Cogn Tech Work (2017) 19:517–528
DOI 10.1007/s10111-017-0428-0

ORIGINAL ARTICLE

Approaches to team performance assessment: a comparison of self-assessment reports and behavioral observer scales

Dennis Andersson[1,2] · Amy Rankin[2] · Darryl Diptee[3]

Received: 25 January 2017 / Accepted: 20 July 2017 / Published online: 5 August 2017

Abstract Human factors research popularly employs perception-based techniques to investigate team performance and its dependency to cognitive processes. Such studies frequently rely upon either observer-based or self-assessment techniques to collect data. In this study, we examined behavioral observer ratings and self-assessment ratings for measuring team performance in virtual teams, with team performance regarded as a combination of task outcome and team cognition. Juxtaposing self-assessments and observer ratings from a quasi-experiment comparing team performance rating techniques reveals that they indeed produce overall similar results, with both singling out teamwork effectiveness ratings as the strongest contributor to overall team performance. However, the comparisons show remarkably low correlation on individual questionnaire items. The most striking difference is that the team members' self-assessments of workload are lower than the corresponding observer ratings. In particular, the self-assessments do not correlate at all with overall team performance, whereas the observers' workload ratings are more consistent with contemporary research that suggests a strong correlation between workload and team performance, suggesting that observer-based techniques are more reliable than self-assessments for assessing workload. For other ratings, the results show that the two techniques are fairly equal, suggesting that the choice between methods to employ can be deferred to other considerations such as obtrusiveness, accessibility, and resource availability.

Keywords Behavioral observation scales · Performance assessment · Self-assessment reports · Team performance

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.

✉ Dennis Andersson
dennis.andersson@foi.se

1 Division for C4ISR, Swedish Defense Research Agency, Box 1165, 581 11 Linköping, Sweden

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

3 Department of Information Science, Naval Postgraduate School, 1 University Circle, Monterey, CA 93943, USA

1 Introduction

Research on teamwork has been plentiful over the last couple of decades, with a lot of attention on understanding what team performance is and how it can be measured (Valentine et al. 2015). In general, five categories of performance measurement methods exist: (1) event-based measurement (EBM), (2) automated performance monitoring (APM), (3) behaviorally anchored rating scales (BARS), (4) behavioral observation scales (BOS), and (5) self-assessment reports (SAR) (Kendall and Salas 2004). In team performance assessment literature, team performance is often seen as a function of one or more of: (1) individual processes, (2) individual outcomes, (3) team processes, and (4) team outcomes (Smith-Jentsch et al. 1998). Traditionally, most research on teamwork has been in the physical realm, on, e.g., health care teams, emergency management, work teams, or military teams or teams in sports. However, in the last decade the virtual domain has received much more attention, with a focus on trying to understand whether there are any differences between how virtual teams function compared to traditional work teams, and what implications it may have on management and leadership (e.g., Krumm et al. 2013). Observing and assessing team cognition in virtual teams is complicated due to the intrinsic nature of working in the digital domain. To overcome this challenge, researchers have attempted to measure team performance in virtual settings using a triangulation of the aforementioned methods (e.g., Granåsen and Andersson 2016; Maynard and Gilson 2014).

A proposed set of best practices for team performance measurements specifies that assessment methods should (1) be designed to focus on processes and outcomes, (2) meet a specific goal, and (3) be linked to the specific scenario or context. Additionally, measurements should focus on observable behaviors and capture multiple levels of performance, i.e., both team and individual performance (Rosen et al. 2008). To reduce measurement errors and get a more robust picture of team performance, triangulation of qualitative and quantitative methods is then recommended to accommodate multiple perspectives into the analysis. Also, if observers are used, they should be trained and use structured observation protocols to enable attention focus and, hence, increase the measurement reliability (Rosen et al. 2008). It is recommended to use metrics with psychometric validity and general applicability to ensure comparability and utility, although the latter requirement may result in metrics becoming too general and ultimately nonsensical (Valentine et al. 2015). Designing a set of valid metrics, while balancing between generalizability and accuracy is a challenge that cannot be underestimated. A set of recommendations has been proposed that intend to give novice researchers guidance into what techniques to employ for assessment of team cognition, for instance when to choose observer-based techniques over self-assessments (Wildman et al. 2013). Among the more successful attempts at creating generic assessment instruments are the Team Diagnostic Survey (Wageman et al. 2005) and the comprehensive Aston Team Performance Inventory (Dawson et al. 2006) which measures team performance on 18 dimensions using an online survey.

To evaluate team performance, there must also be a solid model of what team performance means in the context of a particular study. Despite a plethora of available team performance models, there is no one-size-fits-all approach. Careful consideration is advised to select and tailor the measurement approach to the setting in which it will be used (Kendall and Salas 2004). In an attempt to demystify team cognition and team performance, based on an extensive review of relevant literature, the *Big five* framework proposes five core components of generic teamwork: (1) team leadership, (2) mutual performance monitoring, (3) backup behavior, (4) adaptability, and (5) team orientation (Salas et al. 2005). The main point of the framework is that efficient teams cannot rely solely on task work such as interaction with tools and systems, but must also engage in teamwork activities such as cooperation and communication.

While a complete and agreed-upon set of factors for team performance has not been identified, there are many approaches that highlight various aspects of team performance and its drivers, e.g., a Korean study has shown that culture, creativity, and collaborative practices influence team performance (Yoon et al. 2010). Another comprehensive review of existing models has been presented by NATO[1] together with their attempt at creating an instrument for measuring team effectiveness: *Command Team Effectiveness (CTEF)* (Essens et al. 2005). One of the most complex and ambitious models presented in the NATO review is the QinetiQ model of team effectiveness, emphasizing interactions between teamwork, taskwork, and leadership (Shanahan 2001). The QinetiQ model is arguably too complex to operationalize, with more than 40 different variables that affect team performance directly or indirectly. The somewhat simpler CTEF model is based on subjectively selected variables that have been identified by the NATO HFM-087 workgroup[2] to contribute significantly to command team effectiveness, and to be measurable. In addition, the model tries to reduce the overlap between variables (Essens et al. 2005). The model includes conditions, processes, and outcomes to provide a holistic view on team effectiveness, which distinguishes the model somewhat from other models that focus more explicitly on team performance, such as the *Big five* model (Salas et al. 2005).

