Computers in Human Behavior xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Assessing individual contributions to Collaborative Problem Solving: A network analysis approach

Zachari Swiecki[a,b,*], A.R. Ruis[a], Cayley Farrell[a], David Williamson Shaffer[a,b,c]

[a] Wisconsin Center for Educational Research, University of Wisconsin-Madison, United States
[b] Department of Educational Psychology, University of Wisconsin-Madison, United States
[c] Department of Education, Learning and Philosophy, Aalborg University, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Collaborative Problem Solving
Epistemic Network Analysis
Coding and counting
Assessment
Measurement

ABSTRACT

Collaborative Problem Solving (CPS) is an interactive, interdependent, and temporal process. However, current methods for measuring the CPS processes of individuals, such as coding and counting, treat these processes as sets of isolated and independent events. In contrast, Epistemic Network Analysis (ENA) models how the contributions of a given individual relate to the contributions of others. This article examines the communications of air defense warfare teams from an experiment comparing two different computer-based decision support systems, using this data to ask whether ENA provides a more ecologically valid quantitative model of CPS than coding and counting. Qualitative analysis showed that commanders using one system asked questions to *understand* the tactical situation, while commanders using an experimental system focused more on *actions* in response to the tactical situation. Neither of the coding and counting approaches we tested corroborated these findings with statistically significant results. In contrast, ENA created models of the individual contributions of commanders that (a) showed statistical differences between commanders using the two systems to corroborate the qualitative analysis, and (b) revealed differences in individual performance. This suggests that ENA is a more powerful tool for CPS assessment than coding and counting approaches.

1. Introduction

As the problems faced by society grow more complex, interrelated, and ill-formed, we have increasingly turned to groups and teams for solutions. In response, Collaborative Problem Solving (CPS) has been widely recognized as a vital 21st century skill (Griffin & Care, 2014). Several definitions of CPS exist in the literature, however, most share the view that it is fundamentally socio-cognitive. In CPS, framing, investigating, and solving problems is situated in a collaborative context that involves information sharing, negotiation of meaning, and more broadly, processes which attempt to establish and maintain a shared conception of the problem (Hesse, Care, Buder, Sassenberg, & Griffin, 2015; Miyake & Kirschner, 2014; Roschelle & Teasley, 1995; Rosen, 2015).

Although definitions of CPS share these features, they may disagree on whether the construct should be applied to the individual or the group. For example, Roschelle and Teasley (1995) argue that CPS is consists of solving problems *together* while building a jointly understood problem space. The Organization for Economic Co-operation and Development, on the other hand, draws on literature from CPS, collaborative learning, and organizational psychology, to argue that CPS is the capacity of an *individual* to engage in problem solving while collaborating with others (OECD, 2017). In this paper, we also operationalize CPS at the individual level and define it broadly as the socio-cognitive processes an individual uses to solve problems with others. The importance of examining CPS at the individual level is reflected in recent efforts to develop large-scale assessments of individual CPS process and performance, including the Assessment and Teaching of 21st Century Skills project (Griffin & Care, 2014) and the Programme for International Student Assessment (OECD, 2017).

While these assessments speak to the value of CPS in educational contexts, CPS skills are important in a variety of domains. Military contexts, in particular, often rely on teams to solve problems that are too complex for any individual to solve alone. For example, Hutchins' seminal study of quartermasters in the U.S. Navy (1995) showed that navigation teams have to coordinate tools and information to solve a task with cognitive demands that outpace the capabilities of any one team member. Similarly, Navy air defense warfare (ADW) teams must work together to identify potentially hostile aircraft and defend their ship in situations that are time sensitive and information-dense (Smith,

* Corresponding author. Educational Sciences Building, Room 499, 1025 West Johnson Street, Madison, WI 53706-1706, United States.
E-mail address: swiecki@wisc.edu (Z. Swiecki).

https://doi.org/10.1016/j.chb.2019.01.009
Received 29 May 2018; Received in revised form 29 October 2018; Accepted 13 January 2019
0747-5632/

Johnston, & Paris, 2004).

