140
ER experience	0.046	.763**	1	−0.162	−0.130
TEAM GRS Score	0.628**	−0.175	−0.162	1	.944**
TEAM sum score	0.573**	−0.140	−0.130	.944**	1

*. Significant on a 0.05 level (two-sided); **. Significant on a 0.01 level (two-sided)

Discussion

The performance of teams in this observational study of leadership in ad hoc teams attending to a simulated cardiac arrest with a rare cause varied greatly, with less than half of all teams making the correct diagnosis and calling for ECC. This variation should not be interpreted as a threat to the quality of actual care, because the whole idea of simulation-based training is to move participants out of their comfort zone to stimulate learning [44]. In fact, the rationale behind training for a hypothermic cardiac arrest was that this algorithm differs in important respects (e.g., dosing of adrenaline; frequency of defibrillation; early initiation of ECC) from standard advanced cardiac life support [43], and hypothermic arrests occur regularly (although infrequently) in our catchment area [37, 45, 46]. The scenario thus tested the team leader's ability to lead the team through an algorithm that differs substantially from the much more common standard advanced cardiac life support [47], thus offering a potent opportunity for learning.

This heterogeneity in objective performance across teams is reflected in both, the external observation measures of video recordings and the teams' self-monitoring, a finding that directly answers our second research question. The fact that objective performance measures of resuscitations correlate with scores on the TEAM checklist further validates this instrument for the assessment of resuscitation teams, a finding in line with previous research [9, 36]. It is interesting to note that the TEAM instrument assesses quality of teamwork, not quality of a resuscitation per se, but it seems that good teamwork is a prerequisite for good resuscitation [9, 36]. This finding, which we replicate, is reassuring, because arguably, the sole purpose of attending to patients in cardiac arrest with a team is to provide good resuscitation.

With regard to the first research question, we found moderately high agreement between the external observations and team members' self-monitoring of their teams' teamwork. These findings are well in line with the literature on self-monitoring of one's individual performance on a moment-by-moment basis [31, 35, 39, 40, 48]. Our study extends this conception in two important ways. First, it indicates that the ability to accurately self-evaluate applies to both concurrent monitoring during an event and the time shortly after this event has concluded. Second, it suggests that individuals are capable of monitoring not only their own performance but also that of a team to which they belong. Interestingly, this finding does not hold for team leaders evaluating their team's teamwork.

There are two possible explanations for the latter finding. One is that evaluating a team's teamwork as a team member is a substantially different task than evaluating this team's teamwork as the team leader. Every physician team leader carries the responsibility for their team's teamwork, because arguably a key purpose of instituting a physician team leader is to ensure good teamwork in ad-hoc teams. Given the complexity of leadership [3, 10, 12, 13, 49], particularly in highly dynamic environments such as emergency rooms, leaders may be operating at their full capacity and lack the necessary mental resources for additional meta-cognitive tasks such as self-monitoring. The diffusion of responsibility observable even in small groups [50, 51] may also lead to this

Table 4 Correlations between self-monitoring, leader characteristics and external observations

	Self-monitoring	leader characteristics		external observations	
		Age	ER experience	TEAM GRS Score	TEAM sum score
Self-monitoring	1	0.651**	0.32	0.451	0.347
Age	0.651**	1	0.660**	−0.473*	−0.461*
ER experience	0.320	0.660**	1	−0.163	−0.242
TEAM GRS Score	0.451	−0.473*	−0.163	1	0.945**
TEAM sum score	0.347	−0.461*	−0.242	0.945**	1

*. Significant on a 0.05 level (two-sided); **. Significant on a 0.01 level (two-sided)

diffused responsibility being burdened on the team leader when teams institute such a function explicitly, resulting in further leader overload. It remains an open question whether this phenomenon also affects teams led by experienced non-physicians such as advanced practice nurses. Such nurses may be less tempted than physicians to focus on things other than teamwork in their role as team leaders. Also, nurses in emergency medicine often have considerably longer experience in the field than the physicians they work with. In this study, for example, team members (i.e., mostly nurses) had about twice as much experience in emergency medicine as physician team leaders. Although we cannot exclude that professional experience plays a role in the ability to self-monitor oneself, the literature on self-monitoring suggests otherwise (e.g. [39, 40]).