When designing CTEF, the NATO workgroup included a questionnaire-based measurement instrument that has been used for measurement of military command teams' effectiveness, reportedly with great success (e.g., Thunholm et al. 2014). Several other instruments have been designed for similar purposes, trying to capture team performance in a specific context. One such factor that is commonly attributed as a predictor of team cognition and performance is workload (Bowers et al. 1997). Research suggests that there is a complex relationship between workload and performance, i.e., to some degree there is a positive correlation between workload and performance, but when the workload becomes too high, it induces negative stress and consequently performance will drop (Hancock and Szalma 2008). More refined instruments have been created to measure specific components of teamwork or moderating factors thereof, such as the NASA[3] Task Load Index (NASA TLX), which focuses explicitly on workload by measuring individual (1) mental

[1] North Atlantic Treaty Organization.

[2] NATO Human Factors and Medicine Panel 087.

[3] National Aeronautics and Space Administration.

demand, (2) physical demand, (3) temporal demand, (4) performance, (5) effort, and (6) frustration level on interval scales (Hart and Staveland 1988). The TLX has been used both on individual and team levels; however, a recent study has identified weaknesses in aggregating individual metrics for team analysis and proposes a separate workload index for teamwork-related workload, since metrics designed primarily for individual use are unable to account for interpersonal interactions (Nonose et al. 2016).

1.1 Objectives

This article presents a study designed to compare two generic team performance assessment techniques: *self-assessments reports* (SAR) rated by the team members themselves and *behavioral observer scales* (BOS) rated by expert observers. Team performance is an abstract phenomenon that requires contextualization to be meaningful. In this study, a small and manageable set of measurable variables has been developed that, based on reviewed literature on traditional teams, appears relevant to performance in virtual teams. It is important to note that the intention has not been to include all variables that the literature has to offer, but rather to include common factors that relate to both individual processes and team processes, while being measurable using both SAR and BOS techniques. Specifically, the research attempts to resolve to what level the two methods are interchangeable in terms of expected output.

2 Method

To study the correlation between SAR and BOS ratings, a controlled team performance quasi-experiment was set up with a total of eight sessions, each including one virtual team exposed to tasks that were assessed using both SAR and BOS ratings. The teams were given tasks requiring communication and collaboration via online collaboration tools. With eight teams conducting six separate challenges each, a total of 48 sessions were recorded. Two of the recordings had to be discarded due to methodological errors, leaving the dataset with 46 usable session recordings, each with three pairs of SAR–BOS data for a total of 138 pairs of individual assessments of teamwork in the dataset.

2.1 Study scenario

Each virtual team consisted of one coordinator, interacting with one nearby casualty, and two medical experts assisting the coordinator remotely through online collaboration tools, see Fig. 1. At each of the three locations, an observer was colocated with the local team member. The observers monitored, and assessed, the performance of the team based on observations on the single individual located at their respective site and the interactions between that individual and the rest of the team.

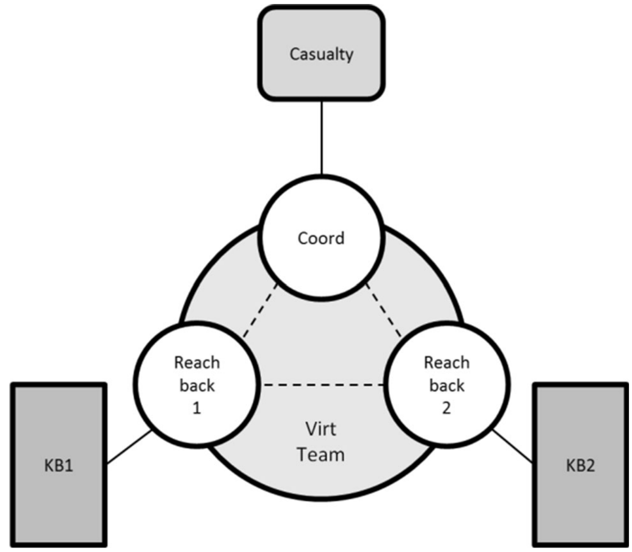

Fig. 1 Each virtual team consisted of one coordinator in direct contact with a casualty and two reachback medical experts with disparate knowledge bases (KB) communicating only through online collaboration tools

All six challenges related to chemical, biological, radiological, and nuclear (CBRN) incidents. The challenges were paired in three scenarios of two parts each: first a diagnosis challenge and then a treatment challenge. The three scenarios were given to the teams in permutated ordering. The scenarios were designed to be similar in an attempt to reduce any impact that the scenario ordering might have on the results.

All three scenarios were inspired by actual CBRN incidents, with one case of radiation poisoning, one of cyanide poisoning, and one of nerve gas poisoning. The scenarios were developed to give a sense of realism, with scripted vital readings extracted from real medical cases or studies. In every scenario, the coordinator encountered a casualty of an unknown chemical or radiological incident. The casualty was roleplayed by the study controllers at the Naval Postgraduate School site, who fed the coordinator scripted readings in response to particular examinations. As the coordinator lacked medical expertise to diagnose and treat the victim, assistance was given by the two remotely located medical experts. To enforce a need for collaboration, the two medical experts were given customized handbooks at the start of the study, each containing partial information on how to diagnose and treat twelve different

CBRN cases. The two handbooks contained some overlapping, and some unique, information. The handbooks were designed so that the only way for the team to successfully diagnose and treat the patient was to combine information from both handbooks based on the coordinator's examination reports. The teams were restricted to the provided tools for communication: an online video conferencing service and a textual chat service.