In this paper, we argue that in addition to being socio-cognitive, CPS is fundamentally interactive, interdependent, and temporal. As we explain in more detail below, individuals communicate with one another, share resources, and use tools while engaging in CPS. These activities mean that CPS is characterized by interactions between people and between tools. However, CPS is distinguished from other interactive settings in the sense that the contributions of a given individual are related to and influenced by the contributions of others. In other words, CPS processes are interdependent. Moreover, CPS processes unfold in time and may vary over the course of problem solving. Thus, CPS also has an important temporal dimension. Together, these three features imply that a key challenge for CPS assessments is measuring the contributions of individuals while accounting for how they relate to the contributions of other members of the team within a given span of time.[1]

Despite these implications, assessments of CPS have traditionally relied on relatively simple measures based on self or peer ratings of CPS performance (Kyllonen, Zhu, & von Davier, 2017). More recently, researchers have begun to focus on the CPS processes of individuals by examining their interactions—conversations, steps taken toward solutions, interface actions, and so on—during problem solving scenarios such as games, simulations, and online projects. While these approaches improve upon previous efforts by focusing on CPS processes, as we discuss in more detail below, the measurement techniques they employ tend to treat these processes as isolated and independent events. Thus, there is still a need for techniques that account for the interactive, interdependent, and temporal nature of CPS.

In this paper, we present a qualitative analysis of the CPS processes of individuals in the context of ADW team training scenarios. This analysis highlights the interactive, interdependent, and temporal features that ecologically valid measures of CPS should account for. Next, we analyze the same data using three quantitative techniques: two based on coding and counting (Chi, 1997; Kapur, 2011; Suthers, 2006)—a common approach to measuring CPS—and one based on Epistemic Network Analysis, or ENA (Shaffer, 2017; Shaffer, Collier, & Ruis, 2016)—an approach to measuring CPS that models the connections between the things individuals say, do, and produce in collaborative settings. Our findings suggest that ENA is a powerful tool for CPS assessment, and that coding and counting approaches are limited in their ability to model important aspects of CPS in a team setting.

2. Theory

2.1. Interaction

Problem solving is one important function of teams. While there are many different kinds of team, for the purposes of this paper, we define a *team* as a group of individuals with distinct roles who work together to reach a common objective (Salas, Dickinson, Converse, & Tannenbaum, 1992). When individuals on teams solve problems together, they interact with one another and with tools to do so. Their processes include *taskwork*—individual interactions with "tasks, tools, machines and systems" (Bowers, Braun, & Morgan, 1997; 90). But they also include *teamwork*—interactions that facilitate taskwork by exchanging information or coordinating behavior (Marks, Mathieu, & Zaccaro, 2001). As this understanding of CPS indicates, team processes are not simply the sum of individual actions; rather, individual actions *interact* with one another, creating a context independent of any single individual.

2.2. Interdependence

As several researchers argue (see, for example, Kozlowski & Ilgen, 2006), in a team setting, the interactions between individuals reify to form emergent states: stable behavioral, cognitive, or motivational states that arise through repeated interaction and, in turn, influence future interactions. For example, Reimann, Yacef, and Kay (2011) argue that a team's history of action is a resource that influences their subsequent actions. Similarly, Clark (1996) argues that as collaborative activities unfold, teams add information to the *common ground*—the set of shared knowledge and experiences that exist between people when they interact. The contents of the common ground influence subsequent actions and how those actions are interpreted (Dillenbourg, 1999).

The things that individuals contribute constitute the team's *discourse*: the actual things members of the team say, the concepts they use, the actions they take, and the artifacts they produce. But critically, because interaction produces a common ground, the discourse of the team is *interdependent*: the actions of one member directly affect the processes of other individuals on the team. For example, when one team member needs input from another in order to complete their taskwork, or when different team members must share limited resources, the actions of each individual are contingent on the actions of others. In other words, both taskwork (the actions of each team member toward accomplishing a task) and teamwork (the actions of each team member toward managing the processes of collaboration) imply that teams are interdependent interactive systems.

