Cognition, Technology & Work (2019) 21:253–274
https://doi.org/10.1007/s10111-018-0499-6

ORIGINAL ARTICLE

Towards automated assessment of team performance by mimicking expert observers' ratings

Dennis Granåsen[1,2]

Received: 28 August 2017 / Accepted: 19 June 2018 / Published online: 29 June 2018

Abstract
Automation is the holy grail of performance assessment. Cheap and reliable automated systems that produce consistent feedback on performance. Many such systems have been proposed that accurately measure the state of a product or the outcome of a process. Procedural faults can be detected and even mitigated without the need for human interference. In production industry and professional sports, this is a natural part of business. However, in macrocognitive team performance studies, human appraisal is still king. This study investigates the reliability of human observers as assessors of performance among virtual teams, and what they base their assessments on when only able to monitor one of the team members at a time. The results show that expert observers put a lot of emphasis on task outcomes and on communication and are generally reliable raters of team performance, but there are several aspects that they cannot rate reliably under these circumstances, e.g., team workload, stress, and collaborative problem-solving. Through simple algorithms, this study shows that by capturing task scores and different quantitative communication metrics, team performance ratings can be estimated to closely match how the expert observers assess team performance in a virtual team setting. The implication of the study is that numeric team performance estimations can be acquired by automated systems, with reasonable accuracy and reliability compared to observer ratings.

Keywords Team performance · Performance assessment · Automation

1 Introduction

Automated performance assessment systems can save time and money by reducing the need for human assessors (Clauser et al. 2000), but developing such systems can be costly in itself. Reusability is therefore a key to making it worthwhile investing in automated systems for team performance assessment. In team performance research, several proposed automated performance assessment systems are deeply integrated with reusable simulators. This approach dramatically reduces the scope of team performance situations that the model needs to handle (Clauser et al. 2000; De Penning et al. 2009; Abbott et al. 2015). Outside these narrow constraints, reusable models of team performance for automated assessment are more fiction than reality. In this study, the results of an experiment are put forth, showing that a linear relationship exists between the observer ratings and recorded metrics of the teams' behaviors. It is thereby demonstrated that observers' ratings of team performance are systematic and candidate for automation. This holds true even when the observers are encouraged to go beyond task performance in their ratings. Automated team performance assessment systems may reduce the need for observers when conducting team training, which makes it a valid objective for organizations that engage in a lot of team training, e.g., emergency management institutes.

Claiming that assessment of team performance can be automated may seem controversial to some. Is it really feasible? Is there any reason to? Horace would probably have said that *grammatici certant, et adhuc sub judice lis est*[1]. It really depends on what is meant by the concept of team performance. In many scenarios, the end result is all that matters,

✉ Dennis Granåsen
dennis.granasen@foi.se

[1] Division for C4ISR, Swedish Defense Research Agency, 164 90 Stockholm, Sweden

[2] Department for Computer and Information Science, Linköpings Universitet, 581 83 Linköping, Sweden

[1] Quote from Ars Poetica (19 AD), verse 78. Roughly translates to "Critics contend, and still disputed by the judge". Sometimes simplified as: "On that point the scholars disagree".

prescribing a utilitarian perspective through which teams are measured by their productivity or the effects they cause on an external system. Such definitions have the benefit of being tangible and quantifiable and are common e.g., in production industry (Molleman and Slomp 1999) and sports (Papps et al. 2011; Scoppa 2015). Similar utilitarian perspectives have been used in the military domain to define team performance when assessing training of certain mission critical or high-risk tasks, e.g., military operations in urban terrain (Sadagic et al. 2013), aviation (Deaton et al. 2007), and cyber security training (Abbott et al. 2015; Granåsen and Andersson 2016). Automated measurement of performance in such settings is both commonplace and sound, with the caveat that for such automation to be feasible, and the training sites need to be prepared with infrastructure and technology to collect data and feed it to assessment algorithms. Automated team performance assessment can thus be a real money-saver as it allows managers to continuously monitor employees at a cheap cost, albeit not without possible moral and ethical implications (Kizza and Ssanyu 2005; Andersson 2011).

There are reasons to be more skeptical towards automated team performance assessment when infrastructure is lacking, but also when the concept of team performance itself is expanded beyond the utilitarian perspective, e.g., by including team behaviors, development, social factors, and psychological ditto. Humans have a fantastic ability of processing qualitative information and distinguishing between proper and improper behavior, but computers need clearly defined models and detectable triggers to produce the same results. It is, therefore, rare that team process metrics are incorporated in automated assessment systems in any meaningful way (Fischer and Mandl 2005; Ifenthaler 2014). Contemporary literature has very little to offer in this area, but there are some notable exceptions, like Kaur and Sood (2015) who created an experimental system for rating job performance based on game theory and data mining of multiple sensor input combined with historical data. Their results show that automated performance appraisal systems can indeed be more holistic and produce adequately accurate appraisals with less need for humans in the loop. It should be remembered that their study focused on assessing individual employees' performance as opposed to team performance. Yet, their results are interesting as the proposed performance model combines a set of indicators that are not directly tied to individual effectiveness. Their model thus serves as an excellent example of automated performance assessment with a wider scope than effectiveness. Such wider scopes are necessary in order for automation to become relevant for a wider type of team tasks. Examples include tactical training of emergency management teams and military units, where indirect factors such as risk minimization, learning, transparency, and team-building may all be just as important as the direct task outcomes (Subramaniam et al. 2010).

As with all computerized calculations, automated team performance assessment systems are only as good as their input and algorithms, and such systems can only operate on data that it has been programmed to interpret. Thus, automated team performance assessment techniques can never be as generic and flexible as human assessors, except for limited case-specific scenarios such as those typically used in microworld studies (Sapateiro et al. 2017). However, automation may cut costs for team training, since it reduces the need for expert raters, and it can improve accuracy by eliminating the random errors that human decisions inherently lead to (Clauser et al. 2000; Harik et al. 2013). Furthermore, automated performance appraisal systems can be combined with computerized feedback to create fully automated training systems (Ong 2007). There is also an objectivity argument, stemming from the fact that computers are number-crunching machines without prejudice (Gierl et al. 2014). Most automated performance assessment techniques that are in use today are case-specific (Harik et al. 2013) and based on revenue or other task-specific metrics (Yen et al. 2017).

As mentioned in the initial paragraph of this article, the objective of this study is to show that it is possible to automate team performance assessments that go beyond task outcomes. To demonstrate such a system, there has to be a baseline to compare against, something that determines what constitutes good performance. A common approach to team training is to let observers/trainers assess the performance and give their feedback during or immediately after the training session (Calvo-Merino et al. 2010). To do so, these observers/trainers sometimes have structured protocols and key indicators to look for; sometimes they do not. Regardless, observer-based team performance assessments rely on subjective evaluation of observations. Since observer ratings are so common, they serve as a reasonable baseline for automated team performance assessment systems to compare against. The critical reader might argue that this leads to a definition, wherein team performance is defined as "team performance rated highly by observers". This may seem unpractical and unorthodox at first, and even lead to circular reasoning. However, if the observers are allowed to freely interpret what constitutes good team performance, then this is actually a practical and sound definition as it allows the observers to study the team's actions and behaviors and incorporate them in their assessment to the degree that they think make sense; it is thus a way of utilizing their tacit knowledge of what team performance is instead of enforcing a theoretical construct which they may or may not agree with. This is beneficial as it is quite often the case that observers/trainers have experience from similar situations to the one they assess, and are, therefore, well equipped to determine what constitutes good performance there and then (Healey et al. 2004; Rosen et al. 2008). A potential problem with that approach is that different observers may emphasize

different behaviors, and as a consequence, this may generate organizational dissonance (Festinger 1957; Vanderhaegen and Carsten 2017).

1.1 Scope

A choice was made to scope this study to virtual teams, defined as a geographically distributed group of two or more individuals, with certain roles and specialized tasks, who collaborate primarily using electronic communication aids (Lipnack and Stamps 2000). More specifically, the studied teams collaborate in synchronous mode, meaning that they utilize in low-latency communication systems such as video and voice conferencing systems—as opposed to high-latency communication such as e-mail and messaging. The virtual team setting is motivated by: (1) controllable communication and interaction paths that can be recorded with relatively little effort and (2) a steadily increasing deployment of virtual teams replacing traditional teams in workplaces all around the world (Ferrazzi 2014), making virtual team studies highly relevant to keep team research up to date with reality.

Furthermore, the concept of team performance has so many definitions and components that it is virtually impossible to cover all (Shanahan 2001; NATO RTO HFM-087 2005). Consequently, only a select few parameters have been included to test the observers' reliability, focusing on behaviors such as coordination, information exchange and backing-up behavior. This reduction of parameters is consistent with current recommendations to reduce the workload for observers to increase their reliability (Salas et al. 2017). Long-term issues such as cohesion and trust are recognized as potentially important (Maccoun et al. 2005; Furumo and Pearson 2006; Paul et al. 2016) but have been excluded from this research for two reasons: (1) evolution of long-term team development effects are expected to be difficult to observe during the short life-span of the ad-hoc experiment teams, and (2) the focus of the observer reliability study is the circumstance under which the observers operate (unable to monitor the entire team). One might argue that a full investigation into the consequences of observers operating under these circumstances would be helpful, and that is probably true. However, as a first step, this study sets to investigate a few select team performance behaviors only.

2 Research backgrounds

Contemporary literature is somewhat confusing regarding the distinction between team performance and team effectiveness. Compare, e.g., Salas, Sims and Burke's (2005) definition "Team performance accounts for the outcomes of the team's actions regardless of how the team may have accomplished the task. Conversely, team effectiveness takes a more holistic perspective in considering not only whether the team performed (e.g., completed the team task) but also how the team interacted (i.e., team processes, teamwork) to achieve the team outcome" with the diametrically opposed view presented by Salas, Cooke and Rosen (2008) that "performance is the activities engaged in while completing a task, and effectiveness involves an appraisal of the outcomes of that activity". This study complies with the latter definition, i.e., prescribing performance as activities and effectiveness as a measure of the outcomes of those activities. Assessment of team performance is by that definition more a matter of evaluating the journey rather than the destination. More often than not, this type of team performance assessments are done by instructors and dedicated observers who compare their observations with some interpretation of appropriate behavior (Wildman et al. 2013). It is not uncommon that expert observers' team performance assessments weigh in both team processes and end results (effectiveness) as there is undoubtedly a causal relationship between performance and effectiveness (Annett et al. 2000).

2.1 Models of team performance and effectiveness

Dickinson and McIntyre (1997) proposed a framework describing teamwork as a process consisting of seven so-called "components": communication, team orientation, team leadership, monitoring, feedback, backup, and coordination. They propose that these seven components are important in any team task and recommend using checklists and numerical scales, such as behavioral observation scales and behavioral summary scales, for measuring these components. Salas et al. (2005) reviewed concurrent literature and used it to form a framework that describes the essentials of teamwork as team leadership, mutual performance monitoring, backup behavior, adaptability and team orientation. This widely cited "big five" framework specifically emphasizes that there is more to effective teamwork than sheer task-work, e.g., cooperation, communication and leadership. Although not validated, this model functions as a conceptual description of generic team performance, usable for development of performance metric frameworks, with the caveat that careful specification is needed to make such generic models applicable in actual team performance scenarios (Dickinson and McIntyre 1997).

In an attempt to integrate several popular teamwork models into one comprehensive model, Rousseau et al. (2006) analyzed differences and similarities between 29 different frameworks published between 1984 and 2005. Their concluding model specifies teamwork as a combination of team maintenance management and performance-regulating behaviors. The maintenance behaviors helps the team evolve and stay coherent, while performance-regulating behaviors are essential when performing teamwork.

Performance-regulating behaviors can be subdivided into preparation of work accomplishment, task-related collaborative behaviors, work-assessment behaviors and team adjustment behaviors. Each of these behavioral dimensions is in turn specified as a number of more tangible behaviors: mission analysis, goal specification and planning (preparation of work accomplishment behaviors); coordination, cooperation and information exchange (task-related collaborative behaviors); performance monitoring and systems monitoring (work-assessment behaviors); and backing-up behaviors, intra-team coaching, collaborative problem-solving and team practice innovation (team adjustment behaviors). There is no denying that the model is generic and as such requires further specification to become applicable in real-world scenarios. However, the specified behaviors are tangible enough to serve as a starting point for defining measurable indicators in any behavior-oriented teamwork-assessment study. While more recent meta-reviews and studies have been published on teamwork (Salas et al. 2008; Jouanne et al. 2017; McEwan et al. 2017), Rousseau et al.'s (2006) model still holds merit as one of the most tangible and descriptive models of generic teamwork.