Another reason for that we did not observe adequate self-monitoring in team leaders may simply be the small sample size of just 22 team leaders. This sample may be too small to achieve significance when testing the likelihood of a correlation occurring by chance. However, we did find significant correlations between team performance and leaders' age at this sample size. Sample size remains a notorious challenge in small group research [2], and this study is no exception.

Other limitations of this study result from its observational design, which was chosen to mimic real-world circumstances as closely as possible, but which rendered it impossible to control for and/or manipulate variables that may affect performance, such as team size, structure, or heterogeneity [52]. This limitation is, at the same time, one of the key strengths of the study: real-world ad hoc teams are also diverse with respect to all these variables and change on a frequent basis. The only such variable we found to affect teamwork was the age of the team leader. We can only speculate that this is because younger team leaders may have more recently received training on resuscitation guidelines, but this finding requires closer investigation. In addition, other factors that may affect teamwork such as institutional culture and environment were not varied between teams in this study, potentially limiting generalizability of our findings to environments substantially different from the one investigated here. Last, the simulation is limited by the fact that a hint was necessary to tell participants that their hands went cold during chest compression, because the simulator used cannot be programmed to feel cold.

Conclusion

Team members seem to have better insight into their teams' teamwork than team leaders, as indicated by the moderately high correlation of their self-monitoring ratings with external ratings of their teamwork. As a practical consequence, the decision to debrief after a resuscitation and the debriefing itself should be informed by team members, not just leaders. External ratings of teamwork as recorded with the TEAM instrument were substantially correlated to objective measures of team performance, a finding that further adds to the validity of the TEAM assessment instrument.

Abbreviations

ISS: Injury Severity Score; ECC: Extracorporeal circulation; TEAM: Team Emergency Assessment Measure; GRS: Global rating scale; ICC: Intraclass correlation coefficient; ANP: Advanced nurse practitioners; CRM: Crew resource management

Acknowledgements

We would like to thank all participants of the study, the team of simulation instructor at the Department of Emergency Medicine Inselspital University Hospital Bern and Susannah Goss for language editing the manuscript.

Authors' contributions

SCH, DLO, AE, TCS and WEH designed the study. SCH, DLO, and TCS collected the data, JF, JEK and WEH analysed the data, all authors interpreted the findings, SCH and WEH wrote the first version of the manuscript and all authors reviewed the manuscript and made revisions in important intellectual content. All authors read and approved the final version to be submitted for peer review.

Funding

Departmental research funds of the Department of Emergency Medicine at Inselspital Bern University Hospital. JEK has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 894536 – "TeamUp".

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request for researchers eligible to work with human data according to Swiss legislation. Eligibility will be determined by the ethics committee of the Canton Bern.

Ethics approval and consent to participate

The ethics committee of the Canton of Bern deemed the study to be exempt from full ethical review (Req-2017-00968) because it did not involve patients. All participants provided written informed consent for their data to be used in the study. As an incentive, all participants were entered in a lottery for the chance to win one of two tablet computers.

Consent for publication

Not applicable.

Competing interests

TCS has received research grants or lecture fees from Bayer, Boehringer Ingelheim, and Daiichi-Sankyo and the Gottfried and Julia Bangerter-Rhyner-Foundation. WEH has received speaker fees from the AO Foundation Zürich and research funding from Mundipharma Medical Basel. WEH is an associate editor at BMC Medical Education. All other authors have nothing to disclose.

Author details

[1]Department of Emergency Medicine, Inselspital University Hospital Bern, Freiburgstrasse 16c, 3010 Bern, Switzerland. [2]Simulated Patient Program, Office of the Vice Dean for Teaching and Learning, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany. [3]Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany. [4]Lernzentrum, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.