2.3. Temporality

The creation of emergent states and specifically of common ground mean that CPS has an important *temporal* dimension: events at any point in time are influenced by prior actions. However, this temporality is not unbounded. The influence of prior activity and shared understanding does not always span the entire history of team actions and interactions. Shaffer (2017) argues that actions and interactions are interpreted with respect to the *recent temporal context*, or the immediately preceding events (see also, Suthers & Desiato, 2012). Similarly, Halpin and von Davier (2017) argue that team interactions are *temporally clustered*, meaning that the actions of a given part of the team make certain actions by other parts of the team more or less likely in the near future. For example, when one team member asks a question, other team members are likely to respond soon after. Thus, while team processes are composed of complex interactions among individuals, the interactions that are most relevant to the interdependence of team processes are locally bounded.

2.4. Modeling CPS

The interactive, interdependent, and temporal nature of teams means that assessing the CPS processes of individuals is particularly challenging. Just as team processes consist of entangled interactions among individuals over time, the evidence teams produce is similarly entangled. A key challenge for measuring CPS, then, is to assess individual contributions in a way that accounts for how they are related to the contributions of other members of the team within a given span of time.

2.5. Coding and counting approaches

Despite these challenges, extant techniques for measuring CPS treat individual contributions as isolated and independent events. Quantitative content analysis, or *coding and counting* (Chi, 1997; Kapur, 2011; Suthers, 2006) is a widely used method that measures collaborative processes by counting the qualitative codes, survey scores, or

[1] In this paper we focus on CPS because of its particular importance in educational and professional settings (as described above). There are certainly other socio-cognitive processes that are also interactive, interdependent, and temporal, and the results we present here might generalize to those settings. However, addressing the full range of socio-cognitive settings that involve interaction, interdependence, and temporality is beyond the scope of this paper.

observer ratings attributed to individuals in collaborative activities. Researchers then use these measures to make claims about CPS processes or outcomes. For example, Hao, Liu, von Davier, Kyllonen, and Kitchen (2016) applied a coding and counting approach to the chat utterances of individuals who worked in dyads to solve problems. The authors coded the utterances for four CPS skills and examined the relationship between the frequency of these codes and outcome measures.

In some cases, researchers compute counts for a relatively large number of code variables, leading to data sets with high dimensionality. For example, to study the processes of individuals in a wiki-based CPS project, Prokofieva (2013) used an automated coding scheme to code for 15 actions such as comments and edits. To analyze this data, the author counted the action codes for each individual and applied the multivariate technique Principal Components Analysis (PCA) to reduce the dimensionality of the data to two constructs: actions taken on their own contributions and actions taken on the contributions of others. [2]

Coding and counting methods are useful because they are easy to implement and easy to understand. Of course, collecting and coding data is non-trivial, but once these processes are complete, counts and proportions are simple to calculate. Researchers can use these results alone or in combination with more sophisticated techniques to find differences between individuals or teams and explain variation in other outcome measures.

While coding and counting methods are useful, in the context of CPS they either ignore or attempt to control for the very phenomena they are trying to assess. These methods analyze the contributions of individuals to CPS, but they do so without taking into account how those contributions were connected to and dependent on the recent activity of the team. In other words, coding and counting assumes that team discourse can modeled by looking at the frequency of individual team member contributions without reference to the contributions of others on the team. This means that such techniques ignore important assessment evidence for individuals by treating this evidence as if it were produced by isolated events.

It follows that if the goal of CPS assessments is to make claims about the processes or skills of individuals, then coding and counting methods are a poor way to justify those claims. The processes of individuals in CPS contexts are interactive, interdependent, and temporally bounded, implying the same for the evidence they produce during assessments. Because coding and counting methods ignore these features, assessments that rely on them are misaligned. Even if they correspond accurately to external evaluations of team performance in some cases, they do so with a flawed model of the situation, raising both theoretical and practical concerns about their validity (Mislevy, 1996).