The backbone of effective teamwork is the task-related collaborative behaviors, exemplified as coordination, cooperation and information exchange (Rousseau et al. 2006). Coordination can be defined as the sequencing, synchronization and integration of team members' activities to ensure task accomplishment (Cannon-Bowers et al. 1995; Rousseau et al. 2006). It is the result of either an explicit effort by the team members which requires some kind of interaction between the members, or something that happens implicitly as an effect of standardized procedures and shared mental models (Johnson et al. 2011; Rutherford 2017). Cooperation, by contrast, is defined as multiple team members working together on the same task accomplishment (Yeatts and Hyten 1998).

In addition to continuously monitoring their own actions, an adaptive team needs to be vigilant towards external system changes such as resource depletion, environmental changes, organizational changes and updated stakeholder requirements (Cannon-Bowers et al. 1995; Salas et al. 2005; Rousseau et al. 2006). Behaviors that enable performance adjustment include: (1) backing-up behavior, i.e., team members helping each other complete tasks (Dickinson and McIntyre 1997); (2) intra-team coaching, i.e., team members assisting each other with feedback and advice (Dickinson and McIntyre 1997); (3) collaborative problem-solving, i.e., multiple team members working actively together to diagnose and resolve a situation (Rousseau et al. 2006); and (4) team practice innovation, i.e., introducing new work practices, developing novel solutions and finding innovative ways of improving their performance (Rousseau et al. 2006).

Several other models of team performance focus on mental models, shared understanding of the situation, and workload (Cannon-Bowers and Salas 1990; Bowers et al. 1997; Kraiger and Wenzel 1997; Mohammed and Dumville 2001; Berggren and Johansson 2010; DeChurch and Mesmer-Magnus 2010; Espevik et al. 2011; Johnson et al. 2011; Mjelde and Smith 2013). It is also a well-known fact that workload and stress affect performance, both on the individual and the team level (Hart and Staveland 1988; Robert and Hockey 1997; Weaver et al. 2001). Although individuals react differently to stress, it has been identified that high stress and high workload correlate negatively not only with task-work efficiency, but also with communication and positively with teamwork (Rasmussen and Jeppesen 2006).

Countless other attributes have been tested in different models of team performance and effectiveness. Such models are often tailor-made for a specific scenario and may, therefore, be hard to generalize. Nevertheless, they might be worth studying as they represent a buffet of ideas for hungry team performance analysts to revel in. Some notable examples cover culture, creativity and collaborative practices (Yoon et al. 2010); team leader behavior (Kolb 1995); team composition and heterogeneity (Temkin-Greener et al. 2004); and trust (Martínez-Miranda and Pavón 2012). With the plethora of available performance and effectiveness indicators and metrics, it becomes even more important to keep the original objective in mind, i.e., to make sure that the final model reflects performance in the situation it is being designed for.

2.2 Automated vs manual team performance assessment techniques

Kendall and Salas (2004) identified automated performance monitoring as an established performance measurement technique, complementing the more human-oriented techniques such as self-assessments and observation-based evaluation. Automated team performance assessment is a way of reducing instructor workload, improving training quality, standardizing assessments and reducing training costs (Kendall and Salas 2004; Deaton et al. 2007). Such automated systems are typically implemented with a predetermined set of task outcomes or behaviors in mind (Ceschi et al. 2014).

There has been plenty research on automated task-based team performance assessment in micro-world studies (Brehmer and Dörner 1993; Johansson et al. 2003; Cooke et al. 2004; Dubé et al. 2011; Persson and Rigas 2014), military simulations (Martin and Foltz 2004; Frank et al. 2008; LaVoie et al. 2008) and cyber defense exercises (Brueckner et al. 2008; Geers 2010; Granåsen and Andersson 2016). Task-based metrics are highly relevant and can accurately reflect effectiveness in their specific simulated environments. However, they are not designed for capturing the cognitive

and developmental aspects of the performance, nor for evaluating team behaviors.

Automated performance assessment has also been proposed at instrumented training sites, e.g., using simple heuristics that reward predefined tactical behaviors, risk aversion, dispersion, mobility to assess the performance of military tactical maneuvers (Sadagic et al. 2013). As useful and robust as that type of evaluation system is, there is no denying that observer-based assessments are more flexible and better prepared to handle indicators that fall outside the scope of the automated system's logging capabilities, e.g., as a consequence of improvisation or unanticipated decisions.

Lawson et al. (2017) conducted a systematic review of academic literature from 1990 and later, in search for military relevant computerized solutions that can assess team performance. It is worth noting that their inclusion criteria went beyond task-based outcome metrics by covering also cognition, coordination, decision-making and resource management. Their literature search identified 571 potentially relevant articles, which after abstract-screening resulted in 57 described solutions for measuring team performance. These 57 could be further narrowed down to only seven after having sorted out those that did not describe actual assessment systems, and those that were impossible to transfer to a context even slightly different from the one they were originally tested in (Lawson et al. 2017). Despite the abundance of team performance research available, truly automated team performance assessment systems that stretch beyond assessment of task outcomes are thus rare.

The opposite approach to automation is to base performance assessments on observer ratings and self-assessments. Such evaluations are prone to subjective interpretations and bias, but their popularity reside in the ease of understanding and applying such methods, they can be used almost regardless of assessment purpose, and with very little technical resources (Wildman et al. 2013). Furthermore, the assessment can take team development and other factors that are difficult to measure into consideration, such as trust and cohesion (Riegelsberger et al. 2003; Tabassi et al. 2014; Paul et al. 2016; Alsharo et al. 2017). Observation-based and self-assessment based techniques are sometimes seen as the only ways to get insight into team-cognitive processes, and they are equally applicable to team- and individual-level constructs (Baker and Salas 1992). With team members' ratings specifically, but also observers' ratings, aggregating scores is a popular method to calculate team-level ratings, although such methods should be used with care as the theoretical rationale must be ensured before aggregating individual ratings to the team-level (Tesluk et al. 1997). It is thus important to carefully validate the team-level variables and the way they are measured before jumping to conclusions. It is quite possible, for instance, that even though most team members excel at their individual tasks, a team still fails to solve the tasks it was designed for, e.g., due to poor leadership, coordination or strategy.

Baker and Salas (1992) established early that observers may meritoriously identify and rate behavioral measures of team performance as long as the sought behavioral cues are clearly described and the measures reliable. They went even further and claimed that "there is no escaping observation" (Baker and Salas 1992). The idea that observers are necessary to evaluate teamwork was repeated 16 years later by Rosen et al. (2008) who also advised that trained observers should use structured protocols to guide them on the sought assessment criteria. Such protocols can help observers give consistent and comparable ratings and make reliable assessments, but also remind them of relevant performance aspects for the specific assessment (Rosen et al. 2008).

Team members' self-assessments is the primary alternative to observation, although the team members' perceptions are at risk of being biased, especially if they fail to interpret the situation and, therefore, are unaware of poor decisions and missed opportunities (Breugst et al. 2012). As such, team members' self-assessment of team performance have been reported to have low reliability, due to considerably lower inter-rater agreement compared to observers' ditto (Brannick et al. 1993). However, in the special case of virtual teams and in-situ observers, self-assessments can indeed correlate with on-site observers' ratings of virtual teams' performance, effectiveness and communication efficiency (Andersson et al. 2017).

Möller (2000) reported that expert observers in the field of medicine are biased by their experience and their familiarity with the rating scales they are using. Likewise he noted that the subjects (patients) can be biased by disillusioned self-conceptions. The same type of dissonance between experts and users are likely to exist in many other areas, where behaviors are rated based on subjective viewpoints, as is often the case for teamwork (Vanderhaegen and Carsten 2017). With the right tools and the right mindset, these divergent views can be combined to create new knowledge and provide a better assessment (Festinger 1957; Vanderhaegen and Carsten 2017). The classical after-action review (AAR) methodology is an example approach to facilitate such positive knowledge creation (Rankin et al. 1995; Morrison and Meliza 1999; U.S. Army Combined Arms Center 2011). It is worth noting that automation has a role also in AAR:s, with researchers striving to integrate an automatically generated baseline/ground truth to focus discussions around (Frank et al. 2008; LaVoie et al. 2008; Sadagic et al. 2013).

2.3 Team performance metrics

Team performance measurement instruments ideally account for both processes and outcomes, are targeted at specific

goals, are adapted to the context in which they are intended to be used, focus on observable behaviors, and care for both team and individual performance (Kendall and Salas 2004; Rosen et al. 2008; Marlow et al. 2018). Regardless of measurement technique, the assessments need a baseline to compare against, as well as potential indicators to observe and measure.

Information exchange, or communication, is commonly referenced as a feature of generic team performance that is measurable (Atanasova and Senn 2011; Stainback 2011; Sudhakar et al. 2011; Macht et al. 2014). Attempts have been made to semi-automatically derive metrics from communication to make conclusions on teamwork, although most such attempts have employed content-driven analysis which traditionally requires human pre-processing of communication data, e.g., coding, before it can be fed to automated analysis functions (Kiekel et al. 2001; Hetrick et al. 2002; Cooke et al. 2004). Research on speech recognition and semantic analysis is rapidly pushing the limits with the advance of virtual personal assistants, enabling far more advanced automated analysis than was possible just a few years ago (Këpuska and Bohouta 2018). With the advance of sophisticated speech-recognition technology, automatic communication transcription and coding, e.g., with respect to emotion, is becoming increasingly more feasible (Wang et al. 2018). Once technology allows, communication-based metrics may add another dimension to automated team performance assessment systems (Foltz et al. 2003; Lavoie et al. 2008; Stein et al. 2013).

As there are established metrics for both workload (physical and mental) and stress, it is not uncommon that they are included in team performance metrics, if nothing else as a baseline for interpreting other results (Wildman et al. 2013). Continuous assessment of the team's work enables them to track their progress towards reaching their goals (Salas et al. 2005), but also to monitor stress and workload to enable performance adjustments (Rasmussen and Jeppesen 2006).

For understanding the inner processes of the team, particularly rapid response ad-hoc teams, a research field has emerged on macrocognition in teams, building on Hutchins' (1995) work on distributed cognition. The aim is to, among other things, map individual cognitive processes to team-based ditto to identify gaps, overlaps, and unique features of cognition in teams vs individuals (Letsky and Warner 2008). This research can coarsely be divided into two directions: the shared cognition path and the interactive team cognition path (Berggren 2016). The shared cognition direction assumes that there are, at any given time, values and knowledge which need to be shared among team members. From this assumption, shared team cognition can be measured as snapshots by estimating the level of agreement among the members (Langan-Fox et al. 2001; MacKenzie et al. 2007; McComb 2008). The trick is to identify what perceptions the team members need to share, and consequently are relevant to measure. The team interaction perspective, on the other hand, is based on the reasoning that the team cognition exists within the team itself, as processes and interactions between its members (Cooke et al. 2013). Team interaction measurement methods typically employ standard communication and social network analysis techniques (Soós and Juhász 2011; Andres 2013).

2.4 Virtual teams

While team performance and effectiveness have been much researched for several decades—it is in the twenty-first century that virtual teams have gained attention of researchers worldwide. A lot of the research on traditional teams has been confirmed applicable through comparative studies that examine the relationship between different aspects of traditional team performance models and virtual teams, e.g., communication (Berry 2011; Morgan et al. 2014), team composition/configuration (Turel and Zhang 2010), and shared mental models (Espevik et al. 2011; Maynard and Gilson 2014). However, the asynchronous nature of many virtual teams renders traditional work patterns obsolete (Berry 2011) and adds an increased correlation between explicit knowledge sharing and effectiveness (Pangil and Chan 2014). It is, therefore, not surprising that synchronous collaboration technologies have been found beneficial to virtual teamwork (Baker 2002), and likewise that collaborative support tools like shared digital whiteboards have been found to instrumental to improve collective decision-making in synchronous virtual teams (Curtis et al. 2017).

The physical distance and dependency on information and communication technology (ICT) make virtual teams stand out from traditional teams (Berry 2011). The relationship between communication and performance in virtual teams is not easily understood as demonstrated by Chang, Hung and Hsieh (2014). In their study they found evidence that high communication quality can negatively impact performance, perhaps because the team members are discouraged and become suspicious when team members are paying too much attention to details that they themselves think are easily settled. Another explanation may be that coherent teams that have developed a high level of interpersonal trust, experience lesser need for elaborate communication (Jarvenpaa et al. 2004).

Trust seems to be more difficult to establish in virtual teams than in traditional work teams, perhaps because the lack of physical interaction (Jarvenpaa et al. 2004). Time and resources spent on building trust relationships are, therefore, well invested to create virtual teams that collaborate effectively (Ford et al. 2017); however, other studies have found conflicting evidence suggesting that trust does not necessarily increase team effectiveness (Alsharo et al. 2017).