As the coordinator was the only person able to see and interact with the casualty, the experts communicated instructions for acquiring specific information relevant for them to be able to diagnose the patient, e.g., blood pressure, oxygen saturation, pulse, respiratory rate, and dilation. When the experts had acquired enough information, or after a maximum of 6 min, the team had an additional 2 min to discuss and agree on a single diagnosis. Before time ran out, the teams provided their primary diagnosis and up to two alternative hypotheses that they were not able to rule out.

After completion of the diagnosis questionnaire, the correct diagnosis was revealed to give all teams the same ability to proceed based on the correct diagnosis regardless of the outcome of the preceding challenge. Thereafter, the treatment challenge began. The procedure for the treatment challenge was similar to that of the diagnosis: The medical experts had sections of the treatment manual that were incomplete and needed to combine their knowledge to instruct the coordinator to conduct the right steps to treat the casualty. The number of tasks that the team should complete varied between six and nine, depending on the diagnosis.

For each challenge, time to completion was registered (or failure to complete), as well as the outcome of their tasks. These results constitute an objective metric of the teams' task performances. The objective scores were used for cross-referencing and correlation analysis with team performance, as positive task outcome was expected to correlate with good team performance.

The study was designed to impose ecological validity through medical accuracy and realistic scenarios. The virtual teams consisted of one forward agent (the coordinator) and multiple remote reachback experts (the medical experts). The setup thus resembles the US Navy battlefield medical scenarios where remote collaboration and distributed command and control have been investigated (e.g., Bordetsky and Netzer 2010). Similar remote just-in-time expertise has been proposed as a potential life-save in the emergency management domain, exemplified, e.g., by the Tokyo subway bombing in 1996 in which a remote physician was able to identify that the perceived bomb attack was actually a chemical attack and consequently could prevent further disaster by informing the emergency services before they entered the area (MacKenzie et al. 2007).

2.2 Participants

Eight teams of three members were selected from student volunteers. For the coordinator role, US military students at the Naval Postgraduate School (NPS) in Monterey, California, with documented skills in leadership and experience from military teamwork in the Navy, Army, and Marines. Two of these eight students were females, and six were males. Two NPS students reported previous experience from activities relating to CBRN incidents. The age range of these students was 25–40 years. The reachback roles were manned by senior nursing students at the Hartnell College in Salinas, California. This group of students was mixed gender, and they ranged between the ages of 20 and 30. As senior students in a relatively small college class, they did have prior experience of working together. None of the 16 nursing students reported any prior experience or knowledge from CBRN diagnosis and treatment.

The observer roles at the reachback sites were manned by one nursing dean and two nursing instructors from Hartnell College, all females with plenty of experience assessing nursing teamwork. Since they had several years of documented experience from teaching and grading students in nursing tasks involving teamwork, they were deemed well-equipped to assess the teams' performances in this study. At the Hartnell site, two of the three available observers were actively involved in each challenge. At the Naval Postgraduate site, one observer was used for the duration of the study. This observer was a male graduate student at the NPS Information Sciences department with a special interest in studying military team behavior and a high proficiency in using online collaboration tools for virtual teams.

2.3 Data collection

Four main data sources were used during the study: (1) team members' self-assessments, (2) observers' ratings, (3) communication recordings, and (4) outcomes-based task score.

The self-assessments were conducted as a 16-question survey after each challenge, using five-point Likert scales to rate their own interactions with the team and the task. The same 16 questions were repeated after all six challenges.

The observers were instructed to monitor team performance (explained as a combination of teamwork and task-based outcomes) and continuously take notes during the challenges. Following each challenge, they answered post-challenge surveys with a set of 14 BOS questions. These questions were formulated in the same way as the self-assessment questions, also using a five-point Likert scale. It should be noted that the observers were allowed to read the questions prior to the study, in order to increase their understanding of where to focus their attention while monitoring the challenges. The scenarios were kept to a maximum of 8 min and performed back-to-back in one session. During short breaks between each challenge, the participants completed post-challenge surveys immediately after finishing their tasks in order to reduce the risk that they would forget or neglect performance trends (Kendall and Salas 2004). Each observer survey was paired with the corresponding team member survey after completion to allow pairwise comparisons.

After each team had completed the sixth challenge, both the team members and the observers answered an additional survey of post-study questions to complete the data collection. In addition, a pre-study survey was completed by each participant prior to the first scenario, in order to collect some minimal demographics. The timing of each survey in relation to the scenarios is depicted in Fig. 2 below.

The post-challenge survey consisted of 16 and 14 questions, for team members and observers, respectively. All questions were inspired by the Crew Awareness Scale (McGuinness and Foy 2000), NASA task load index (TLX) (Hart and Staveland 1988), and Bushe and Coetzer's survey instrument (1995). These instruments were selected as a baseline for the surveys both because they were available and familiar to the study controllers, and because they were considered feasible to use for assessment, in terms of effort.

Table 1 presents the survey items for both the team members (labeled x) and for the observers (labeled y). Question x_0 used an 11-point Likert scale, whereas all other questions used five-point Likert scales. The scale was designed so that a high score is expected to correlate with strong team performance for all items, except items 1, 2, and 10 where the opposite relationships were expected. All questions are related to individual or team performance. There is a significant overlap in what the team members and the observers were asked to assess, as depicted by the table.