Received: 20 August 2019 Accepted: 19 June 2020
Published online: 23 June 2020

References

1. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. In: Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care [Internet]. Washington, D.C: National Academies Press; 2015. [cited 2017 Dec 15]. Available from: http://www.nap.edu/catalog/21794.
2. Hautz WE, Kämmer JE, Exadaktylos A, Hautz SC. How thinking about groups is different from groupthink. Med Educ. 2017;51(2):229.
3. Tschan F, Semmer NK, Gautschi D, Hunziker P, Spychiger M, Marsch SU. Leading to recovery: group performance and coordinative activities in medical emergency driven groups. Hum Perform. 2006;19(3):277–304.
4. Tschan F, Semmer NK, Gurtner A, Bizzari L, Spychiger M, Breuer M, et al. Explicit reasoning, confirmation Bias, and illusory Transactive memory: a simulation study of group medical decision making. Small Group Res. 2009;40(3):271–300.
5. Williams KA, Rose WD, Simon R. Teamwork in emergency medical services. Air Med J. 1999;18(4):149–53.
6. Rosenman ED, Ilgen JS, Shandro JR, Harper AL, Fernandez R. A Systematic Review of Tools Used to Assess Team Leadership in Health Care Action Teams. Acad Med. 2015;90(10):1408–22.
7. Leonard M, Graham S, Bonacum D. The human factor: the critical importance ofeffective teamwork and communication in providing safe care. Qual Safe Health Care. 2004;13:i85–90.
8. Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature: teamwork and patient safety. Acta Anaesthesiol Scand. 2009 Feb;53(2):143–51.
9. Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, et al. Rating medical emergency teamwork performance: development of the TEAM emergency assessment measure (TEAM). Resuscitation. 2010;81(4):446–52.
10. Tiel Groenestege-Kreb D, van Maarseveen O, Leenen L. Trauma team. Br J Anaesth. 2014;113(2):258–65.
11. Saxena A, Desanghere L, Stobart K, Walker K. Goleman's leadership styles at different hierarchical levels in medical education. BMC Med Educ. 2017;17(1):169.
12. Marks MA, Mathieu JE, Zaccaro SJ. A temporally based framework and taxonomy of team processes. Acad Manag Rev. 2001;26(3):356–76.
13. Ford K, Menchine M, Burner E, Arora S, Inaba K, Demetriades D, et al. Leadership and teamwork in trauma and resuscitation. West J Emerg Med. 2016;17(5):549–56.
14. Stroben F, Schröder T, Dannenberg KA, Thomas A, Exadaktylos A, Hautz WE. A simulated night shift in the emergency room increases students' self-efficacy independent of role taking over during simulation. BMC Med Educ [Internet]. 2016;16(1) Available from: http://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-016-0699-9, [cited 2016 Jul 19].
15. Fernandez Castelao E, Boos M, Ringer C, Eich C, Russo SG. Effect of CRM team leader training on team performance and leadership behavior in simulated cardiac arrest scenarios: a prospective, randomized, controlled study. BMC Med Educ. 2015;15:116.
16. Ryan A, Rizwan R, Williams B, Benscoter A, Cooper DS, Iliopoulos I. Simulation training improves resuscitation team leadership skills of nurse practitioners. J Pediatr Health Care Off Publ Natl Assoc Pediatr Nurse Assoc Pract. 2019;33(3):280–7.
17. Rosenman ED, Vrablik MC, Brolliar SM, Chipman AK, Fernandez R. Targeted simulation-based leadership training for trauma team leaders. West J Emerg Med. 2019;20(3):520–6.
18. Buyck M, Manzano S, Haddad K, Moncousin A-C, Galetto-Lacour A, Blondon K, et al. Effects of blindfold on leadership in pediatric resuscitation simulation: a randomized trial. Front Pediatr. 2019;7:10.
19. Roberts NK, Williams RG, Schwind CJ, Sutyak JA, McDowell C, Griffen D, et al. The impact of brief team communication, leadership and team behavior training on ad hoc team performance in trauma care settings. Am J Surg. 2014;207(2):170–8.
20. Gross B, Rusin L, Kiesewetter J, Zottmann JM, Fischer MR, Prückner S, et al. Crew resource management training in healthcare: a systematic review of intervention design, training conditions and evaluation. BMJ Open. 2019 Feb;9(2):e025247.
21. Armenia S, Thangamathesvaran L, Caine A, King N, Kunac A, Merchant A. The role of high-Fidelity team-based simulation in acute care settings: a systematic review. Surg J. 2018;04(03):e136–51.