Virtual teams also have a tradition of having less formal leadership structures compared to traditional teams, although some sort of emergent structures, e.g., information leadership, can often be observed and have a positive impact on team effectiveness (O'Mahony and Ferraro 2007). Ziek and Smulowitz (2014) report that the leadership that seems to have the most positive effect on virtual teams' effectiveness is characterized by the ability of asking the right questions, setting and communicating goals and visions, in addition to the ability of demonstrating insight and imagination.

A more holistic approach has been proposed to compensate for the lack of observation opportunities in the cyber domain, by triangulating different assessment methods such as observer ratings, self-assessments and outcome-based task scores (Granåsen and Andersson 2016). Findings highlight the problems of relying too much on either approach as they all say something about the team's performance but independently neither say enough. Such findings are in line with the general advise to always have a clearly stated objective and adapt the measurement models carefully to the domain when engaging in team performance assessment (Kendall and Salas 2004; Rosen et al. 2008).

2.5 Synthesis and research gap

The presented research background reveals that team performance is a thriving research field with plenty of models that try to explain team performance, and an ever-growing set of measurement approaches for different scenarios. A recurring feature of a large set of the explored research is that team performance evaluations are conducted through observer ratings and self-assessment reports (Wildman et al. 2013). These techniques have been well researched and can now be considered mature.

Automated team performance assessment literature advocates the use of computer programs to calculate performance ratings. An immediate effect of automated performance ratings is the reduced need for observers, which may prove cost-effective for organizations that engage in a lot of team performance reviews. Another benefit is that algorithms can produce deterministic and reliable ratings, however, that assumes that models and sensor data perfectly describe the team performance, which may seem like a utopian dream. Therefore, it is not surprising that literature on automated performance rating systems focus on simple performance models and metrics, such as measures of effectiveness calculated by training simulators (Lawson et al. 2017). The consequence is that most proposed automated team performance rating systems have low fidelity, focus only on task effectiveness, and are unable to account for unanticipated events in their respective reviews (Dorsey et al. 2009). Ideally, automated systems should produce ratings of team performance that account for attitudes, behaviors and cognition as well as effectiveness. As there are no gold standards for how to rate team performance, generating ratings that are on par with observer ratings is a reasonable first goal. To date there has been very little research published on this topic.

To conduct such research, a natural first step is to determine the baseline, i.e., how the observers rate team performance. Research suggests that observers should review attitudes, behaviors and cognition, both on team- and individual levels, when rating team performance as these are an integral part of performance (Salas et al. 2017). Having done that, expert observers are able to compare their observations with their expectations and rate the observed team performance. While this is certainly possible in many contexts, the virtual team format does not lend itself to observation as easily since team members may be both physically and temporally separated from each other. The observers thus have to choose whether to focus their observations on a part of the team, or to rely on remote techniques such as video feeds, to attempt to observe all sub-teams. Both methods are likely to impede the observers' ability of assessing the team performance, yet to date there is little or no research published on this particular topic that can help establish which observation strategy is better for virtual team observation.

3 The study

The central research gap concerns how to broaden the scope of automated team performance assessment systems. To fill that gap is a huge challenge that goes well beyond the scope of what this study can achieve. However, a small part of it can be approached by testing one methodology to broaden the scope. This study employs an approach to let observers rate virtual team performance to create a baseline to compare automated rating systems against. As task score is the most commonly used metric for automatic team performance analysis, this was used as the benchmark to beat. The tested automation model consists of a simple linear regression on easily captured team performance indicators that complemented the benchmark model.

To follow the advice of Salas et al. (Salas et al. 2017), the observers' rating protocols were kept small with questions centered around workload, backing-up behavior, stress, coordination, task-work efficiency, collaborative problem-solving and information exchange. All these performance indicators have been selected from the performance models discussed in the previous section, based on expected applicability to generic virtual teamwork scenarios and ease of measurement.

The validity of using observer ratings as a baseline is dependent upon their reliability. If their ratings are unreliable, then they make no sense to use as a baseline. In this study, the observers operated in-situ, one observer per

team site. An alternative approach would be to use off-site observers using remote monitoring technology to follow the teamwork at all locations simultaneously. Ideally both methods should be tested and compared. However, due to budget restrictions only the in-situ mode could be tested in this study.

As task effectiveness is the norm of today's automated team performance assessment systems that would be the natural baseline for any attempt to improve the predictive power of an automated system. In a virtual team setting, the only way the team members can collaborate is by communicating through the provided communication tools. Therefore, it makes sense to capture and integrate that communication into the assessment model. In fact, that may be the only thing that matters, as it is the only means by which the team members interact in a virtual team. Hence, the quantitative metrics are all centered around effectiveness and communication.

A simple statistical approach was then employed to create a model that matches the observer ratings with a reasonable accuracy. It is well acknowledged that any such model would be highly contextual and quite uninteresting outside the narrow scope of its' applicability. For this reason, there is a value in using simple models and cheap technology to make the approach a cost-effective alternative to observer-based ratings.

A side effect of virtual teams communicating via ICT is that the communication is already encoded into signals, meaning that recording and analyzing it is a matter of tapping into the communication systems. In a controlled environment this is normally very easy to do, making this data capture cheap. The automated analysis, however, is a different story. Many quantitative communication analysis techniques are based on communication coding, i.e., sorting each utterance/message into one or more categories. The idea is that the number of collected samples in each category says something about the subject at hand. However, no such automatic speech classification technology was available for this study due to the limited resources available, instead this study employs a very simple communication analysis strategy by counting the number of utterances and messages in each performance without trying to analyze their content, combined with the accumulated total time used for speaking. The rationale behind these metrics is that communication is essential to create shared mental models which in turn affects the effectiveness of virtual teams (Maynard and Gilson 2014; Schmidtke and Cummings 2017). An underlying assumption is that all or most communication is directly relevant to the team's performance, i.e., an insignificant amount nonsensical communication. However, one must be careful with such metrics as it is recognized that if the team spends all time communicating then they may not have time to solve their tasks. Albeit these crude metrics have been recognized as potentially problematic, they were chosen because of being easily accessible with available technology and deemed interesting enough. A positive spin on these limitations is that simple solutions are often cheaper to acquire—lowering the threshold for organizations that look to automate their team performance assessments but shun the investment costs. To summarize, this study addresses two research questions:

RQ1 How reliable are human observers' ratings of teamwork in virtual teams, when only able to monitor team members at one site?

RQ2 To what extent can the observers' ratings of team performance be replicated with enough accuracy using only automatically quantifiable metrics?

4 Method

An experimental study was designed to let in-situ observers monitor a virtual team's performance and compare their assessments with an objective task score. The scenario was designed to give each team member a specific role and a unique set of clues needed to solve the task, restricting their communication to the provided tools for audio–video and textual chat. As such, the setup rewarded communicative and collaborative teams and punished teams that relied upon individual problem-solving. The tasks revolved around chemical, biological, radiological and nuclear (CBRN) incidents, simulated through scripted roleplaying by the experiment facilitators. Each observer was only able to directly monitor the performance of one team member per performance session. All assessments on the team level, therefore, had to be done by interpreting that team member's interactions with the team.

The study setup was designed to impose ecological validity through medical accuracy and realistic scenarios. The virtual teams each consisted of one forward agent (the coordinator) and two remote reachback experts (the medical experts) intended to mimic the setup experimented with, e.g., for remote radiation detection (Bordetsky et al. 2007) and robot-assisted humanitarian search and rescue operations (Murphy et al. 2004). The following sections present a brief overview of the experiment. A more thorough write-up of the experiment setup, scenario, challenges, participants, and data collection has been presented in another study based on the same data set (Andersson et al. 2017). The following section summarizes the methodology from the sister article, and complements it by adding information relevant for this study only, particularly regarding analysis.

Table 1 Description of rating items in the observers' team performance assessment protocol

#	Question	Behavior(s)/team performance correlate(s)
q_1	High individual workload	Workload
q_2	High team workload	Workload
q_3	Evenly distributed workload	Coordination, backing-up
q_4	Low stress level for the subject	Stress
q_5	Low stress level for the team	Stress
q_6	Effective teamwork	Backing-up, coordination, task-work efficiency
q_7	Equal participation	Collaborative problem-solving
q_8	Active in decision-making (DM) process	Collaborative problem-solving
q_9	All members active in DM	Collaborative problem-solving
q_{10}	The team was wasting time	Task-work efficiency
q_{11}	Team coordination	Coordination
q_{12}	Clear communication	Information exchange
q_{13}	Efficient communication	Coordination, information exchange
q_{14}	Team performance	Overall

4.1 Experiment setup

The experiment consisted of eight teams performing six different CBRN-related challenges each. The challenges were grouped in sets of two with each pair representing one CBRN-scenario and consisting of one diagnosis followed by one treatment challenge. Both types of challenges revolved around the same type of problem with the team members needing to share information and solve a logical deduction problem to find the correct diagnosis and treatment instructions, respectively. From a problem-solving perspective, all challenges were very similar. All teams were composed of three members located at different sites and communicating only through online collaboration tools. Two of the team members were nursing students, given the task to act as medical reachback experts. One military student acted as the forward operator facing medical challenges for which it required the assistance of the medical experts. One observer was stationed at each location, given the task of monitoring teamwork by observing only the one team member at that location. The same three observers rated the 19 first team performances, after which one of the observers had to be replaced for logistical reasons. After that substitution there were no more changes in the observer lineup for the remainder of the experiment.

The in-situ observers were guided by a protocol consisting of 14 items to rate on a 5-point Likert scale. Each of the 14 items correspond to one identified behavior or team performance correlate, as presented in Table 1. It is assumed that the relative distances between each step on the rating scale are equal, thus that the scale is linear from 1 to 5. Additionally, the observers were given the task to make notes of just about anything that they thought could be noteworthy in relation to the performances. The mean observer ratings ($\bar{q}$) on each question were used as the team-level ratings. Items 1–13 were selected as a representation of reasonable metrics of team performance correlates and behaviors, based on the presented research background. The rating items were taken from Bushe and Coetzer's (1995) survey instrument and the Crew Awareness Scale Only (McGuinness and Foy 2000) and edited after an initial pilot test. The full wordings of all rating items are listed in the sister study from the same experiment (Andersson et al. 2017). Only rating items that relate to behaviors and metrics that were thought to be relevant for the designed task were included in the final protocol, including workload (q_1, q_2), backing-up behavior (q_3, q_6), stress (q_4, q_5), coordination (q_3, q_6, q_{11}, q_{13}), task-work efficiency (q_6, q_{10}), collaborative problem-solving (q_7, q_8, q_9), and information exchange (q_{12}, q_{13}). It is worth noting that this selection is not meant to be a comprehensive list of team performance indicators but rather a sample of what team performance analysts may want to study, and what the pilot testers of this study thought they would be able to rate. The final item (q_{14}) in the rating protocol is an overall rating of the team's performance "all things considered". The observers were instructed to rate this as they saw fit after having observed the performance. The team mean of this overall performance rating ($\bar{q}_{14}$) serves as the baseline performance rating for each completed team challenge.

Upon completion of the challenge, the teams were instructed to report their diagnosis, and if they were not certain, up to two alternative diagnoses. From a cognitive perspective task outcome is just one of many dimensions of team performance (Granåsen and Andersson 2016). It cannot be denied though that the outcome is often the reason for the team performing anything at all, and as such it is hard to imagine a performance assessment method that completely neglects it. Therefore, a metric was designed to capture the task outcome with slightly higher accuracy than just success or fail. The task outcome (x_1) for the diagnosis challenge was

specified as $10-3\times n$ points for correct diagnosis, where n corresponds to the number of alternative diagnoses that the team were not able to completely rule out. If the primary diagnosis was incorrect, but the correct diagnosis was mentioned among the alternative diagnoses, then the team was awarded $3-n$ points. Failure to name the correct diagnosis at all resulted in 0 points. For the treatment challenges, the task score was calculated as $10-n$, were n corresponds to the number of mistakes noted by the experiment controllers. A mistake was defined as failure to complete one step of the treatment program or performing a treatment step that was not part of the correct program.