In addition to the collected survey data, intra-team communication (auditory and textual chats) was recorded for the entire duration of each challenge. These data were collected by tapping into the collaboration tool that the team used to communicate during the challenges. These recorded interactions may prove useful for future analysis; however, they have not been included in this study and are mentioned here for the sake of completeness only. Likewise, x_0, x_{14} and x_{15} have not been used in this study. For each challenge, an observer assessment score of the overall performance (z_{obs}) was calculated as the mean of the y_{16} ratings for that challenge.

The fourth and final data source is the objective outcomes-based task performance measure, which consists of a record of whether the challenge was successfully completed and the time it took the team to complete the task. For the diagnosis challenges, the teams were given a base score of 10 points if their main hypothesis was the correct diagnosis. Three points were deducted per erroneous alternative hypothesis (n_a) that they could not rule out (max. 3), and thus task score = 10 – 3 * n_a. If their primary hypothesis was incorrect, but their alternative hypotheses contained the correct diagnosis, they were instead given 3 points minus the number of erroneous alternative hypotheses: task score = 3 − n_a. A complete failure to diagnose the patient resulted in a task score of zero. In each diagnosis scenario, the base score was adjusted by a time score that was calculated as time score = (1 − t_D/120) * 10, where t_D corresponds to the time needed (in seconds) for the team to agree on a primary diagnosis, after the time to interact with the patient had run out, max 120 s.

The treatment challenges consisted of a set of tasks that the experts needed to convey to the coordinator, which then had to be performed on the patient. The base performance score for the treatment challenge was 10 points, with 1 point deducted for every error that occurred (n_e), where an error was either a failure to conduct one task, incomplete/failed attempt at conducting one task, or completion of a task that was not a part of the scripted treatment program: task score = 10 − n_e. The treatment challenge base score was then moderated by time just as for the diagnosis challenges: time score = (1 − t_T/360) * 10, where t_T represents the time needed to complete the treatment challenge (max 360 s).

Both in the case of diagnosis and treatment challenges, the total score (z_{task}) was calculated as task score + time score. The aggregated score represents a speed/accuracy trade-off that is commonplace both in work situations and sports (Fairbrother 2010), with the caveat that the designed

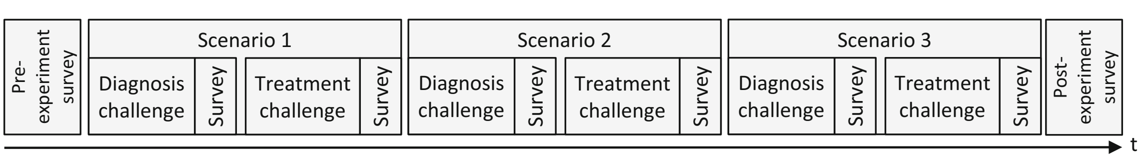

Fig. 2 Each team was exposed to three scenarios consisting of one diagnosis challenge and one treatment challenge. Surveys were completed after each completed challenge in addition to the pre- and post-study surveys

Table 1 Full wordings for the SAR and BOS questions

SAR	BOS	Wording as surveyed	Unit of analysis
x_0		How confident are you personally in your team's diagnosis/treatment?	Individual
To what extent do you agree with the following statements			
x_1	y_1	*The workload was high for [me individually/the subject]*	*Individual*
x_2	y_2	*The workload was high for the team*	*Team*
x_3	y_3	*The workload was evenly distributed within the team*	*Team*
x_4	y_4	*The stress level was low for [me as an individual/the subject]*	*Individual*
x_5	y_5	*The stress level was low for the team members in general*	*Team*
x_6	y_6	*The teamwork was effective*	*Team*
x_7	y_7	*All team members participated equally*	*Team*
x_8	y_8	*[I/The subject] was active in the team's decision-making process*	*Individual*
x_9	y_9	*All team members were active in the team's decision-making process*	*Team*
x_{10}	y_{10}	*[My/The] team was wasting time*	*Team*
x_{11}	y_{11}	*The team's efforts were well coordinated*	*Team*
x_{12}	y_{12}	*The team was clear in its communication*	*Team*
x_{13}	y_{13}	*The team was efficient in its communication*	*Team*
x_{14}		I am satisfied with the team's performance	Individual
x_{15}		I felt part of the team	Individual
	y_{16}	The team performed well	Team

The questions rated by both methods (1–13) are marked italic. Different wordings are indicated by brackets, with SAR followed by BOS after the solidus

metric does not necessarily correspond to the prioritizations made by the teams and as such the scoring system. These data were captured by the experiment leaders and kept for reference and benchmarking.

2.4 Analysis

Before any analysis, the scale was inverted for questions 1, 2, and 10 in both sets, to get a homogenous dataset where positive correlations are expected to have a positive impact on team performance. The team members' self-assessment scores were compared to the observers' ratings for all process variables (variables 1–13 in Table 1). The comparison was done by calculating the bivariate correlation coefficients on each set of variables. Correlation strengths have been categorized according to Table 2 below. t tests were then computed to see whether there are significant differences in how the choice of method affects the results, at $\alpha = .05$. The p values where adjusted using the Holm–Bonferroni method to compensate for the multiple hypotheses problem, to reduce the risk for false positives (Holm 1979).

Table 2 Intervals of $|r|$ used for analysis of correlation strength between individual items rated using SAR and BOS methods

Correlation strength	Interval
Strong	$\lvert r\rvert \geq .90$
High	$.70 \leq \lvert r\rvert < .90$
Moderate	$.50 \leq \lvert r\rvert < .70$
Low	$.30 \leq \lvert r\rvert < .50$
Weak	$.15 \leq \lvert r\rvert < .30$
None	$\lvert r\rvert < .15$

The objective team performance score reflects only task-based aspects of team performance. As a contrast to this measurement, the subjective y_{16} variable represents the observers' overall assessments of team performance. The aggregation of y_{16} for all observers relating to one team represents a more comprehensive take on team performance, with the caveat that the ratings are subjective. The task scores have been compared with the aggregated observer scores to determine whether there is indeed a positive correlation between team performance as interpreted by the observers and task performance. This correlation test was done at the team level, since team is the unit of analysis that both metrics were designed for. Thus, each completed challenge produced one tuple of data for a total of $N = 46$ samples.