22. Miller D, Crandall C, Washington C, McLaughlin S. Improving teamwork and communication in trauma care through in situ simulations. Acad Emerg Med Off J Soc Acad Emerg Med. 2012;19(5):608–12.
23. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa critical care medicine, high-Fidelity simulation, and crisis resource management I study. Crit Care Med. 2006;34(8):2167–74.
24. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' non-technical skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth. 2003;90(5):580–8.
25. Kolbe M, Burtscher MJ, Manser T. Co-ACT—a framework for observing coordination behaviour in acute care teams. BMJ Qual Saf. 2013;22(7):596–605.
26. DeMoor S, Abdel-Rehim S, Olmsted R, Myers JG, Parker-Raley J. Evaluating trauma team performance in a level I trauma center: validation of the trauma team communication assessment (TTCA-24). J Trauma Acute Care Surg. 2017;83(1):159–64.
27. Greif R, Lockey AS, Conaghan P, Lippert A, De Vries W, Monsieurs KG, et al. European resuscitation council guidelines for resuscitation 2015: section 10. Education and implementation of resuscitation. Resuscitation. 2015;95:288–301.
28. Morrison LJ, Neumar RW, Zimmerman JL, Link MS, Newby LK, McMullan PW, et al. Strategies for improving survival after in-hospital cardiac arrest in the United States: 2013 consensus recommendations: a consensus statement from the American Heart Association. Circulation. 2013;127(14):1538–63.
29. Freytag J, Stroben F, Hautz WE, Schauber SK, Kämmer JE. Rating the quality of teamwork-a comparison of novice and expert ratings using the TEAM emergency assessment measure (TEAM) in simulated emergencies. Scand J Trauma Resusc Emerg Med. 2019;27(1):12.
30. Mullan PC, Wuestner E, Kerr TD, Christopher DP, Patel B. Implementation of an in situ qualitative debriefing tool for resuscitations. Resuscitation. 2013;84(7):946–51.
31. McConnell MM, Regehr G, Wood TJ, Eva KW. Self-monitoring and its relationship to medical knowledge. Adv Health Sci Educ. 2012;17(3):311–23.
32. Tweed M, Purdie G, Wilkinson T. Low performing students have insightfulness when they reflect-in-action. Med Educ. 2017;51(3):316–23.
33. Moulton C, Regehr G, Lingard L, Merritt C, MacRae H. 'Slowing down when you should': initiators and influences of the transition from the routine to the effortful. J Gastrointest Surg. 2010;14(6):1019–26.
34. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA. 2006;296(9):1094–102.
35. Eva KW, Regehr G. 'I'll never play professional football' and other fallacies of self-assessment. J Contin Educ Heal Prof. 2008;28(1):14–9.
36. Cooper S, Cant R, Connell C, Sims L, Porter JE, Symmons M, et al. Measuring teamwork performance: validity testing of the TEAM emergency assessment measure (TEAM) with clinical resuscitation teams. Resuscitation. 2016;101:97–101.
37. Exadaktylos A, Hautz WE. Emergency medicine in Switzerland. ICU Manag Pract. 2015;15(4):160–2.
38. Oppenheimer DM. The secret life of fluency. Trends Cogn Sci. 2008;12(6):237–41.
39. Hautz WE, Schubert S, Schauber SK, Kunina-Habenicht O, Hautz SC, Kämmer JE, et al. Accuracy of self-monitoring: does experience, ability or case difficulty matter? Med Educ. 2019;53(7):735–44.
40. Kämmer JE, Hautz WE, März M. Self-monitoring accuracy does not increase throughout undergraduate medical education. Med Educ. 2020;54(4):320–7.
41. Valentine MA, Nembhard IM, Edmondson AC. Measuring teamwork in health care settings: a review of survey instruments. Med Care. 2014;1.
42. Field JM, Hazinski MF, Sayre MR, Chameides L, Schexnayder SM, Hemphill R, et al. Part 1: Executive Summary: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122(18_suppl_3):S640–56.
43. Vanden Hoek TL, Morrison LJ, Shuster M, Donnino M, Sinz E, Lavonas EJ, et al. Part 12: Cardiac Arrest in Special Situations: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122(18_suppl_3):S829–61.
44. Meyer O. Simulators Don't teach - Lernprozesse und Simulation. In: St. Pierre M, Breuer G, editors. Simulation in der Medizin [Internet]. Berlin, Heidelberg:

Springer Berlin Heidelberg; 2013 [cited 2019 Aug 14]. Available from: http://link.springer.com/10.1007/978-3-642-29436-5.
45. Mendonca M, Hautz WE, Schucht P, Reineke D, Kadner A, Lehmann B, et al. Extracorporeal life support in hypothermic cardiac arrest: reconsidering trauma as an absolute contraindication. Resuscitation. 2019;135:228–9.
46. Winkler B, Jenni HJ, Gygax E, Schnüriger B, Seidl C, Erdoes G, et al. Minimally invasive extracorporeal circulation resuscitation in hypothermic cardiac arrest. Perfusion. 2016;31(6):489–94.
47. Sauter TC, Iten N, Schwab PR, Hautz WE, Ricklin ME, Exadaktylos AK. Out-of-hospital cardiac arrests in Switzerland: Predictors for emergency department mortality in patients with ROSC or on-going CPR on admission to the emergency department. Schäfer A, editor. PLoS One. 2017;12(11):e0188180.
48. Pusic MV, Chiaramonte R, Gladding S, Andrews JS, Pecaric MR, Boutis K. Accuracy of self-monitoring during learning of radiograph interpretation. Med Educ. 2015;49(8):838–46.
49. Härgestam M, Lindkvist M, Jacobsson M, Brulin C, Hultin M. Trauma teams and time to early management during in situ trauma team training. BMJ Open. 2016;6(1):e009911.
50. Aronson E, Wilson TD, Akert RM, Reiss M. Sozialpsychologie. 8., aktualisierte Auflage. Hallbergmoos: Pearson; 2014. p. 678. (ps Psychologie).
51. Darley JM, Latané B. Bystander intervention in emergencies: diffusion of responsibility. J Pers Soc Psychol. 1968;8(4):377–83.
52. Hautz WE, Hautz SC, Kämmer JE. Whether two heads are better than one is the wrong question (though sometimes they are). Adv Health Sci Educ Theory Pract. 2020;6.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5 Curriculum vitae

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

6 Publications and scientific contributions

6.1 Articles

Freytag J, Stroben F, Hautz WE, Penders D, Kämmer, JE. Effect of using a cognitive aid on content and feasibility of debriefings of simulated emergencies. GMS J Med Educ (under review).

Kienle R*, Freytag J*, Lück S, Eberz P, Langenbeck S, Sehy V, Hitzblech, T. Communication Skills Training in Undergraduate Medical Education at Charité –Universitätsmedizin Berlin. GMS J Med Educ (forthcoming).