As the motivation for this study was to identify generic approaches for automated team performance assessment, the task outcome metric (which is always contextual) was complemented with easily quantified constructed metrics that can be captured with contemporary technology. Communication frequency was selected as a candidate metric as it has been shown to correlate with team performance outcomes in a military flight simulator training experiment (Brannick et al. 1993). In the virtual team experiment subject to this study, the team's geographical distribution and dependency on online collaboration tools for coordination make it natural to include information exchange as one of the primary performance indicators. Four quantitative metrics were designed around communication frequency. For each challenge the number of spoken utterances (x_2) were counted and recorded as a measure of intra-team communication, using digital recordings captured by the online communication tool. An utterance was defined as a continuous stream of words and phrases spoken by one individual conveying one message. An utterance was classified as completed when interrupted by another individual, being followed by a notable pause or the speaker changed subject. The total time the team spent talking (x_3) was also measured during each challenge. As the teams were also allowed to communicate via textual chat, the number of written messages (x_4) sent over the chat were counted during each challenge. A written message was defined as one chunk of text sent in one batch. A combined indicator was constructed by summing the total number of spoken and written messages (x_6).

In addition to the task outcomes, the team's efficiency was measured by registering the time to completion (x_5) using a stopwatch (or failure to complete). For the diagnosis challenges, this time corresponds to the duration from when a team was no longer allowed to interact with the patient, until consensus was reached on a primary diagnosis. For the treatment challenges, the time corresponds to the duration from when the treatment started until it was completed.

The objective performance metrics are summarized in Table 2 below. The metrics were designed to be as non-intrusive as possible towards the team members to gain increased acceptance, and consequently no physiological instruments were allowed. Furthermore, although highly relevant, data stream processing techniques such as automated speech analysis, voice pitch analysis, and video analysis were excluded as the researchers did not have access to any such tools with enough accuracy to be useful for the intended purpose.

Table 2 Objective performance metrics

#	Metric
x_1	Outcome-based task score
x_2	Number of spoken utterances/messages
x_3	Accumulated time for spoken communication
x_4	Number of written messages
x_5	Total time for completing challenge
x_6	Total number of spoken and written messages

All designed x variables are objectively measurable and as such are undisputable in terms of correctness. The captured metrics have been linearly scaled and translated to the same 1–5 scale as the observer ratings used for comparability, where 1 corresponds to the lowest measured value and 5 to the highest.

As the six challenges were designed to be similar and comparable, the teams were expected to learn and improve their performance with each iteration. To balance for this learning effect, the scenario ordering was permutated differently for the teams, with the limitation that treatment challenges always followed after a diagnosis challenge. As there were only six possible permutations and eight performing teams, a perfectly balanced experiment could not be created. The order in which the scenario and challenges were administered for each team is presented in Table 3. To further limit the learning effect, the scoring system was not explained to the teams—and their scores were never revealed to them. Instead the teams were given instructions to solve their tasks in any way they saw fit, using only the provided materials and communication tools.

4.2 Analysis methodology

The distribution of all observer ratings, task scores, communication and interaction metrics were first investigated through descriptive statistics, with calculated means, standard deviations, skewness and kurtosis. Thereafter, x_1 and $\bar{q}_{14}$ were checked for correlation with all other metrics (x_2–x_6 and $\bar{q}_1$–$\bar{q}_{13}$) to get an indicator to which teamwork aspects are reflected in the task scores and in the observers' overall ratings of team performance, respectively. The task score metric (x_1) reflects the utilitarian perspective, where only the outer effects are considered, while the mean observers' overall ratings ($\bar{q}_{14}$) represent the opposite view that team performance is best estimated using expert observers' tacit understanding of what it is.

Table 3 Order of scenarios and challenges for each team

#	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
1	Cd	Cd	Nd	Rd	Rd	Nd	Cd	Rd
2	Ct	Ct	Nt	Rt	Rt	Nt	Ct	Rt
3	Nd	Rd	Rd	Cd	Nd	Cd	Rd	Nd
4	Nt	Rt	Rt	Ct	Nt	Ct	Rt	Nt
5	Rd	Nd	Cd	Nd	Cd	Rd	Nd	Cd
6	Rt	Nt	Ct	Nt	Ct	Rt	Nt	Ct

Scenarios: *C* cyanide, *N* nerve gas, *R* radiation scenario

Challenge types: *d* diagnosis, *t* treatment

Since high consistency between different raters is an indicator of high reliability, RQ1 was approached by calculating the inter-rater agreement between the observers. The calculations used the two-way mixed-effects consistency model for the intraclass correlation coefficient (Koo and Li 2016). The mixed-effects model was preferred over random effects, since the raters were handpicked. As one of the raters was replaced midway through the experiment, the data set had to be partitioned for the reliability calculations.

RQ2 was approached by creating linear regression models of the collected metrics (x_1–x_6) towards the mean observer score $\bar{q}_{14}$ for each performance. The constructed metrics were designed to be automatically quantifiable and related to task outcome, time efficiency and information exchange quantity. To explore if these metrics can say something about overall team performance, they were all fitted against $\bar{q}_{14}$ using multiple linear regression after the outliers had been removed, determined by Cook's distance (Cook 1977) exceeding three times the mean ($D_i > 3 \times \bar{D}$). As the study was explorative, an alternative outlier heuristic was thereafter designed as the result of careful inspection of the data to find discrepancies that were not reflected by Cook's distance.

All combinations including x_1 and one or more of x_2–x_6 were fitted against $\bar{q}_{14}$ and compared using the adjusted R^2 fitness criterion. The appropriateness of the selected model was determined by correlating its predicted values with actual $\bar{q}_{14}$ ratings using Pearson's *r*. It has been recognized that there are no guarantees for such relationships to exist, and also that there might be non-linear relationships, e.g., logistic and polynomial. However, linearity was preferred for simplicity's sake.

5 Results

Of the 48 recorded team performances, two were deleted due to methodological errors. The remaining 46 samples show that the teams generally performed well on completing their challenges according to the objective task-based performance metric x_1 ($M = 7.50$, SD = 3.17 on the scale from 0 to 10). For further analyses, the scale has been transformed to align with the 5-point scales used for the observer ratings ($M = 4.00$, SD = 1.27). The mean observer ratings for each challenge ($\bar{q}_{14}$) were slightly more conservative ($M = 3.76$, SD = 0.89). Thus, there is a small offset between the two scales. Pearson's correlation coefficient between x_1 and $\bar{q}_{14}$ is $r(44) = 0.51$, $p < .01$, confirming that observer ratings of virtual team performance correlate positively with successful task performance.

Table 4 summarizes the means (*M*), standard deviations (SD), standard error of the mean (SEM), skewness (Skew) and excess kurtosis (ExKu) of the mean observer ratings ($\bar{q}_n$) and performance metrics (x_n). Table 4 shows that most of the distributions are left-skewed (Skew < 0), and several also draw towards being platykurtic (ExKu < 0). Identified skewed distributions are classified as non-normal when the magnitude exceeds the double standard error for skewness (SES), i.e., 0.70 for $N = 46$. The same method has been used to determine non-normal kurtosis, i.e., comparing the magnitude to the double standard error for kurtosis (SEK), which is 1.38 for $N = 46$. The statistics show that most distributions fall within the criteria for normal distributions. The challenge completion time, however, stands out as the only one with non-normal kurtosis. Additionally, the workload variables are right-skewed, and observer ratings of team performance and the task performance score are left-skewed. The variables on communication efficiency and clarity are left-skewed.

All variables have been compared for correlation with $\bar{q}_{14}$ and x_1, see Table 5. The table shows that all observer ratings correlate moderately or strongly with their overall team performance assessment ($\bar{q}_{14}$). There is also moderate correlation between the communication and coordination variables ($\bar{q}_6$, $\bar{q}_{11}$, $\bar{q}_{12}$ and $\bar{q}_{13}$) and the task score (x_1). The stress variables ($\bar{q}_4$ and $\bar{q}_5$) and individual workload ($\bar{q}_1$) correlate moderately with the task score. While communication is undoubtedly the primary means to solve the challenges given, the identified negative correlations between x_1 and x_2, as well as between x_1 and x_6, suggest that too many message exchanges could also be an indication of disagreement, failed problem-solving, ineffective teamwork, or possibly that the team put more emphasis on strategizing

Table 4 Descriptive statistics of mean observer ratings, task scores, communication and interaction metrics

#	Question	*M*	SD	SEM	Skew	ExKu
$\bar{q}_1$	High individual workload	3.12	0.55	0.08	1.05*	0.83
$\bar{q}_2$	High team workload	2.99	0.58	0.09	0.81*	0.91
$\bar{q}_3$	Evenly distributed workload	3.48	0.69	0.10	−0.61	−0.74
$\bar{q}_4$	Low stress level for the subject	3.03	0.68	0.10	0.10	−0.91
$\bar{q}_5$	Low stress level for the team	3.17	0.66	0.10	0.03	−0.60
$\bar{q}_6$	Effective teamwork	3.62	0.95	0.14	−0.35	−1.31
$\bar{q}_7$	Equal participation	3.51	0.79	0.12	−0.52	−0.45
$\bar{q}_8$	Active in DM process	3.88	0.58	0.09	−0.03	−0.76
$\bar{q}_9$	All members active in DM	3.71	0.70	0.10	−0.59	−0.15
$\bar{q}_{10}$	The team was wasting time	3.78	0.78	0.11	−0.45	−0.97
$\bar{q}_{11}$	Team coordination	3.57	0.91	0.13	−0.33	−0.97
$\bar{q}_{12}$	Clear communication	3.59	0.95	0.14	−0.55	−0.66
$\bar{q}_{13}$	Efficient communication	3.59	0.93	0.14	−0.50	−0.61
$\bar{q}_{14}$	Team performance	3.76	0.89	0.14	−0.77*	−0.36
x_1	Task score	4.00	1.28	0.19	−1.29*	0.51
x_2	Number of utterances	3.13	0.78	0.11	−0.43	0.77
x_3	Communication time	3.09	0.76	0.11	−0.42	0.87
x_4	Number of written messages	1.82	1.23	0.18	1.33*	0.65
x_5	Challenge completion time	4.00	0.92	0.14	−1.91*	2.77*
x_6	Total number of messages	2.71	0.92	0.14	0.44	0.12

* $|\text{Skew}| > 2 \times \text{SES}$ or $|\text{ExKu}| > 2 \times SEK$

Table 5 Pearson's correlation coefficients, *r* (44), between mean observer ratings of team performance ($\bar{q}_{14}$), task scores (x_1) and all other team-level metrics ($\bar{q}_{1-13}$ and x_{2-4})

	$\bar{q}_{14}$	x_1
$\bar{q}_1$	−0.52***	−0.37**
$\bar{q}_2$	−0.48**	−0.24*
$\bar{q}_3$	0.68***	0.27*
$\bar{q}_4$	0.64***	0.33**
$\bar{q}_5$	0.66***	0.33**
$\bar{q}_6$	0.93*****	0.45**
$\bar{q}_7$	0.79****	0.24*
$\bar{q}_8$	0.66***	0.23*
$\bar{q}_9$	0.80****	0.27*
$\bar{q}_{10}$	−0.83****	−0.40**
$\bar{q}_{11}$	0.93*****	0.48**
$\bar{q}_{12}$	0.93*****	0.46**
$\bar{q}_{13}$	0.95*****	0.49**
x_2	−0.14	−0.46**
x_3	0.13	−0.15
x_4	−0.10	−0.04
x_5	−0.21*	−0.31**
x_6	−0.23*	−0.57***
Correlation strengths		
*Weak		$0.15 \le \lvert r \rvert < 0.30$
**Low		$0.30 \le \lvert r \rvert < 0.50$
***Moderate		$0.50 \le \lvert r \rvert < 0.70$
****High		$0.70 \le \lvert r \rvert < 0.90$
*****Strong		$\lvert r \rvert \ge 0.90$

and team-building. At this level of analysis, determining the actual reason behind the obtained results would require thorough content analysis of the recorded communication.