2.6. Temporal models

There are a number of methods that attempt to account for the interaction, interdependence, and temporality of CPS. One method, Lag Sequential Analysis (Bakeman & Gottman, 1997) can model CPS processes using the transitional probabilities between sequences of collaborative actions. For example, Kapur (2011) used Lag Sequential Analysis to compare the processes of teams who solved well-formed versus ill-formed problems. Another method, Sequential Pattern Mining (Srikant & Agrawal, 1996) can model CPS interactions by mining data for frequent sequences of collaborative actions. For example, Perera, Kay, Koprinska, Yacef, and Zaïane (2009) compared the sequences of actions that high and low achieving teams made as they participated in an online software development project. Similarly, Maldano, Kay, Al-Qaraghuli, and Kharrufa (2011) compared the sequences of actions that high and low achieving teams made as they used a digital tabletop tool to collaboratively solve problems.

While these two methods can account for the interactive, interdependent, and temporal nature of CPS contexts, they can only—at least in their current implementations— account for these factors at a single level of analysis: these methods can model the actions of an individual while accounting for their own prior actions, or they can model team actions while accounting for the prior actions of the team. However, they cannot model the actions of a particular individual while simultaneously taking into account the prior actions of that individual *and* the rest of the team.

2.7. Models of interaction

Social Network Analysis (SNA) is one method that can situate the actions of individuals in the context of their team. SNA can be used to model patterns of interaction among individuals, for example, which individuals are communicating, the frequency of communication, and the order in which the communication occurs, while simultaneously modeling properties of the team as a whole, such as cohesion (Haythornthwaite, 1996; Sweet, 2017). However, SNA only measures the *structure* of interactions between individuals on a team, not the *content* of those interactions. Because team interactions are influenced by the type of information shared, the content communicated, and the cognitive processes used by the individual team members (Gašević, Joksimović, Eagan, & Shaffer, 2019), a content neutral method like SNA provides a limited model of CPS processes.

An alternative model of interactivity in CPS is ENA. Like SNA, ENA builds network models that describe interactive, interdependent, and temporal phenomena. However, whereas SNA builds networks that describe interactions among *individuals*, ENA builds networks that describe interactions among the *discourse* of individuals. Because ENA models interactions between the discourse of an individual *and the discourse of others* within a given recent temporal context (Shaffer, 2017; Siebert-Evenstone et al., 2017), this method can account for the interactivity, interdependence, and temporality of CPS processes at the individual level of analysis while also modeling the relationships between the actions of one individual and the other members of a team. This affordance of ENA allows us to meaningfully compare coding and counting to a more theoretically valid approach for assessing the CPS processes of individuals.

2.8. Novel contributions

There is already an extensive body of work using ENA, including applications of ENA to CPS contexts. For example, ENA has been used to analyze CPS data from simulations of professional practice in domains such as engineering (Arastoopour, Swiecki, Chesler, & Shaffer, 2015), urban planning (Bagley & Shaffer, 2015), and medicine (Sullivan et al., 2018). However, this study extends prior applications of ENA to CPS in two important ways. First, the primary goal of this paper is to investigate how measuring the contributions of individuals in the context of their teams (using ENA) compares to measuring the contributions of individuals as isolated events (using coding and counting). While prior work by Siebert-Evenstone et al. (2017) has *suggested* that this affordance of ENA is valuable in collaborative contexts, this claim has not been empirically tested. Second, in this study, we apply ENA to a novel and important context. In particular, we analyze the communications of ADW teams as they managed potentially hostile air contacts in simulated peacekeeping scenarios. We chose the context of ADW teams because it is particularly well suited to assessments of CPS. In ADW teams, individuals have well-defined roles, and their activities are

[2] Another approach is to highly script the CPS activity. For example, the Programme for International Student Assessment (OECD, 2017) controls CPS scenarios by limiting team size, composition, and action. Teams consist of one human and at least one computer-controlled agent. The actions of the human are constrained to a small set of choices which allows researchers to associate each action with a particular CPS skill. These actions are then scored and counted to yield a measure of CPS skill. In other words, such assessments combine highly scripted scenarios with a coding and counting approach.

complex, interdependent, and time-sensitive (Paris, Johnston, & Reeves, 2000). Moreover, there exists a robust set of data on ADW team performance in a high fidelity simulation. This allows us to go beyond previous studies of ENA as a tool for examining collaboration by examining the *performance of professionals* in a high-performance domain, rather than student learners.