Hautz C, Oberholzer DL, Freytag J, Exadaktylos A, Kämmer JE, Sauter TC, Hautz WE. An observational study of self-monitoring in ad hoc health care teams. BMC Med Educ 2020, 20 (201). IF = 1.870

Freytag J, Stroben F, Hautz WE, Schauber, SK, Kämmer JE. Rating the quality of teamwork—a comparison of novice and expert ratings using the Team Emergency Assessment Measure (TEAM) in simulated emergencies. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 2019, 27(12). IF = 2.036

Freytag J, Hölzer H, Sonntag U. Adherence to Trained Standards After a Faculty Development Workshop on "Teaching With Simulated Patients". GMS J Med Educ 2017, 34(4): Doc45.

Hölzer H, Freytag J, Sonntag U. Faculty Development for Small-Group-Teaching with Simulated Patients (SP) - Design and Evaluation of a Competency-based Workshop. GMS J Med Educ 2017, 34(4): Doc42.

Freytag J*, Stroben F*, Hautz WE, Eisenmann D, Kämmer JE. Improving patient safety through better teamwork: How effective are different methods of simulation debriefing? A pragmatic, prospective and randomized study. BMJ Open 2017, 7:e015977. IF = 2.562

**Shared authorship*

6.2 Talks

Hitzblech T, Lück S, Eberz P, Freytag J. [„How can I become more confident in giving feedback?“ – Approaches to train and consolidate feedback skills of simulated patients]. Paper presented at: iSLS 2019. Proceedings of the 14th International Skills Lab Symposium; 2019 Mar 29-30; Neuruppin/Brandenburg an der Havel, Germany. Düsseldorf: German Medical Science GMS Publishing House; 2019. p. 1. German.

Buchmann M, Hitzblech T, Freytag J, Eberz P. [About the difficulties to classify students‘ communicative commitments to change – Categorisation of commitments to change after SP encounters]. Paper presented at: iSLS 2019. Proceedings of the 14th International Skills Lab Symposium; 2019 Mar 29-30; Neuruppin/Brandenburg an der Havel, Germany. Düsseldorf: German Medical Science GMS Publishing House; 2019. p. 1-2. German.

Freytag J, Stroben F, Hautz W, Eisenmann D, Kämmer J. How to debrief non-technical skills? – Comparison of two debriefing methods during a large emergency simulation for undergraduates. Paper presented at: 13th International Skills Lab Symposium; 2018 Mar 23-24; Maribor, Slovenia.

Freytag J, Hölzer H. [Development of a training concept to deepen simulated patients' feedback skills]. Paper presented at: GMA 2016. Proceedings of the annual conference of the German association for medical education (GMA); 2016 Sep 14-17; Bern, Switzerland. Düsseldorf: German Medical Science GMS Publishing House; 2016. p. 209. German.

Freytag J, Stroben F. [Effective teamwork? – Implementation of a feedback concept targeting team cooperation in a simulated night shift for medical students]. Paper presented at: 11th International Skills Lab Symposium; 2016 Mar 18-19; Essen, Germany. German.

6.3 Posters

Freytag J, Lück S, Buchmann M, Stroben F, Eberz P, Hitzblech T. Committing to change what do students want to improve after a peer teaching course with cases from general practice? Poster presented at: AMEE 2019. Proceedings of the AMEE annual conference; 2019 Aug 24-28; Vienna, Austria. Dundee, UK: AMEE, 2019. p. 1866.

Lück S, Freytag J, Eberz P, Hitzblech T. It's still challenging! - New concepts for training simulated patients (SP) in the art of giving feedback. Poster presented at: AMEE 2019. Proceedings of the AMEE annual conference; 2019 Aug 24-28; Vienna, Austria. Dundee, UK: AMEE, 2019. p. 125.

Freytag J, Stroben F. [Written feedback after simulation scenarios – a possibility to enhance the learning effect?]. Poster presented at: GMA 2018. Proceedings of the annual conference of the German association for medical education (GMA); 2018 Sep 19-22; Vienna, Austria. Düsseldorf: German Medical Science GMS Publishing House; 2018. p. 123-4. German.