5.1 Observer assessment: inter-rater agreement

Table 6 shows the inter-rater agreement among the observers rating team performance. The table has been split into two halves due to the replacement of one observer (cases 1–19 and cases 20–46, respectively). The top of the table displays the degree of freedom (*df*) for each data set. The between-judge df is calculated as $n-1$, while the residual *df* is $(n-1)\times(k-1)$, where n is the number of cases per data set (19 and 27, respectively) and k is the number of raters (3) (Shrout and Fleiss 1979). Questions q_1, q_4 and q_8 have been excluded from this summary, since these ratings concern individuals and not teams, thus the observers' ratings for these questions cannot be compared as they were in fact rating different individuals. The two-way mixed-effects consistency is reported once per data set for each remaining question. From the low agreement on team workload (q_2) it is apparent that the observers were unable to deduce a coherent view while distributed and only able to directly observe one of the team members. They seemed to have struggled also with team stress, time utilization and team member participation (q_5, q_7, q_9 and q_{10}). On communication- and coordination-related questions (q_6, q_{11}, q_{12} and q_{13}) on the other hand, the observers' inter-rater agreement was

Table 6 Two-way mixed average measures of intraclass correlation (ICC3k)

		Cases 1–19 *df* (18,36)	Cases 20–46 *df* (26,52)	Overall
q_2	High team workload	0.42*	−0.60	Poor
q_3	Evenly distributed workload	0.69**	0.63**	Good
q_5	Low stress level for the team	0.46*	0.44*	Fair
q_6	Effective teamwork	0.91***	0.76***	Excellent
q_7	Equal participation	0.50*	0.61**	Fair
q_9	All members active in DM	0.74**	0.57*	Fair
q_{10}	The team was wasting time	0.83***	0.58*	Fair
q_{11}	Team coordination	0.81***	0.81***	Excellent
q_{12}	Clear communication	0.88***	0.77***	Excellent
q_{13}	Efficient communication	0.86***	0.71**	Good
q_{14}	Team performance	0.85***	0.80***	Excellent

Intraclass correlation coefficient strengths (Cicchetti 1994)		Overall rating	
*Fair	$0.40 \le ICC < 0.60$	Fair	Both > 0.4
**Good	$0.60 \le ICC < 0.75$	Good	Both > 0.6
***Excellent	$0.75 \le ICC$	Excellent	Both > 0.75

Data has been partitioned into two subsets, due to the replacement of one rater after case 19. The lowest of the two ICC values was used for the overall rating

high, as for the questions on overall team performance (q_{14}). The study has thereby confirmed that for ratings related to communication, coordination and overall virtual team performance, the in-situ observer ratings are reliable, but for workload ratings they are not.

5.2 Outlier detection

Figure 1 presents Cook's distance (D) between the model-predicted values for each case and the corresponding observer ratings ($\bar{q}_{14}$). With the selected cutoff at $D_i > 3 \times \bar{D}$, there are two outliers: cases 7 and 19.

Table 7 shows that both outliers are from a team's first diagnosis challenge. Team 2 (case 7) seemed to have problems with teamwork and communication ($\bar{q}_6$, $\bar{q}_7$, $\bar{q}_{11}$, $\bar{q}_{12}$ and $\bar{q}_{13}$ are all low), yet the team successfully solved the challenge with a perfect score. This phenomenon is hard to explain, but clearly the tasks were solvable with a less than optimal approach to communication. Team 4 (case 19) has the opposite relationship, with fairly good ratings but lowest possible task score.

Unfortunately, the observers did not record any comments to their observations on case 7, however, for case 19, they gave comments, as listed in Table 8:

The observer comments suggest that the team performed ok to get all the necessary pieces of information, but had problems realizing how to combine the information to solve the challenge. The observers acknowledged the partial success and gave average scores thanks to good collaboration on information management, but the objective performance metric dismissed their performance due to the poor collaborative problem-solving.

5.3 Alternative outlier heuristics

Only five of the diagnosis challenges received (normalized) task scores less than two (cases 9, 19, 25, 29 and 41). Manual inspection of the data reveals that the observers generally did not penalize poor task-work. In fact, three of these five were rewarded above-average ratings on $\bar{q}_{14}$, interestingly four are the first performance iteration of a team which suggests that the observers has lower expectations on the teams for their first performances. Task scores and observer ratings thus correlate badly at low task scores ($x_1 < 2$). Further investigation reveals that there are additional cases that scored below average on task score, but were rewarded above average on observer rating. For values near the average score this is not very spectacular, but two of these cases (24 and 42) had an absolute difference larger than one meaning that the observers found their performance acceptable despite the low task scores. On the other end of the spectrum there were four cases (2, 7, 28 and 37) that received task scores above four but were rewarded below average by the observers, and thereby fulfilling the condition of absolute difference larger than one. Three of these four got the maximum possible task score. In these cases, the observers thus did not reward the teams for their achievements. However, in total, there were 32 cases with a task score greater than four, and most of them were rewarded by the observers so the four outliers do stand out. The explanation could be that the tasks were

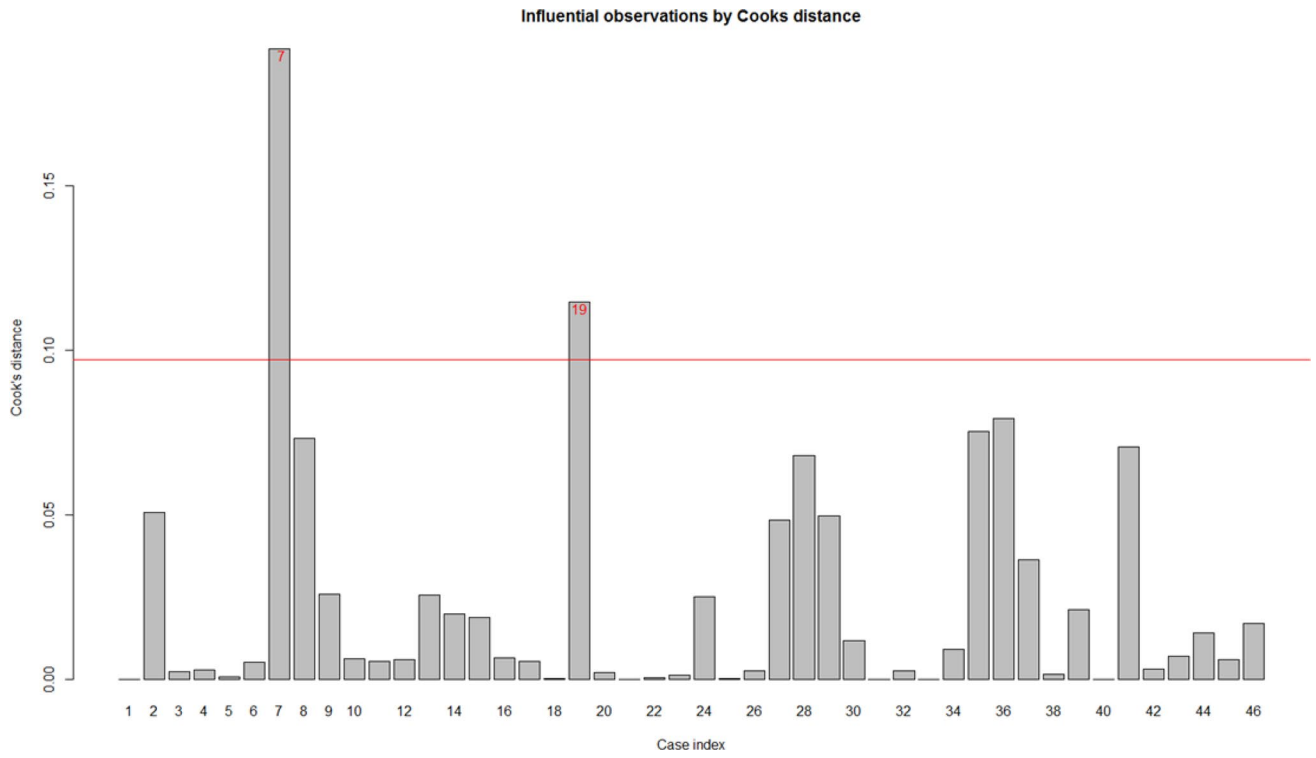

Fig. 1 Cook's distance (D_i) calculated between pairs of mean observer rating ($\bar{q}_{14}$) and model-predicted value. The horizontal line marks the cutoff at $D > 3 \times \bar{D}$

Table 7 Observer ratings, and performance ratings x_1, x_2, and x_3 of the identified outliers

#	T	I	S	$\bar{q}_2$	$\bar{q}_3$	$\bar{q}_5$	$\bar{q}_6$	$\bar{q}_7$	$\bar{q}_9$	$\bar{q}_{10}$	$\bar{q}_{11}$	$\bar{q}_{12}$	$\bar{q}_{13}$	$\bar{q}_{14}$	x_1	x_2	x_3
7	2	1	Cd	1.7	2.0	2.3	2.0	1.7	2.0	2.7	2.0	1.7	1.7	2.3	5.0	4.0	3.5
19	4	1	Rd	3.0	2.3	3.0	4.0	3.0	3.0	4.0	3.0	4.0	3.3	3.7	1.0	5.0	3.5

Case number (#), team id (T), performance iteration (I), scenario (S, see Table 3)

Table 8 Observer comments for case 19 (outlier)

Observer	Comment
1	"Subject was constantly glancing at matrix to try to make a decision diagnosis. Subject was not asking many questions, but appeared to be process the information heard to contribute to the diagnosis. Subject was slightly nervous as evidenced by hands shaking a little bit."
2	"Team exchanged ideas while the coordinator checked on victim. Made good use of time (asked double questions). Asked great questions ... very "in charge" during the diagnosis"
3	"They got the wrong diagnosis, it seemed expert #1 were much more active than #2. Focusing on the bruises seems to confuse the team. They developed a strategy to ask each other what information they had."

too easy (as hinted by the fact that the mean task score was 4.00) and that the teams were able to achieve good scores despite poor teamwork.

As summarized in Table 9, 11 cases stand out as either:

- having very low task scores ($x_1 < 2$),
- having poor task scores ($x_1 < 3$), above-average observer rating ($\bar{q}_{14} \geq 3$) and a difference of more than one, or
- vice versa with high task score ($x_1 > 4$) and low observer rating ($\bar{q}_{14} \leq 3$).

Under the conditions imposed by removing these outliers, there is high correlation between task scores and observer ratings of team performance, $r(33) = 0.83$, $p < .01$.

Table 9 Outliers by the alternative heuristic

						Outlier condition		
#	T	I	S	$\bar{q}_{14}$	x_1	$x_1 < 2$	$x_1 < 3$ and $\bar{q}_{14} \geq 3$ and $\lvert x_1 - \bar{q}_{14} \rvert > 1$	$x_1 > 4$ and $\bar{q}_{14} \leq 3$
2	1	2	Ct	2.7	4.2			✓
7	2	1	Cd	2.3	5.0			✓
9	2	3	Rd	2.0	1.0	✓		
19	4	1	Rd	3.7	1.0	✓	✓	
24	4	6	Nt	4.3	2.6		✓	
25	5	1	Rd	2.7	1.4	✓		
28	5	5	Cd	2.7	5.0			✓
29	6	1	Nd	3.3	1.0	✓	✓	
37	7	3	Rd	3.0	5.0			✓
41	8	1	Rd	3.7	1.0	✓	✓	
42	8	2	Rt	3.7	2.6		✓	

Case number (#), team id (T), performance iteration (I), scenario (S, see Table 3)

5.4 Multiple linear regression

To determine what aspects of the team's performance are relevant to the observers' tacit understanding of team performance, the observer score ($\bar{q}_{14}$) was fitted against the task score (x_1) and all possible combinations of the collected communication and interaction metrics (x_{2-6}). For this multiple linear regression analysis, the two outliers were removed and all resulting regression models were compared for highest adjusted R^2 identifying the best model of $\bar{q}_{14}$ as proportional to x_1 and x_3 at $F(2,41) = 13.96$, $p < .001$, with adjusted R^2 of 0.38. Although only 38% of the variance in the observers' ratings is explained by task-work, it is an increase from the baseline task score-only model, i.e., $\bar{q}_{14}$ as proportional to only x_1 having $F(1,42) = 22.47$, $p < .001$ with adjusted $R^2 = 0.33$. A better result was achieved by applying the alternative heuristic for removing outliers presented above, the best model fits $\bar{q}_{14}$ as proportional to x_1, x_2 and x_3 at $F(3,31) = 32.45$, $p < .001$, with adjusted R^2 of 0.74. The accuracy of this model is very high compared to the baseline; however, the caveat is that the model could not be used for almost 25% of the performances (11 of 46). This obviously reduces validity, but the finding is still interesting as it shows that under the right circumstances, team performance assessment can be automated with high accuracy and high reliability. The circumstances in this case being that the teams should not be total novices at the task, the tasks should be challenging enough that teams cannot master them without effort, and alternative assessments are needed when the outcome of the task-work is too poor. These restrictions are not very severe as it is quite reasonable to expect teams to receive other training/instruction before relying on automated performance assessment systems, nor is it typically very complicated to design systems that override the linear assessment model under certain conditions such as failure to complete the given task.

The regression coefficients (β) for the model depends on the scale used for the model inputs, as such they are highly dependent on context. For reference, Table 10 presents these coefficients for the second model together with their standard errors. The predicted values of the model for each performance (excluding the outliers) are shown, as gray circles in Fig. 2, while the black dots show the corresponding task scores (x_1). Note that since a discrete rating scale was used, there are multiple cases with the same mean observer rating. Pearson's correlation coefficient between the model-predicted values and the observer ratings is $r(33) = 0.87$, $p < .01$. Including any of the three other coefficients, x_{4-6}, results in over-fitting.