The independent SAR and BOS variables were then fitted to both the z_{task} and the z_{obs} scores through multiple regressions. The glmulti package (Calcagno and de Mazancourt 2010) for R was used to identify the best-fit regression model using exhaustive search optimizing on the Akaike information criterion (AIC) instead (Akaike 1974).

All results in this study were computed using IBM SPSS and R.

3 Results

The results section objectively describes significants and other noteworthy results, followed by undisputable interpretations. The first part of the section focuses on relationships between team members' self-assessments and observer ratings of specific phenomena. The second part focuses on how the measurements fit in with the overall assessment of team performance.

Table 3 below presents the descriptive statistics for SAR (x) and BOS (y) ratings of each individual team process item from the surveys. All items, using both methods, produced ratings over the entire scale. Table 3 further presents the computed correlations, Pearson's r, between SAR and BOS ratings and finally a t test identifying systematic differences between the two rating systems. Data show that 10 of the 13 variables are correlating ($|r| \geq .15$). Seven correlation scores are classified as low according to Table 1, in order of correlation strength: *coordination efficiency*, *communication efficiency, communication clarity, teamwork effectiveness*, *time utilization, participation equality*, and *team participation in decision making (DM)*. The other three identified correlations are classified as weak, in strength order: *workload distribution*, *team's aggregated stress level*, *individual participation in DM*. Interestingly, the variables for *workload* and *individual stress* do not give evidence to any correlations between the self-assessments and the observer ratings ($|r| < .15$). The paired t tests with Holm–Bonferroni adjusted probabilities ($\tilde{p}$) show a tendency that the team members rate *participation equality* higher than the observers do ($\Delta M_7 = .28$); however, at $\tilde{p} = .055$ this difference is not strong enough to be statistically significant. No other differences could be statistically verified between the ratings produced by the two methods on any of the questions.

Table 4 Descriptive statistics for overall team performance metrics, normalized to a five-point scale

Team performance metric	N	min	max	M	SD
z_{obs}	46	1.67	5.00	3.76	.89
z_{norm}	46	1.00	5.00	2.82	1.02

The z_{task} and the z_{obs} metrics represent different views of what team performance is with z_{obs} representing the view that observers are the best jurors of team performance; and z_{task} represents the view that team performance ought to be judged by its produced outcome. The two collected metrics have been compared at the team level across all challenges, to show that a moderate positive correlation ($r = .54$, $df = 45$) exists between the two investigated team performance scores. The descriptive of both metrics (after normalization to fit the one–five scale used for the surveys) is presented in Table 4. The normalized task performance metric, z_{norm}, was calculated through linear transformation of z_{task}.

Table 5 presents correlation between each individually measured value and the overall performance scores, with z_{obs} on the left-hand side and z_{task} on the right-hand side. The table shows that all measurements, except x_1 and x_2, correlate significantly with the mean observer rating. High correlation was identified with four variables ($y_{6,11–13}$), moderate correlation with nine variables ($x_{6–7,11–13}$,

Table 3 Descriptive statistics, pairwise correlations, and paired t tests between process variables assessed both by observers and by team members ($N = 138$, $df = 137$)

n	Description	M_x	SD_x	N_y	M_y	SD_y	R	ΔM	$\tilde{p}$
1	Low individual workload	3.33	.97	138	3.12	.88	−.01	−.21	.764
2	Low team workload	3.21	.98	138	2.99	.96	−.14	−.22	.812
3	Even workload distribution	3.59	.83	138	3.48	.92	.29*	.12	1.00
4	Low individual stress level	3.21	1.04	138	3.03	1.07	.07	.18	1.00
5	Low team's aggregated stress level	3.22	.89	138	3.17	1.02	.23*	.04	1.00
6	Effective teamwork	3.75	.85	138	3.62	1.13	.39**	.12	1.00
7	Equal participation among team members	3.78	.73	138	3.51	1.12	.34**	.28	.055
8	Individual participation in decision making	3.83	.76	138	3.88	.86	.23*	−.05	1.00
9	Team's equal participation in decision making	3.82	.73	138	3.71	.92	.31**	.11	1.00
10	Time utilization	3.72	1.01	138	3.78	1.21	.37**	.06	1.00
11	Well-coordinated team efforts	3.62	.89	138	3.57	1.09	.49**	.04	1.00
12	Clear communication	3.67	.91	138	3.59	1.13	.39**	.08	1.00
13	Efficient communication	3.61	.95	138	3.59	1.15	.45**	.02	1.00

Correlation strengths (see Table 2): * weak; ** low

Table 5 Pairwise correlations (Pearson's r) between process variables assessed both by observers and by team members and overall team performance metrics z_{obs} and z_{task}, respectively ($df = 138$)

n	Description	x, z_{obs}	y, z_{obs}	x, z_{task}	y, z_{task}
1	Low individual workload	.11	.32**	.00	.25*
2	Low team workload	.14	.29*	.04	.16*
3	Even workload distribution	.31**	.50***	.09	.25*
4	Low individual stress level	.29*	.41**	.15	.19*
5	Low team's aggregated stress level	.38**	.42**	.20*	.23*
6	**Effective teamwork**	**.60*****	**.77******	**.44****	**.46****
7	Equal participation among team members	.52***	.56***	.24*	.21*
8	Individual participation in decision making	.34**	.44**	.28*	.25*
9	Team's equal participation in decision making	.48**	.61***	.19*	.26*
10	Time utilization	.41**	.53***	.27*	.30*
11	**Well-coordinated team efforts**	**.60*****	**.77******	**.39****	**.44****
12	**Clear communication**	**.52*****	**.78******	**.40****	**.42****
13	**Efficient communication**	**.63*****	**.77******	**.35****	**.41****

Bold text style indicates the overall strongest correlating variables

Correlation strengths (see Table 2): * weak; ** low; *** moderate; **** high

$y_{3,7,9\text{–}10}$), low correlation with nine ($x_{3,5,8\text{–}10}$, $y_{1,4\text{–}5,8}$), and weak with two (x_4 and y_2). All correlations with the task score are classified as low, weak, or nonexistent, as shown in Table 5, right side. Low correlations can be found with eight of the variables ($x_{6,11\text{–}13}$, $y_{6,11\text{–}13}$) while ($x_{5,7\text{–}10}$, $y_{1,3\text{–}5,7\text{–}10}$) correlate weakly with z_{task}.