Lück S, Eberz P, Freytag J, Hitzblech T. [Simulated patients as SP educators – a successful „train the trainer"-approach at Charité Berlin]. Poster presented at: GMA 2018. Proceedings of the annual conference of the German association for medical education (GMA); 2018 Sep 19-22; Vienna, Austria. Düsseldorf: German Medical Science GMS Publishing House; 2018. p. 122. German.

Kienle R, Freytag J, Hitzblech T. [Motivational Interviewing as teaching content in the undergraduate medical curriculum]. Poster presented at: GMA 2018. Proceedings of the annual conference of the German association for medical education (GMA); 2018 Sep 19-22; Vienna, Austria. Düsseldorf: German Medical Science GMS Publishing House; 2018. p. 185. German.

Eberz P, Freytag J, Lück S, Hitzblech T, Stroben F, Eisenmann DE. [Simulated patients' feedback and communication skills of medical students – mixed-methods-study to identify commitments to change after SP encounters]. Poster presented at: GMA 2018. Proceedings of the annual conference of the German association for medical education (GMA); 2018 Sep 19-22; Vienna, Austria. Düsseldorf: German Medical Science GMS Publishing House; 2018. p. 139. German.

Freytag J, Stroben F. Graphic feedback summaries after large simulation events – a possibility to deepen the learning experience? Poster presented at: AMEE 2018. Proceedings of the AMEE annual conference; 2018 Aug 25-29; Basel, Switzerland. Dundee, UK: AMEE, 2018. p. 97.

Lück S, Freytag J, Eberz P, Hitzblech T. Simulated patients (SP) training other SP – a successful "Train the trainer" approach at Charité Berlin. Poster presented at: AMEE 2018.

Proceedings of the AMEE annual conference; 2018 Aug 25-29; Basel, Switzerland. Dundee, UK: AMEE, 2018. p. 99.

Eberz P, Freytag J, Langenbeck S, Hitzblech T. [Train the trainer at the simulated patient programm: SPs train SPs]. Poster presented at: GMA 2017. Proceedings of the annual conference of the German association for medical education (GMA); 2017 Sep 20-23; Münster, Germany. Düsseldorf: German Medical Science GMS Publishing House; 2017. p. 161. German.

Hitzblech T, Kienle R, Peters H, Freytag J. [Training of teachers' social and communication skills with the participation of medical students]. Poster presented at: GMA 2017. Proceedings of the annual conference of the German association for medical education (GMA); 2017 Sep 20-23; Münster, Germany. Düsseldorf: German Medical Science GMS Publishing House; 2017. p. 115. German.

Freytag J, Stroben F, Hautz WE, Eisenmann D, Kämmer JE. Validation of the German "Teamwork Emergency Assessment Measure" TEAM –a study with student and expert raters. Poster presented at: AMEE 2017. Proceedings of the AMEE annual conference; 2017 Aug 26-30; Helsinki, Finland. Dundee, UK: AMEE, 2017. p. 100.

Ellermeyer T, Ziegeler K, Deselaers T, Freytag J, Dannenberg KA. [Peer teaching course „psychopathogical screening"– Influence of simulated patient encounters on students' expectation of self-efficacy]. Poster presented at: GMA 2016. Proceedings of the annual conference of the German association for medical education (GMA); 2016 Sep 14-17; Bern, Switzerland. Düsseldorf: German Medical Science GMS Publishing House; 2016. p. 232. German.

Freytag J, Stroben F. Implementing Crisis Resource Management (CRM) principles into undergraduate education using a peer-teaching approach. Poster presented at: AMEE 2016. Proceedings of the AMEE annual conference; 2016 Aug 27-31; Barcelona, Spain. Dundee, UK: AMEE, 2016. p. 382.