The regression model shows that the observers' ratings of virtual team performance in the presented experiment can be replicated with acceptable accuracy by combining the calculated task score, the spoken communication frequency and the time spent on verbal communication. There is no evidence to support that the provided metrics of written communication can further improve the model. As the model-predicted values correlate stronger than the task score metrics (0.87 vs 0.51), the regression model is a better predictor of observers' team performance ratings than task score, hence the experiment answers RQ2 by demonstrating that a simple automated solution based on easily accessible effectiveness and communication-based metrics are better at replicating observer ratings than pure task score-based systems.

Table 10 Regression coefficients for the best multiple linear regression model for team performance (as rated by observers) predicted by automated outcome-based and communication-based metrics

	β	SD (β)
(Intercept)	− 1.40	0.59
x_1	1.01	0.11
x_2	0.12	0.12
x_3	0.17	0.10

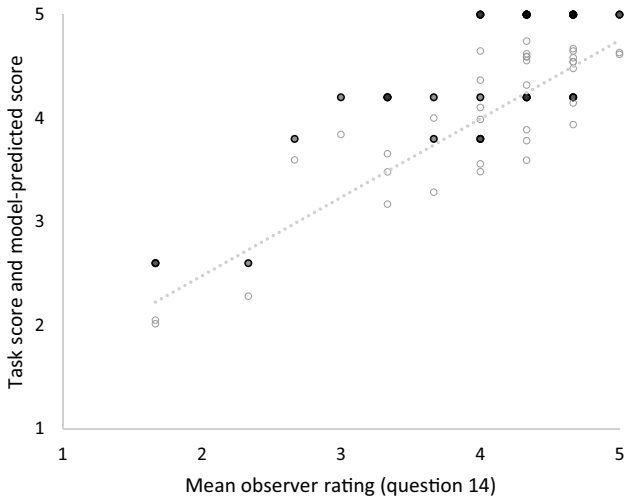

Fig. 2 Regression of team performance. Task performance scores are drawn as black dots and model-predicted team performance scores as gray circles. The dashed line marks the least-squares regression

5.5 Observer comments

The observers' written comments were collected after the final challenge was completed. Their comments are presented in Table 11. Observer #3 noted that observing teamwork by monitoring one individual is a difficult task and gave the advice that virtual teamwork is better observed remotely, where one may be able to observe all team members simultaneously. Two observers commented on the significance of communication, which confirms that observers' interpretation of teamwork in virtual teams is largely a combination of effectiveness and communication (Andersson et al. 2017). Observer #4's comment on the ability to work with high levels of stress is an interesting insight, especially so since many models of teamwork include stress as a factor (NATO RTO HFM-087 2005; NATO RTO HFM-156 2010; Hull et al. 2011).

6 Analysis and discussion

This section presents an analysis of the results in general, followed by sub-sections with discussions relating to each of the two research questions in depth. Additionally, the findings are discussed in relation to the studied literature.

A moderate correlation ($r = .51$) could be identified between the team-level observer rating of team performance and the objective task performance score. The correlation strength gives a hint that observers did account for effectiveness when rating performance. However, since the correlation strength is not very high, they must also have taken other parts of the performance into account, e.g., attitudes, behaviors and cognitions as have been suggested in other studies (e.g., Salas et al. 2017). This motivates the need to complement measures of effectiveness with indirect metrics of team performance in automated performance assessment systems. Regardless, one might have expected even higher correlation, since the task was very isolated and the teams were created for the sole purpose of solving the tasks. Indeed the correlation is high ($r = .83$) under the conditions defined by removing the outlier identified by the alternative heuristic.

The most striking difference found in this study between observers' team performance ratings and the objective task score, occur when teams score exceptionally low on the task score metric, e.g., when they fail to solve their challenges. In these cases, the observer ratings are significantly higher than their strictly task-based counterparts, suggesting that the observers were reluctant to penalize the teams for task failure and chose to focus on positive behaviors. Another potential explanation lies in the fact that many of these low-scoring performances were the first performance of a newly formed team—thus their performances may

Table 11 Observers' summarizing comments

Observer	Comment
1	"It was interesting to observe the teamwork evolve. I placed highest regard on how well and clear the communication was by each person in the team. The health professionals did not consult with each other as much as I expected. Only one group observed was effective as well as efficient in their group communication."
2	"Coordination and communication are the main things I expect to impact my assessments."
3	"Team factors are in general much more important than individual. However, rating the team is harder when focusing on one member. The best spot to assess the team is by observing remotely."
4	"Although stress may be high within a team, if they can remain calm, communicate clearly and work together to complete the task, then all is well."

have been hampered by them having to spend time on team-building, strategizing and task familiarization. Observers that recognized such behaviors as an important part of the teamwork may have opted to reward the teams with better scores than what was covered by the sheer outcome-based scoring heuristics. In addition, the effect could be emphatic from the observers' side, as it is reasonable to expect newly formed teams to struggle with task performance in the forming phase. Perhaps the dissonance between the observers and team members could be resolved by letting them talk through the events in a post-hoc analysis session, e.g., an AAR (Rankin et al. 1995; Morrison and Meliza 1999; U.S. Army Combined Arms Center 2011; Vanderhaegen and Carsten 2017). Further research is needed to resolve this issue.

The observers' overall performance ratings correlate very strongly, $r(44) > 0.93$, with their own ratings of communication, even more so than with the task score. Thus, the observers seem to have focused more on whether their respective teams communicated well than they did on how well they got the job done. This result might be an effect of each observer being able to monitor only one team member. Observer #3's advice to monitor the entire team remotely might have generated other results. However, the inter-rater agreement is highest on these rating items, which confirms that the observers' ratings are indeed reliable and that the observers' interpretations of team performance in virtual teams are heavily influenced by communication, just as in Brannick et al.'s (1993) flight simulation teamwork experiment.

This result highlights that there are many different ways of interpreting and assessing team performance and that obtained results are meaningful only in the context of that interpretation (Kendall and Salas 2004; Rosen et al. 2008; Wildman et al. 2013).

6.1 How reliable are human observers' ratings of teamwork in virtual teams, when only able to monitor team members at one site?

The observers reported that it is hard to observe teamwork when only able to monitor one of the team members. From that standpoint, one might expect the observers' inter-rater reliability to be low. Indeed, this study has shown that the observers' inter-rater is poor on ratings of team workload. The ratings of team stress are slightly better, but still not good. Therefore, measurement of team workload and stress in virtual teams is advisably done through more reliable mechanisms, e.g., using validated instruments for self-assessment such as NASA-TLX (Hart and Staveland 1988) and the Teamwork Workload Scale (Nonose et al. 2016). One observer noted that remote observation might be more appropriate, as the observers then would be able to monitor the whole team. Whether such an approach would actually result in higher reliability has not been investigated.

The agreement on ratings of collaborative problem-solving items (q_7 and q_8) is too low to be classified as good. Again a reasonable explanation might be that the inability to observe the entire team made it difficult for the observers to get a good understanding of how well the team members collaborated.

All other items were rated with good or excellent inter-rater agreement, including the overall team performance rating. The results thus show that observers only able to monitor one member of a virtual team are able to reliably rate overall team performance and performance indicators associated with coordination, backing-up behavior, and information exchange.

6.2 Can human observers' overall ratings of virtual teams' performances be replicated with enough accuracy using only automatically quantifiable metrics?

The traditional approach to automated team performance assessment is based on different metrics for task outcome similar to the task score metric x_1 designed for this study (Ceschi et al. 2014). Some researchers have proposed to dress this with sophisticated automated systems for speech recognition and content analysis (Kiekel et al. 2001; Hetrick et al. 2002; Foltz et al. 2003; Cooke et al. 2004; LaVoie et al. 2008; Wang et al. 2018), and video analysis (Stein et al. 2013). This study took a different approach to use simple technology and work easily captured performance indicators to create a mathematical model that was fitted against the observers' overall team performance assessment. Using multiple linear regression analysis on a subset of the data, the best model identified correlated highly with the observer rating, $r(33) = 0.87$. The resulting model is a linear combination of task performance score, number of verbally communicated messages, and time spent on verbal communication, i.e., one measure of task outcome and two of verbal communication frequency. The main value of this model is that it shows that team performance ratings (as judged by observers) can be quantitatively approximated using only metrics that are easily captured using readily available technology. Overall, this suggests that once calibrated and under the right conditions, automated team performance assessment systems can indeed rate team performance of virtual teams, if the objective is to obtain end results that correspond to how expert observers quantify overall team performance. While it is not certain that advanced speech recognition and semantic analysis capabilities such as the novel virtual personal assistant technology utilizes would improve the model (Këpuska and Bohouta 2018), it is certainly an interesting topic worthy of further investigation. It should be noted that

these results suffer from not having been validated. Cross validation with a larger data set is required to establish whether automation can actually be accomplished using a model such as the one calculated in this study. However, the results are convincing enough to motivate further research.

It should also be noted that although the acquired fit is good (albeit not validated), it has been recognized that other regression equations, e.g., logistic and polynomial, might produce better models and perhaps also be applicable to the data set also covering some of the outlier conditions that were excluded for this analysis. This has not been examined in this study as the aim of the study was to examine if the idea was even feasible while keeping the proposed solution simple, and this simple modeling technique matches this equilibrium nicely.

6.3 Implications

This study confirms that observers include task outcome in their ratings of virtual team performance, but that their ratings are also influenced by teamwork behaviors such as coordination and information exchange. Observers can reliably rate these metrics despite being able to monitor only one team member, which is beneficial, since they can then work in-situ, which increases their ability of picking up certain behaviors that are difficult to capture off-site (Wildman et al. 2013).

The results from this study also show that the idea that "there is no escaping observation" (Baker and Salas 1992; Rosen et al. 2008) is only partly valid. Current technology seems to be far away from the diagnostic capabilities of a human observer, consequently an automated performance assessment system cannot be trusted to generate very detailed feedback on teamwork. However, in some cases, the only result that matters is to get an estimate of the team's overall performance. The study shows that in a virtual team setting such metrics can be defined using a combination of task scores and simple measures of communication frequency. The drawback is that the resulting model will only ever be valid for the specific type of tasks that it was designed for, but on the other hand there are several use cases of performance assessment, where the same scenario is used repeatedly, e.g., training, education, and certification.

It is imperative to acknowledge that the model presented in this article is not very relevant in itself as the task it was constructed for is artificial. The value of this work lies in the process reaching the model, the fact that it can be done, and quite easily at that. There is nothing that says similar models cannot be created for other, more relevant, team performance scenarios. With a bank of reliable observer ratings and samples to train the algorithms, this study shows that a model can be calculated that mimics observer ratings of team performance. The next step is to validate such a model, e.g., with cross validation, to show that the model works on arbitrary inputs. It is recognized that the model is tied to the team performance scenario, and that other scenarios require other inputs and other modeling techniques. However, the main take-away message from this study is that these models are easy to generate.

7 Conclusion

The study has confirmed that expert observers' ratings of virtual team performance are reliable for indicators associated with coordination, backing-up behavior, and information exchange—but not for team workload. Task outcome is indeed a huge portion of team effectiveness and team performance, and this study confirms that expert observers take a more holistic approach to team performance assessment. However, while the aggregated observer ratings of team performance correlate with the task score, there are also notable and systematic differences such as when task scores are low and when teams are newly formed and untrained at the tasks they are performing. The observer ratings are heavily influenced by intra-team communication, suggesting that observers regard team cognition as an important part of team performance.

Other studies have pointed at observers as the main source for team performance appraisal, and also found that team workload and stress are important aspects of performance. This study found that among synchronous virtual teams, such ratings are not very reliable when the observers are not able to monitor the team in its entirety. This finding is important as it highlights the need for alternative methods for assessing team workload and stress in this setting, and their relevance to performance.

With simple frequency metrics added to the task score, a regression model was able to predict the observers' performance ratings of the teams with high accuracy under certain conditions. The implication of this result is that when algorithms have been trained and the performance is centered around communication and collaborative problem-solving, and it is possible to mimic observer ratings of synchronous virtual team performances by combining outcome-based task scores with quantified metrics of communication. It is reasonable to expect that similar results can be obtained for other task types, by identifying the most relevant quantifiable metrics. Thus, the study shows that automated team performance assessment systems can be trained to numerically analyze performance of synchronous virtual teams in much the same way as expert observers do. The results of this study motivate further research into the development of a new type of training systems for teamwork in virtual environments, with support for automated on-the-fly

performance evaluation and feedback, thereby reducing the resource cost for training virtual teams.