Table 6 shows multiple regression models from the team members' (x) and observers' (y) assessments of each independent variable fitted against the dependent variable z_{obs}. The underlying assumption for the linear regression is that the respondents have interpreted the Likert scales as points on a linear continuum. The best identified model for x contains seven of the original 13 variables, $F_x(7, 130) = 21.82$, $p < .001$, with adjusted $R^2 = .52$. The best model generated for y contains four variables, $F_y(4, 133) = 75.59$, $p < .001$, with adjusted $R^2 = .69$. Homoscedasticity was confirmed for both models by using the Breusch–Pagan test (1979), $\chi_x(1) = 6.73$, $p = .009$ and $\chi_y(1) = 12.62$, $p < .001$. Visual inspection of the QQ-plots confirms normality of the studentized residuals for both models (Atkinson 1987), see Fig. 3.

Table 7 shows the regression models generated using the same methodology as above, but toward z_{task} instead of z_{obs}. The z_{task} models display less predictive power, as may be expected since the models were fitted against a metric that focuses solely on output (task performance) and consequently does not include anything relating to team cognition: $F_x(4, 133) = 11.81$, $p < .001$, with adjusted $R^2 = .24$ and $F_y(4, 133) = 11.59$, $p < .001$, with adjusted $R^2 = .24$. Homoscedasticity could not be confirmed for either model using the Breusch–Pagan test (1979), $\chi_x(1) = .012$, $p = .914$ and $\chi_y(1) = 1.07$, $p < .301$. The QQ-plots of the studentized residuals for both models are provided in Fig. 4; however, their interpretations are irrelevant since the models are being rejected due to heteroscedasticity.

Table 6 Multiple linear regression from variables $x_{1\text{–}13}$ and $y_{1\text{–}13}$ to z_{obs}

n	Description	β_x	SD_x	p_x	β_y	SD_y	p_y
	(Intercept)	.16	.35	.648	1.59	.25	<.001*
1	Low individual workload				.16	.05	.002*
3	Even workload distribution	−.24	.08	.006*			
4	Low individual stress level	.18	.05	.001*			
6	Effective teamwork	.24	.11	.029*	.29	.07	<.001*
7	Equal participation among team members	.24	.11	.035*			
9	Team participation in decision making (high)	.19	.10	.063	.17	.06	.007*
11	Well-coordinated team efforts	.22	.11	.040*	.26	.07	<.001*
13	Efficient communication	.15	.10	.156			

The regression coefficient (β), standard deviation (SD), and p value are listed for each variable included in the final model

* Significant at $p < .05$

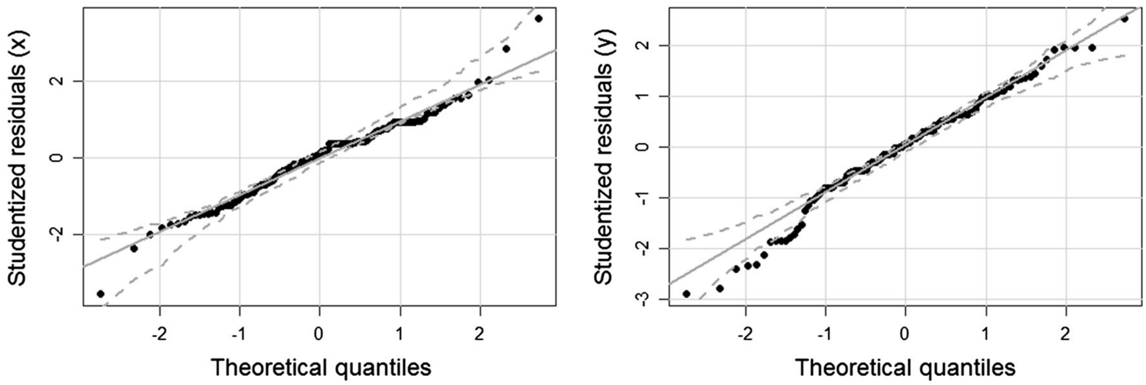

Fig. 3 QQ-plots of the studentized residuals for the best-fit regression models of team members' self-assessment parameters (*left*) and observers' ratings (ratings) against the mean overall observer score

Table 7 Multiple linear regression from variables x_{1-13} and y_{1-13} to z_{task}

N	Description	β_x	SD_x	p_x	β_y	SD_y	p_y
	(Intercept)	−3.12	2.48	.211	4.53	2.08	.031*
1	Low individual workload				.75	.44	.093
3	Even workload distribution	−1.53	.55	.006*			
4	Low individual stress level	.81	.37	.029*			
6	Effective teamwork	2.93	.55	<.001*	1.66	.64	.011*
7	Equal participation among team members				−.86	.47	.069
8	Individual participation in decision making	1.08	.57	.061			
11	Efficient coordination				1.05	.62	.092

The regression coefficient (β), standard deviation (SD), and p value are listed for each variable included in the final models