Freytag J, Sonntag U. [Student ratings of teaching quality in a course on communication and social skills]. Poster presented at: annual conference of the German association for medical education (GMA); 2015 Sep 30 - Oct 3; Leipzig, Germany. German.

Freytag J, Sonntag U. Quality of teaching performance: Teachers' self-assessment in comparison to learners' and observers' views. Poster presented at: AMEE 2015. Proceedings of the AMEE annual conference; 2015 Sep 4-9; Glasgow, United Kingdom. Dundee, UK: AMEE, 2015. p. 347.

Freytag J, Kampel B, Fritz AH, Strohmer R, Thrien C. [Experience report from the implementation of a new tool to assess the quality of SP feedback]. Poster presented at: 9th International Skills Lab Symposium; 2015 Mar 13-14; Leipzig/Halle, Germany. German.

Freytag J, Sonntag U. [Evaluation of faculty development workshops for teachers working with simulated patients]. Poster presented at: annual conference of the German association for medical education (GMA); 2014 Sep 25-27; Hamburg, Germany. German.

6.4 Workshops

Freytag J, Penders D. [Debriefing, (dialogical) feedback, learning conversation… it's all Greek to me!] Workshop at: iSLS 2019. Proceedings of the 14th International Skills Lab

Symposium; 2019 Mar 29-30; Neuruppin/Brandenburg an der Havel, Germany. Düsseldorf: German Medical Science GMS Publishing House; 2019. p. 19-20. German.

Stroben F, Freytag J, Dannenberg KA. [Aspects of case design for emergency medicine simulations – Realism vs. cost efficiency]. Workshop at: 12th Internationales Skills Lab Symposium; 2017 Mar 31 - April 1; Erlangen, Germany. German.

7 Acknowledgements

I want to thank everyone, who supported me in the process of writing this dissertation. Being able to study the complex and interesting topic of teamwork has been a privilege.

First, I want to thank my supervisor Prof. Dr. Adelheid Kuhlmey for her interest in my research topic, for giving me the opportunity to start this project and her support throughout the process.

I would like to express deep gratitude to Dr. Juliane Kämmer for always having an open ear and for all the helpful feedback and tips in many meetings, (skype) calls and emails. I would like to thank you for sharing your experience as a researcher and your ideas with me. It was a pleasure to have you as a co-author and mentor.

I am also very thankful to my co-author and friend Dr. Fabian Stroben with whom I started on this path of research in medical education for the great teamwork on many projects, his analytic mind and his help on many occasions. Sincere thanks to Dr. Dorothea Penders for the support from across the corridor, to Prof. Dr. Wolf Hautz for his ability to spark many peoples' – and also my interest in medical education and research, to Prof. Dr. Stefan Schauber for his support with statistical questions and to all my other co-authors: Dr. Stefanie Hautz, PD Dr. Thomas Sauter, Daniel Oberholzer and Prof. Dr. Aristomenis Exadaktylos.

This dissertation evolved out of my work at the simulated patient program, where I have, respectively had the joy of working with Susanne Lück, Dr. Rolf Kienle, Karoline Fritz, Peter Eberz, Dr. Henrike Hölzer, Dr. Tanja Hitzblech and many more, who have always supported my research interests. Additionally, I want to thank all (former) peer tutors of our local skills lab "Lernzentrum", with whom I had the pleasure to co-organise the nightshift simulations and many other projects. Especially I want to mention Dr. Katja Moschberger, David Steinbart, Jan Gerken and Dr. Hanno Heuzeroth.

I want to thank all my friends from different areas and chapters of my life for supporting me (…yes you, Big 5 + IQ!). Many thanks to my flatmates Julian and Felix for both joining my complaints about writing blocks or too much workload and taking my mind off these things. These thanks also include my quarantine flatmate Meike to whom I am grateful for the emotional support and the last proofread. I want to thank my parents and my sister Claudia for being there for me, as they have always been, and being proud of me – no matter what.