2.9. Prior analyses of the data

The data we use here was collected as part of the Tactical Decision Making Under Stress (TADMUS) program (Cannon-Bowers & Salas, 1998). One goal of this program was to understand the decision making processes of ADW teams as they managed deceptive situations, such as concealed attacks, decoys, and non-hostile aircraft responding to in-flight emergencies. An important outcome of the TADMUS program was the development of a decision support system for the commanding officers on ADW teams. These computer displays were designed to aid the decision making processes of commanders by lowering the demands on their attention and memory in high stress situations (Johnston, Fiore, Paris, & Smith, 2013).

A study by Morrison, Kelly, Moore, and Hutchins (1996) examined the effect of the decision support system on team communication by comparing teams with access to the system (experimental condition) to those without it (control condition). Using a coding and counting approach, these researchers found that teams in the experimental condition were statistically significantly less likely to talk about sensor-based information, for example, speed, location, radar type, compared to teams in the control condition. In addition, no statistically significant differences between the two conditions were found in terms of the number of tactical actions they ordered or the number of clarifications they had to make about the tactical situation. The authors did find that teams in the experimental condition clarified less information about aircraft behavior and identity in particular, but these specific differences were not reported as statistically significant. More importantly, the results described above were examined at the team level, not the individual level. Here, we build on this prior work[3] to illustrate the value that ENA can add over a traditional coding and counting methods, but we do so by examining how the contributions of individual team members, particularly the commanding officers, were affected by the presence or absence of the decision support system.

Prior work (see, for example, Kapur, 2011) has compared coding and counting to other methods for assessing CPS. Here, we take a similar approach by comparing coding and counting to ENA. Specifically, we conducted a qualitative analysis of the communications of ADW teams to understand the interactive, interdependent, and temporal nature of this CPS context and highlight important differences between commanders in the experimental and control conditions. Next, we assessed CPS using two coding and counting analyses to model the qualitative differences we found between commanders in the two conditions: (1) a univariate analysis, which uses the cumulative counts of different kinds of coded talk, and (2) a multivariate analysis using PCA on the same code variables. Finally, we conducted an ENA analysis and compared the results to both the counting and coding approaches, and to the original qualitative analysis.

[3] Entin and Serfaty (1999) used similar data to investigate the communications of ADW teams and their commanders, and developed a measure of *anticipation*. However, their analysis accounted for the ratio of information exchanged from commanders to team members and team members to commanders, but did not examine the specific kinds of team communication, which our qualitative analysis (see below) suggested was an important difference between teams.

3. Methods

3.1. Data

As part of the TADMUS project, 16 ADW teams participated in four simulated training scenarios to test the effect of a decision support system and critical thinking-and-teamwork training on team performance (see Johnston et al., 2013 for a more complete description). The goal for each team during the scenarios was to perform the *detect-to-engage sequence*. In brief, this entails developing and maintaining awareness of a tactical situation by detecting and identifying vessels, or *tracks*, assessing whether tracks are threats, and finally, deciding whether to take action towards tracks, for example, warning, covering with weapons, or engaging (Paris et al., 2000).

Each team consisted of six participants assigned to particular roles. Two participants held command roles and the remaining four held supporting roles. All participants had approximately the same military rank and similar levels of experience. The two command roles were the Commanding Officer (CO) and the Tactical Action Officer (TAO). The CO and TAO were responsible for defending the ship and making tactical decisions regarding actions toward tracks. The support roles consisted of the Identification Supervisor (IDS), the Air Warfare Coordinator (ADWC), the Tactical Information Coordinator (TIC), and the Electronic Warfare Supervisor (EWS). Their main function was to report the information necessary for the commanders to develop and maintain awareness of the tactical situation and make decisions. Their duties included reporting the detection of tracks, determining the identification of tracks (e.g., aircraft type), sending warning to tracks, and in some cases, recommending tactical actions to their supervisors.