Acknowledgments The author would like to thank fellow researchers Amy Rankin and Darryl Diptee for their help in setting up the experiment, as well as nursing teachers Barbara Dunham and Debra Kaczmar at Hartnell College in Salinas, CA, for their massive contribution by providing access to their classes of students, assisting with the setup, and reviewing the scenarios. Also the fellow software engineering experimentation students Kaveh Razavi and Jason Massey at George Mason University, as well as Prof. Jeff Offutt at the same institution, contributed by reviewing and commenting on the research design and study setup. Last, but not least, the authors would like to extend a thank you to the helpful students at Hartnell College and Naval Postgraduate School in Monterey, CA, whose contributions enabled this study. An IRB approval (IRB# NPS.2012.0042-IR-EP7-A) was granted by the Naval Postgraduate School Institutional Review Board for the study presented herein.

Disclaimer This work relates to Department of the Navy Grant N62909-11-1-7019 issued by Office of Naval Research Global. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein.

References

Abbott R, McClain J, Anderson B et al (2015) Automated performance assessment in cyber training exercises. In: Proceedings of the 2015 Interservice/Industry Training Systems and Education Conference (I/ITSEC). Orlando, FL

Alsharo M, Gregg D, Ramirez R (2017) Virtual team effectiveness: The role of knowledge sharing and trust. Inf Manag 54:479–490. https://doi.org/10.1016/j.im.2016.10.005

Andersson D (2011) Privacy and distributed tactical operations evaluation. In: Proceedings of the 4th international conference on advances in human-oriented and personalized mechanisms, technologies, and services. Barcelona, Spain

Andersson D, Rankin A, Diptee D (2017) Approaches to team performance assessment: a comparison of self-assessment reports and behavioral observer scales. Cognit Technol Work 19:517–528. https://doi.org/10.1007/s10111-017-0428-0

Andres HP (2013) Team cognition using collaborative technology: a behavioral analysis. J Manag Psychol J Manag Psychol Iss J Manag Psychol 28:38–54. https://doi.org/10.1108/02683941311298850

Annett J, Cunningham D, Mathias-Jones P (2000) A method for measuring team skills. Ergonomics 43:1076–1094

Atanasova Y, Senn C (2011) Global customer team design: dimensions, determinants, and performance outcomes. Ind Mark Manag 40:278–289. https://doi.org/10.1016/j.indmarman.2010.08.011

Baker G (2002) The effects of synchronous collaborative technologies on decision making: a study of virtual teams. Inf Resour Manag J 15:79–93. https://doi.org/10.4018/irmj.2002100106

Baker DP, Salas E (1992) Principles for measuring teamwork skills. Hum Factors 34:469–475

Berggren P (2016) Assessing shared strategic understanding. Linköping University, Linköping

Berggren P, Johansson BJE (2010) Developing an instrument for measuring shared understanding. In: French S, Tomaszewski B, Zobel C (eds) Proceedings of the 7th international ISCRAM conference. Seattle, WA

Berry GR (2011) Enhancing effectiveness on virtual teams: understanding why traditional team skills are insufficient. J Bus Commun 48:186–206. https://doi.org/10.1177/0021943610397270

Bordetsky A, Dougan A, Chiann FY, Kilberg A (2007) TNT maritime interdiction operation experiments: enabling radiation awareness and geographically distributed collaboration for network-centric maritime interdiction operations. In: Proceedings of the 12th international command and control research and technology symposium. Newport, RI

Bowers CA, Braun CC, Morgan BBJ (1997) Team workload: its meaning and measurement. In: Brannick MT, Salas E, Prince C (eds) Team Performance assessment and measurement: theory, methods, and applications. Lawrence Erlbaum, Mahwah, pp 85–108

Brannick MT, Roach RM, Salas E (1993) Understanding team performance: a multimethod study. Hum Perform 6:287–308

Brehmer B, Dörner D (1993) Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study. Comput Hum Behav 9:171–184

Breugst N, Patzelt H, Shepherd DA, Aguinis H (2012) Improves team performance assessment accuracy: evidence from a multilevel study. Acad Manag Learn Educ 11:187–206

Brueckner S, Guaspari D, Adelstein F, Weeks J (2008) Automated computer forensics training in a virtualized environment. Digit Investig 5:105–111. https://doi.org/10.1016/j.diin.2008.05.009

Bushe GR, Coetzer G (1995) Appreciative inquiry as a team development intervention: a controlled experiment. J Appl Behav Sci 31:13–30

Calvo-Merino B, Ehrenberg S, Leung D, Haggard P (2010) Experts see it all: Configural effects in action observation. Psychol Res 74:400–406. https://doi.org/10.1007/s00426-009-0262-y

Cannon-Bowers JA, Salas E (1990) Cognitive psychology and team training: shared mental models in complex systems. In: The fifth annual conference of the society for industrial organization psychology. Miami, FL

Cannon-Bowers JA, Tannenbaum SI, Salas E, Volpe CE (1995) Defining competencies and establishing team training requirements. In: Guzzo RA, Salas E (eds) Team EFFECTIVENESS AND DECISION MAKING IN ORGANIZATIONS. Jossey-Bass, San Francisco, pp 333–380

Ceschi A, Dorofeeva K, Sartori R (2014) Studying teamwork and team climate by using a business simulation: How communication and innovation can improve group learning and decision-making performance. Eur J Train Dev 38:211–230

Chang HH, Hung C-J, Hsieh H-W (2014) Virtual teams: cultural adaptation, communication quality, and interpersonal trust. Total Qual Manag Bus Excell 25:1318–1335. https://doi.org/10.1080/14783363.2012.704274

Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6:284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clauser BE, Harik P, Clyman SG (2000) The generalizability of scores for a performance assessment scored with a computer-automated scoring system. J Educ Meas 37:245–261. https://doi.org/10.1111/j.1745-3984.2000.tb01085.x

Cook RD (1977) Detection of Influential Observation in Linear Regression. Technometrics 19:15–18. https://doi.org/10.2307/1268249

Cooke NJ, Salas E, Kiekel PA, Bell B (2004) Advances in measuring team cognition. In: Salas E, Fiore SM (eds) Team cognition:

understanding the factors that drive process and performance. American Psychological Association, Washington, DC, pp 83–106
Cooke NJ, Gorman JC, Myers CW, Duran JL (2013) Interactive team cognition. Cogn Sci 37:255–285. https://doi.org/10.1111/cogs.12009
Curtis AM, Dennis AR, McNamara KO (2017) From monologue to dialogue: performative objects to promote collective mindfulness in computer-mediated team discussions. MIS Q 41:559–581
De Penning L, Kappé B, Boot E (2009) Automated performance assessment and adaptive training for training simulators with SimSCORM. In: Proceedings of the 2009 interservice/industry training, simulation, and education conference (I/ITSEC). Orlando, FL
Deaton JE, Bell B, Fowlkes JE et al (2007) Enhancing team training and performance with automated performance assessment tools. Int J Aviat Psychol 17:317–331. https://doi.org/10.1080/10508410701527662
DeChurch LA, Mesmer-Magnus JR (2010) The cognitive underpinnings of effective teamwork: a meta-analysis. J Appl Psychol 95:32–53. https://doi.org/10.1037/a0017328
Dickinson TL, McIntyre RM (1997) A conceptual framework for teamwork measurement. In: Brannick MT, Salas E, Prince C (eds) Team Performance assessment and measurement: theory, methods, and applications. Lawrence Erlbaum, Mahwah, pp 19–43
Dorsey D, Russell S, Keil C et al (2009) Measuring teams in action: Automated performance measurement and feedback in simulation-based training. In: Team effectiveness in complex organizations: cross-disciplinary perspectives and approaches. Routledge, New York, pp 351–381
Dubé G, Kramer C, Vachon F, Tremblay S (2011) Measuring the Impact of a collaborative planning support system on crisis management. In: Proceedings of the 8 h international ISCRAM CONFERENCE. Lisbon, Portugal
Espevik R, Johnsen BH, Eid J (2011) Communication and performance in co-located and distributed teams: an issue of shared mental models of team members? Mil Psychol 23:616–638. https://doi.org/10.1080/08995605.2011.616792
Ferrazzi K (2014) Managing yourself. Harv Bus Rev 92:120–123. https://doi.org/10.1016/B978-0-7020-2920-2.50008-X
Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Stanford
Fischer F, Mandl H (2005) knowledge convergence in computer-supported collaborative learning: the role of external representation tools. J Learn Sci 14:405–441. https://doi.org/10.1207/s15327809jls1403
Foltz PW, Laham D, Derr M (2003) Automated speech recognition for modeling team performance. Proc Hum Factors Ergon Soc Annu Meet 47:673–677. https://doi.org/10.1177/154193120304700402
Ford RC, Piccolo RF, Ford LR (2017) Strategies for building effective virtual teams: trust is key. Bus Horiz 60:25–34. https://doi.org/10.1016/j.bushor.2016.08.009
Frank G, Evens N, Hubal R, Whiteford B (2008) Automated, interactive AARs: a positive spin. In: Proceedings of the 2008 interservice/industry training, simulation, and education conference (I/ITSEC). Orlando, FL
Furumo K, Pearson JM (2006) An empirical investigation of how trust, cohesion, and performance vary in virtual and face-to-face teams. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences. Kauai, HI, pp 1–10
Geers K (2010) Live fire exercise: preparing for cyber war. J Homel Secur Emerg Manag. https://doi.org/10.2202/1547-7355.1780
Gierl MJ, Latifi S, Lai H et al (2014) Automated essay scoring and the future of educational assessment in medical education. Med Educ 48:950–962. https://doi.org/10.1111/medu.12517
Granåsen M, Andersson D (2016) Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. Cognit Technol Work 18:121–143. https://doi.org/10.1007/s10111-015-0350-2
Harik P, Baldwin P, Clauser B (2013) Comparison of automated scoring methods for a computerized performance assessment of clinical judgment. Appl Psychol Meas 37:587–597. https://doi.org/10.1177/0146621613493829
Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical results. In: Hancock PA, Meshkati N (eds) Human Mental Workload. North Holland Press, Amsterdam, pp 239–250
Healey AN, Undre S, Vincent CA (2004) Developing observational measures of performance in surgical teams. Qual Saf Health Care 13:33–41. https://doi.org/10.1136/qshc.2004.009936
Hetrick C, Cooper G, Walsh K et al (2002) Evaluating communication in a synthetic task environment. In: Proceedings of the 44th annual conference of the international military testing association (IMTA). Ottawa, Canada
Hull L, Arora S, Kassab E et al (2011) Assessment of stress and teamwork in the operating room: AN exploratory study. Am J Surg 201:24–30. https://doi.org/10.1016/j.amjsurg.2010.07.039
Hutchins E (1995) Cognition in the wild. MIT Press, Cambridge
Ifenthaler D (2014) Toward automated computer-based visualization and assessment of team-based performance. J Educ Psychol 106:651–665. https://doi.org/10.1037/a0035505
Jarvenpaa SL, Shaw TR, Staples DS (2004) Toward contextualized theories of trust: the role of trust in global virtual teams. Inf Syst Res 15:250–267. https://doi.org/10.1287/isre.1040.0028
Johansson B, Persson M, Granlund R, Mattsson P (2003) C3Fire in command and control research. Cognit Technol Work 5:191–196. https://doi.org/10.1007/s10111-003-0127-x
Johnson TE, Top E, Yukselturk E (2011) Team shared mental model as a contributing factor to team performance and students' course satisfaction in blended courses. Comput Hum Behav 27:2330–2338
Jouanne E, Charron C, Chauvin C, Morel G (2017) Correlates of team effectiveness: an exploratory study of fire fighter's operations during emergency situations. Appl Ergon 61:69–77. https://doi.org/10.1016/j.apergo.2017.01.005
Kaur N, Sood SK (2015) A game theoretic approach for an iot-based automated employee performance evaluation. IEEE Syst J 11:1385–1394. https://doi.org/10.1109/JSYST.2015.2469102
Kendall DL, Salas E (2004) Measuring team performance: review of current methods and consideration of future needs. In: Ness JW, Tepe V, Ritzer D (eds) Advances in human performance and cognitive engineering research, vol 5. JAI Press, Amsterdam, pp 307–326
Këpuska V, Bohouta G (2018) Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). 2018 IEEE 8th Annu Comput Commun Work Conf CCWC 2018 2018–Janua. https://doi.org/10.1109/CCWC.2018.8301638
Kiekel PA, Cooke NJ, Foltz PW, Shope SM (2001) Automating measurement of team cognition through analysis of communication data. In: Smith MJ, Salvendy G, Harris D, Koubek RJ (eds) Usability evaluation and interface design. Lawrence Erlbaum Associates, Mahwah, pp 1382–1386
Kizza JM, Ssanyu J (2005) Workplace Surveillance. In: Weckert J (ed) Electronic monitoring in the workplace: controversies and solutions. Idea Group Inc. Publishers, Hershey
Kolb JA (1995) Leader behaviors affecting team performance: similarities and differences between leader/member assessments. J Bus Commun 32:233–248

Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kraiger K, Wenzel LH (1997) Conceptual development and empirical evaluation of measures of shared mental models as indicators of team effectiveness. In: Brannick MT, Salas E, Prince C (eds) Team performance assessment and measurement: theory, methods, and applications. Lawrence Erlbaum, Mahwah, pp 63–84

Langan-Fox J, Wirth A, Code S et al (2001) Analyzing shared and team mental models. Int J Ind Ergon 28:99–112. https://doi.org/10.1016/S0169-8141(01)00016-6

Lavoie N, Foltz P, Rosenstein M et al (2008) Automating convoy training assessment to improve soldier performance. In: Proceedings of the 26th Army Science Conference. Orlando, FL

LaVoie N, Foltz PW, Rosenstein M et al (2008) Automated support for AARs: exploiting communication to assess team performance. In: Proceedings of the 2008 interservice/industry training, simulation, and education conference (I/ITSEC). Orlando, FL

Lawson BD, Britt TW, Kelley AM et al (2017) Computerized tests of team performance and crew coordination suitable for military/aviation settings. Aerosp Med Hum Perform 88:722–729

Letsky M, Warner N (2008) Macrocognition in teams. In: Letsky M, Warner NW, Fiore SM, Smith CAP (eds) Macrocognition in teams: theories and methodologies. Ashgate, Aldershot, pp 1–13

Lipnack J, Stamps J (2000) Virtual teams: people working across boundaries with technology, 2nd edn. John Wiley, New York

Maccoun RJ, Kier E, Belkin A (2005) Does social cohesion determine motivation in combat? An old question with an old answer. Armed Forces Soc 32:1–9

Macht GA, Nembhard DA, Kim JH, Rothrock L (2014) Structural models of extraversion, communication, and team performance. Int J Ind Ergon 44:82–91. https://doi.org/10.1016/j.ergon.2013.10.007

MacKenzie C, Hu PFM, Fausboll C et al (2007) Challenges to remote emergency decision-making for disasters or Homeland Security. Cogn Technol Work 9:15–24. https://doi.org/10.1007/s10111-006-0051-y

Marlow S, Bisbey T, Lacerenza C, Salas E (2018) Performance measures for health care teams: a review

Martin MJ, Foltz PW (2004) Automated team discourse annotation and performance prediction using LSA. In: Proceedings of HLT-NAACL 2004: short papers. association for computational linguistics, pp 97–100

Martínez-Miranda J, Pavón J (2012) Modeling the influence of trust on work team performance. Simul Trans Soc Model Simul Int 88:408–436. https://doi.org/10.1177/0037549711404714

Maynard MT, Gilson LL (2014) The role of shared mental model development in understanding virtual team effectiveness. Gr Organ Manag 39:3–32. https://doi.org/10.1177/1059601113475361

McComb SA (2008) Shared mental models and their convergence. In: Letsky M, Warner NW, Fiore SM, Smith CAP (eds) Macrocognition in teams: theories and methodologies. Ashgate, Aldershot, pp 35–50

McEwan D, Ruissen GR, Eys MA et al (2017) The effectiveness of teamwork training on teamwork behaviors and team performance: a systematic review and meta-analysis of controlled interventions. PLoS One 12:1–24. https://doi.org/10.1371/journal.pone.0169604

McGuinness B, Foy L (2000) A subjective measure of SA: the Crew Awareness Rating Scale (CARS). In: Human performance, situational awareness and automation conference. Savannah, GA

Mjelde FV, Smith K (2013) Performance assessment of military team-training for resilience in complex maritime environments. In: Proceedings of the 57th annual meeting of the human factors and ergonomics society, pp 2116–2120

Mohammed S, Dumville BC (2001) Team mental models in a team knowledge framework: expanding theory and measurement across disciplinary boundaries. J Organ Behav 22:89–106

Molleman E, Slomp J (1999) Functional flexibility and team performance. Int J Prod Res 37:1837–1858. https://doi.org/10.1080/002075499191021

Möller HJ (2000) Rating depressed patients: observer- vs self-assessment. Eur Psychiatry 15:160–172. https://doi.org/10.1016/S0924-9338(00)00229-7

Morgan L, Paucar-Caceres A, Wright G (2014) Leading effective global virtual teams: the consequences of methods of communication. Syst Pract Action Res 27:607–624. https://doi.org/10.1007/s11213-014-9315-2

Morrison JE, Meliza LL (1999) Foundations of the after action review process. U.S. Army Resarch Institute, Alexandria

Murphy RR, Riddle D, Rasmussen E (2004) Robot-assisted medical reachback: a survey of how medical personnel expect to interact with rescue robots. Robot Hum Interact Commun. https://doi.org/10.1109/ROMAN.2004.1374777

NATO RTO HFM-087 (2005) Military command team effectiveness: model and instrument for assessment and improvement. NATO Research and Technology Organisation, Neuilly-sur-Seine Cedex, Frankrike

NATO RTO HFM-156 (2010) Measuring and analyzing command and control performance effectiveness

Nonose K, Yoda Y, Kanno T, Furuta K (2016) An exploratory study: a measure of workload associated with teamwork. Cognit Technol Work 18:351–360. https://doi.org/10.1007/s10111-015-0363-x

O'Mahony S, Ferraro F (2007) The emergence of governance in an open source community. Acad Manag J 50:1079–1106

Ong J (2007) Automated performance assessment and feedback for free-play simulation-based training. Perform Improv 46:24–31. https://doi.org/10.1002/pfi

Pangil F, Chan JM (2014) The mediating effect of knowledge sharing on the relationship between trust and virtual team effectiveness. The mediating effect of knowledge sharing on the relationship between trust and virtual team effectiveness. J Knowl Manag 18:92–106

Papps KL, Bryson A, Gomez R (2011) Heterogeneous worker ability and team-based production: EVIDENCE from major league baseball, 1920–2009. Labour Econ 18:310–319. https://doi.org/10.1016/j.labeco.2010.11.005

Paul R, Drake JR, Liang H (2016) Global virtual team performance: the effect of coordination effectiveness, trust, and team cohesion. IEEE Trans Prof Commun 59:186–202. https://doi.org/10.1109/TPC.2016.2583319

Persson M, Rigas G (2014) Complexity: the dark side of network-centric warfare. Cognit Technol Work 16:103–115. https://doi.org/10.1007/s10111-012-0248-1

Rankin WJ, Gentner FC, Crissey MJ (1995) After action review and debriefing methods: technique and technology. In: Proceedings of the 17th interservice/industry training systems and education conference (I/ITSEC). Alburquerque, NM, pp 252–261

Rasmussen TH, Jeppesen HJ (2006) Teamwork and associated psychological factors: a review. Work Stress 20:105–128. https://doi.org/10.1080/02678370600920262

Riegelsberger J, Sasse MA, McCarthy J (2003) The researcher's dilemma: evaluating trust in computer-mediated communication. Int J Hum Comput Stud 58:759–781

Robert G, Hockey J (1997) Compensatory control in the regulation of human performance under stress and high workload: a cognitive-energetical framework. Biol Psychol 45:73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Rosen MA, Salas E, Wilson KA et al (2008) Measuring team performance in simulation-based training: adopting best practices

for healthcare. Simul Healthc 3:33–41. https://doi.org/10.1097/SIH.0b013e3181626276
Rousseau V, Aubé C, Savoie A (2006) Teamwork behaviors: a review and an integration of frameworks. Small Gr Res 37:540. https://doi.org/10.1177/1046496406293125
Rutherford JS (2017) Monitoring teamwork: a narrative review. Anaesthesia 72:84–94. https://doi.org/10.1111/anae.13744
Sadagic A, Kölsch M, Welch G et al (2013) Smart instrumented training ranges: bringing automated system solutions to support critical domain needs. J Def Model Simul Appl Methodol Technol 10:327–342. https://doi.org/10.1177/1548512912472942
Salas E, Sims DE, Burke CS (2005) Is there a "Big Five" in teamwork? Small Gr Res 36:555–599. https://doi.org/10.1177/1046496405277134
Salas E, Cooke NJ, Rosen MA (2008) On teams, teamwork, and team performance: discoveries and developments. Hum Factors 50:540–547. https://doi.org/10.1518/001872008X288457
Salas E, Reyes DL, Woods AL (2017) The assessment of team performance: observations and needs. In: von Davier A, Zhu M, Kyllonen P (eds) Innovative assessment of collaboration. Springer, Cham, pp 21–36
Sapateiro CM, Antunes P, Johnstone D, Pino JA (2017) Gathering big data for teamwork evaluation with microworlds. Cluster Comput 20:1637–1659. https://doi.org/10.1007/s10586-016-0715-1
Schmidtke JM, Cummings A (2017) The effects of virtualness on teamwork behavioral components: the role of shared mental models. Hum Resour Manag Rev 27:660–677. https://doi.org/10.1016/j.hrmr.2016.12.011
Scoppa V (2015) Fatigue and team performance in Soccer: evidence from the FIFA world cup and the UEFA European championship. J Sport Econ 16:482–507. https://doi.org/10.1177/1527002513502794
Shanahan P (2001) Mapping team performance shaping factors. Fort Halstead, UK
Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 86:420–428. https://doi.org/10.1037/0033-2909.86.2.420
Soós JK, Juhász M (2011) Capturing team performance differences through communication based analyses of team cognition. Period Polytech Soc Manag Sci 18:75–85
Stainback JR (2011) A new lean model: improving team performance through communications efficacy. University of Tennessee, Tennessee
Stein D, Krausz B, Löffler J et al (2013) Automatic audio and video event recognition in an intelligent resource management system. Int J Inf Syst Cris Response Manag 5:1–12. https://doi.org/10.4018/ijiscram.2013100101
Subramaniam C, Ali H, Mohd Shamsudin F (2010) Understanding the antecedents of emergency response: a proposed framework. Disaster Prev Manag An Int J 19:571–581. https://doi.org/10.1108/09653561011091904
Sudhakar GP, Farooq A, Patnaik S (2011) Soft factors affecting the performance of software development teams. Team Perform Manag 17:187–205. https://doi.org/10.1108/13527591111143718
Tabassi AA, Ramli M, Roufechaei KM, Tabasi AA (2014) Team development and performance in construction design teams: an assessment of a hierarchical model with mediating effect of compensation. Constr Manag Econ 32:932–949. https://doi.org/10.1080/01446193.2014.935739
Temkin-Greener H, Gross D, Kunitz SJ, Mukamel D (2004) Measuring Interdisciplinary Team Performance in a Long-Term Care Setting. Source Med Care 42:472–481
Tesluk P, Mathieu JE, Zaccaro SJ, Marks M (1997) Task and aggregation issues in the analysis and assessment of team performance. In: Brannick MT, Salas E, Prince C (eds) Team performance assessment and measurement: theory, methods, and applications. Lawrence Erlbaum, Mahwah, pp 197–224
Turel O, Zhang Y (2010) Does virtual team composition matter? Trait and problem-solving configuration effects on team performance. Behav Inf Technol 29:363–375. https://doi.org/10.1080/01449291003752922
U.S. Army Combined Arms Center (2011) Leader's guide to after-action reviews (AAR). Fort Leavenworth, Kansas
Vanderhaegen F, Carsten O (2017) Can dissonance engineering improve risk analysis of human–machine systems? Cognit Technol Work 19:1–12. https://doi.org/10.1007/s10111-017-0405-7
Wang Z, Zechner K, Sun Y (2018) Monitoring the performance of human and automated scores for spoken responses. Lang Test 35:101–120. https://doi.org/10.1177/0265532216679451
Weaver JL, Bowers CA, Salas E (2001) Stress and teams: performance effects and interventions. In: Hancock PA, Desmond PA (eds) Stress, workload, and fatigue. Lawrence Erlbaum Associates, Mahwah, pp 83–106
Wildman JL, Salas E, Scott CPR (2013) Measuring cognition in teams: a cross-domain review. Hum Factors 56:911–941. https://doi.org/10.1177/0018720813515907
Yeatts DE, Hyten C (1998) High-performing self-managed work teams. SAGE, Thousand Oaks
Yen IL, Bastani F, Huang Y et al (2017) SaaS for automated job performance appraisals using service technologies and big data analytics. In: Proceedings of the IEEE 24th International Conference on Web Services (ICWS), pp 412–419
Yoon SW, Song JH, Lim DH, Joo B-K (2010) Structural determinants of team performance: the mutual influences of learning culture, creativity, and knowledge. Hum Resour Dev Int 13:249–264. https://doi.org/10.1080/13678868.2010.483815
Ziek P, Smulowitz S (2014) The impact of emergent virtual leadership competencies on team effectiveness. Leadersh Organ Dev J 35:106–120