* Significant at $p < .05$

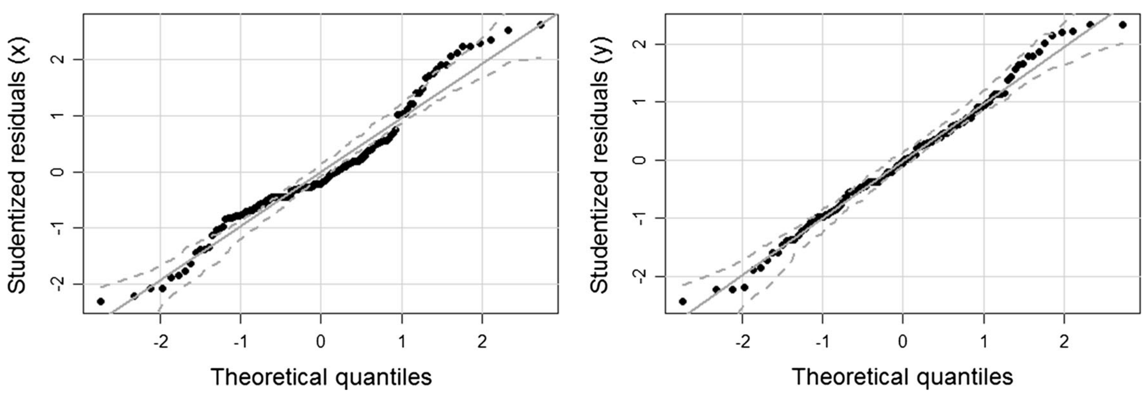

Fig. 4 QQ-plots of the studentized residuals for the best-fit regression models of team members' self-assessment parameters (*left*) and observers' ratings (ratings) against the task score

4 Analysis

The team-level comparison between task performance and observers' overall assessment of team performance unsurprisingly shows moderate positive correlation. This effect is expected since the task outcome is indisputably one of the central components of performance; in fact, it is often the very reason for the performance.

The correlation tests between SAR and BOS variables arguably generate more interesting, and also more surprising, findings. While the statistics show a significant correlation on most comparisons, the correlation strengths are classified as low at best. This implies that on a case-by-case basis one should not expect observers and team members to agree on any of the ratings in the tested surveys; however, when scaled up to a larger population, some level of convergence can be expected. Since there is no easily measured objective ground truth for intrinsic measurements such as the ones in the survey, determining which of the two methods produces more true results is, for the majority of the questions, very hard, if not impossible.

The most striking differences between SAR and BOS outputs are that SAR rate the degree of participation in the team's efforts higher on average than BOS, while BOS seem to rate workload higher than SAR. By the assumption that there is a real correlation between team performance and workload, the data suggest that the observers' ratings are more accurate on this particular point, since BOS data display at least some correlation against the collected team performance scores, whereas the team members' self-assessments of workload do not correlate with either. For the ratings of team members' individual participation in the teams' efforts, the results are less clear, making it difficult to assess the validity of these measurements.

The regression analysis showed that the individual SAR and BOS ratings can be used to predict the observers' overall team performance ratings with reasonable accuracy, but not task scores. This result is expected since the task score is affected by many factors that the ratings did not capture very well, e.g., learning, strategy, problem-solving. Also, the task score metric did not capture anything that relates to the internal processes of teamwork, which further may have contributed to the deviation between ratings and task scores. The relatively high accuracy of the two regression models fitted against z_{obs} suggests that despite their individual differences, both SAR and BOS techniques make sense for measuring team performance. For them to reflect task performance though, one must carefully define what is meant by task performance and align the SAR and BOS surveys thereafter.

Teamwork, coordination, and communication have been shown to be the strongest contributors to overall team performance in the virtual team setting that the experiment scenario provides. Of these, teamwork effectiveness has the largest β coefficient both in the SAR and the BOS dataset. Although BOS ratings of communication efficiency and clarity both have high correlation with the observer's overall performance ratings, the clarity rating did not improve the regression models, which is reasonable since the clarity may be interpreted as a precondition to efficient communication. Interestingly, the SAR ratings indicate a negative relationship between even workload distribution and team performance. This effect is hard to interpret as the SAR ratings on individual and team workload do not correlate with team performance. Also noteworthy is that the BOS ratings on low individual workload is a predictor of team performance, i.e., overloaded team members have a negative impact on team performance.

The inability of finding regression models toward the task score shows that the task scores depend largely on factors that the SAR and BOS variables did not capture. Despite these discrepancies, the observers were able to incorporate task outcome in their final team performance verdict, as shown by the relatively high correlation between the observers' overall team performance ratings and the task scores.

5 Conclusions

The comparisons presented in this article between SAR and BOS techniques show that the two methods do produce different results, but also that the results are related to each other. Both methods put teamwork effectiveness at the top of the list of the most important variables for determining team performance. The most striking differences that can be statistically verified are that team members seem to underestimate their workload, while proficient observers rate the same phenomenon more adroitly. Further, that the self-assessment ratings of workload do not correlate with either performance metric in this setting is an important finding as it is a popular assumption in human factors research that such a correlation exists and that SAR is an appropriate tool to capture that workload. For the other investigated questionnaire items, the correlations between SAR and BOS ratings are surprisingly low, but it is hard to determine whether either method is superior in terms of accuracy.

One of the main problems (not only with the calculations presented in this article, but with the team performance assessment field as a whole) is that there is no agreed-upon definition of what team performance really is, nor any agreed-upon metrics for quantifying it (Valentine et al. 2015). Without such a baseline, it is impossible to determine which model is more accurate in describing reality. Therefore, before establishing any kind of recommendations on which technique to apply in a given situation, the first questions to ask should be what questions the team performance assessment is meant to answer, and under what constraints the assessments are to be made.