The teams were divided into two conditions with eight teams in each condition. The conditions differed with respect to the technological support and training provided to the command roles on each team. Commanders in the control condition had access to standard technology referred to as the command and decision simulation watch-station. This watch-station provided basic information about track identification and behavior. Commanders in the experimental condition had access to a decision support system designed to provide information about track identification, behavior, threat levels, and actions taken toward the track. This system was designed to support the decision making of commanders and reduce the cognitive effort involved in maintaining tactical awareness. It was developed in response to prior work that found that the existing watch stations failed to meet the needs of commanders in high stress situations (Hutchins & Kowalski, 1993; Hutchins, Morrison, & Kelly, 1996).

In addition, all team members in the experimental condition received computer-based training to develop critical thinking and teamwork skills, and face-to-face training to develop their teamwork and after action review skills. Teamwork training included emphasizing information exchange among team members, providing backup and support to team members, using proper communications phraseology, providing situation updates, and establishing task priorities. Control condition teams received their typical combat training curriculum.

Each team, regardless of condition, participated in the same four scenarios during the experiment. Each scenario was designed to take 30 min to complete, and the order in which teams completed the scenarios was counterbalanced using a Latin square. The four scenarios differed with respect to their geographical location, types of tracks present, and the behavior that tracks demonstrated. During each scenario, teams performed the detect-to-engage sequence for multiple tracks in situations that simulated the essential characteristics of littoral warfare operations. Each scenario required individual team members to process and communicate information within their team. In addition, team members could also communicate with personnel outside of their team, such as external commanding officers and tracks. Individuals playing these external roles used scenario-specific scripts to standardize these communications as much as possible. Analyses by Johnston et al.

(2013) suggested that these scenarios were roughly equivalent in terms of the workload they placed on teams. For a more detailed description of the scenarios, see Johnston, Poirier, and Smith-Jentsch (1998) and Hall, Dwyer, Cannon-Bowers, Salas, and Volpe (1993).

Transcripts from each team/scenario combination were analyzed for this study.[4] The dataset consists of 63 transcripts from the 16 teams: 32 from the experimental condition and 31 from the control,[5] and has been analyzed in several prior studies (Foltz & Martin, 2008; Foltz, Martin, Abdelali, Rosenstein, & Oberbreckling, 2006; Johnston et al., 2013; Smith et al., 2004). Here, the transcripts were segmented into lines corresponding to turns of talk, for a total of 12,027 lines. Our units of analysis were the individual ADW team members across their four training scenarios. In total, the analysis included 94 individuals.[6] In light of the experimental design, we grouped individuals according to their experimental condition and their duties on the team: command or support. In what follows, we focus the analysis on the 29 individuals who held command roles—16 in the experimental condition and 13 in the control—because the experiment was designed to affect their performance directly.

3.2. Qualitative analysis

To investigate how the individual contributions of the commanders differed between conditions, we analyzed the transcripts qualitatively using the codes described in Table 1 below. Not being experts in naval warfare, our first step in developing these codes was to familiarize ourselves with the different stages of the detect-to-engage sequence for ADW teams as described in the existing literature (see, for example, Paris et al., 2000). We then conducted a grounded analysis of the data (Glaser & Strauss, 1967), and triangulated our findings with prior qualitative analyses that have been conducted on similar data (see, for example, Morrison et al., 1996). These codes capture aspects of the socio-cognitive nature of CPS in the ADW context. However, we note that although taskwork and teamwork are hypothesized to be important dimensions of CPS, we did not distinguish between the two in this analysis because of the nature of the data being analyzed. Specifically, because we were modeling the talk between team members, every action was both a contribution to the solution of the problem (i.e., taskwork) and simultaneously a contribution to the team process (teamwork). For example, reporting Track Behavior is an important problem step in the detect-to-engage sequence, but it also shares important information that members of the team can use to develop and maintain a shared understanding of the problem.