Measuring team performance is important to ensure development and improvement for team-focused organizations. While team performance is a non-trivial construct, recent research has provided a plethora of models and methods for measure various aspects of team performance. Self-assessment and observer-based (SAR and BOS) techniques are among the most commonly used methods to conduct such assessments. Since these methods are inherently subjective, the choice of method will have an influence on the result, an effect that is often neglected or overlooked.

Despite the generally low correlations and sometimes even minor inconsistencies between individual questionnaire items, the SAR and BOS techniques both appear accurate enough to be practically useful for team performance assessment. The key to get meaningful assessments, as pointed out, e.g., by Wildman et al. (2013), is to match metrics with purpose. To add to that perspective, our results indicate that when perception-based measurement techniques are appropriate, the choice between observer-based and self-reported measurements will impact the results, but for team performance in general there is no universal benchmark to say that either method is more accurate than the other. Therefore, for items that are reasonable to examine using either of the two techniques, the choice can be deferred to other practical issues such as accessibility, obtrusiveness, privacy infringement, and resource availability. Perhaps the most fundamental issue of concern for this choice is how the cognitive processes are affected, both by the self-reflective processes triggered by self-reporting (Wildman et al. 2013), and by the moderating effect that observation may have on performance (e.g., Merrett 2006).

Acknowledgements The authors would like to thank 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. In addition, Mr. Per-Anders Oskarsson at the Swedish Defense Research Agency assisted with advice on statistical analysis. 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.

Funding This study was supported by the US Office of Naval Research Global, Arlington, VA [Grant N62909-11-1-7019], and the Swedish Civil Contingencies Agency, Stockholm, Sweden [Grant 2012-3380].

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Atkinson AC (1987) Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. Oxford University Press, Oxford

Bordetsky A, Netzer D (2010) Testbed for tactical networking and collaboration. Int C2 J 4(3). http://dodccrp.org/files/IC2J_v4n3_B_Bordetsky.pdf

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

Breusch TS, Pagan AR (1979) A simple test for heteroscedasticity and random coefficient variation. Econometrica 47(5):1287–1294

Bushe GR, Coetzer G (1995) Appreciative inquiry as a team development intervention: a controlled experiment. J Appl Behav Sci 31(1):13–30

Calcagno V, de Mazancourt C (2010) Glmulti: an R package for easy automated model selection with (generalized) linear models. J Stat Softw 34(12):1–29. https://www.jstatsoft.org/index.php/jss/article/view/v034i12/v34i12.pdf

Dawson JF, West MA, Markiewicz L (2006) Aston team performance inventory: management set. ASE, London

Essens P, Vogelaar A, Mylle J, Blendell C, Paris C, Halpin S, Baranski J (2005) Military command team effectiveness: model and instrument for assessment and improvement (RTO-TR-HFM-087). NATO Research and Technology Organisation, Neuilly-sur-Seine Cedex

Fairbrother JT (2010) Fundamentals of motor behavior. Human Kinetics, Champaign

Granåsen M, Andersson D (2016) Measuring team effectiveness in cyber-defense exercises: a cross-disciplinary case study. Cogn Technol Work 18(1):121–143

Hancock PA, Szalma JL (2008) Stress and performance. In: Hancock PA, Szalma JL (eds) Performance under stress. Ashgate, Aldershot, pp 1–18

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

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70

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

Krumm S, Terwiel K, Hertel G (2013) Challenges in norm formation and adherence: the knowledge, skills, and ability requirements of virtual and traditional cross-cultural teams. J Pers Psychol 12(1):33–44

MacKenzie C, Hu PFM, Fausboll C, Nerlich M, Benner T, Gagliano D, Whitlock W, Lam D, Xiao Y (2007) Challenges to remote emergency decision-making for disasters or Homeland Security. Cogn Technol Work 9(1):15–24

Maynard MT, Gilson LL (2014) The role of shared mental model development in understanding virtual team effectiveness. Gr Organ Manag 39(1):3–32

McGuinness B, Foy L (2000) A subjective measure of SA: the Crew Awareness Rating Scale (CARS). Presented at the HPSAA conference, Savannah, GA

Merrett F (2006) Reflections on the hawthorne effect. Educ Psychol 26(1):143–146

Nonose K, Yoda Y, Kanno T, Furuta K (2016) An exploratory study: a measure of workload associated with teamwork. Cogn Technol Work 18(2):351–360

Rosen MA, Salas E, Wilson KA, King HB, Salisbury ML, Augenstein JS et al (2008) Measuring team performance in simulation-based

training: adopting best practices for healthcare. Simul Healthc 3(1):33–41
Salas E, Sims DE, Burke CS (2005) Is there a "big five" in teamwork? Small Gr Res 36(5):555–599
Shanahan P (2001) Mapping team performance shaping factors. QinetiQ, Fort Halstead, UK (**unpublished report**)
Smith-Jentsch KA, Zeisig RL, Acton B, McPherson JA (1998) Team dimensional training: a strategy for guided team self-correction. In: Cannon-Bowers JA, Salas E (eds) Making decisions under stress: implications for individual and team training. American Psychological Association, Washington, pp 271–297
Thunholm P, Berggren P, Wikberg P (2014) Evaluating the effectiveness of an armoured brigade staff. In: Berggren P, Nählinder S, Svensson E (eds) Assessing command and control effectiveness: dealing with a changing world. Ashgate, Farnham, pp 141–160
Valentine MA, Nembhard IM, Edmondson AC (2015) Measuring teamwork in health care settings: a review of survey instruments. Med Care 53(4):16–30
Wageman R, Hackman JR, Lehman E (2005) Team diagnostic survey: development of an instrument. J Appl Behav Sci 41(4):373–398
Wildman JL, Salas E, Scott CPR (2013) Measuring cognition in teams: a cross-domain review. Hum Factors 56(5):911–941
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 Res Dev Int 13(3):249–264