To code the data, we developed an automated coding scheme using the *nCodeR* package for the statistical programming language R ((Marquart, Swiecki, Eagan, & Shaffer, 2018). We used *nCodeR* to develop automated classifiers for each of the codes in Table 1 using regular expression lists. Turns of talk that matched one or more regular expressions associated with a code were annotated with that code. *nCodeR* uses two interrater reliability statistics to establish code reliability: Cohen's kappa, and Shaffer's rho (Eagan, Rogers, Pozen, Marquart, & Shaffer, 2016; Author et al., 2017a). The kappa statistic measures the agreement between two raters, while accounting for agreement due to chance. Rho measures whether the level of agreement found for a sample coded by two raters generalizes to the rest of the dataset. In this study, we used the standard threshold for kappa (> 0.65) and a rho threshold of < 0.05 to determine whether the agreement between human coding and an automated classifier was suitable.[7]

To create valid and reliable codes, we assessed concept validity by requiring that two human raters achieve acceptable measures of kappa and rho, and reliability by requiring that both human raters independently achieve acceptable measures of kappa and rho compared to the automated classifier. Table 2 shows interrater reliability measures, kappa (rho), for each code.

After validating the automated classifier for each code, we used the automated classifiers to code the data.

3.3. Code counts

For each commander, we calculated raw counts for the separate codes. Normality checks—Shapiro Wilks tests and Q-Q plots—on the code counts suggested the distributions for several codes were non-normal. To account for non-normality, small sample sizes, and to apply a consistent statistical test across cases, we used two-sample Mann-Whitney *U* tests to compare the distribution of code counts for each code between the conditions. We also conducted post-hoc power analyses on these tests using the statistical software G*Power (Faul, Erdfelder, Lang, & Buchner, 2007).

3.4. Principal Components Analysis

To investigate whether a quantitative approach that accounts for linear relationships between multiple variables could capture the qualitative differences we found between commanders in the two conditions, we performed a PCA on the code count matrix. PCA is a multivariate statistical technique that reduces the dimensionality of highly correlated data by finding linear combinations of the original dimensions that maximize the variance accounted for in the data (Bartholomew, Steele, Galbraith, & Moustaki, 2008). Because we were interested in measuring the individual contributions of commanders in the context of their teammates, we applied PCA to the count matrix for all subjects in the study.

Prior to conducting the PCA, we performed a Bartlett's test of sphericity to test whether the variables in the dataset were sufficiently correlated to warrant using PCA. The test was significant at $p < 0.05$. As part of the PCA method, the count matrix was mean-centered and each dimension was scaled to have unit variance. This scaling ensures that the variance maximized by PCA will not be due to differences in the ranges of the original variables.

To compare the individual contributions of the commanders between conditions, we analyzed the PCA scores for commanders on the dimensions, i.e., principal components with eigenvalues greater than one. In this case, the PCA returned three dimensions that meet this criteria. Normality checks on these scores suggested that distributions of scores on two of the three dimensions were non-normal. To account for non-normality, small sample sizes, and to apply a consistent statistical test across cases, we applied two-sample Mann-Whitney *U* tests to

[4] As part of the original experiment, each team/scenario combination was also scored using the Air Defense Warfare Team Observation Measure (Johnston, Smith-Jentsch, & Cannon-Bowers, 1997) which provides scores on four dimensions of teamwork behavior. We did not analyze those outcome variables here because they were measured at the team level, and our interest was in differences between individual commanders in the two conditions.

[5] Each team had 4 transcripts, one for each scenario they completed. A transcript for one scenario from a control condition team was missing from the dataset.

[6] Turns of talk for three COs in the control condition were missing from the data set, indicating that these COs were either absent from the scenarios or did not choose to speak.

[7] Rho follows the same logic as a standard hypothesis test and its interpretation is similar to that of a *p* value. In our study, the null hypothesis for a test using rho is that, for a given code, the sample of data coded by two raters was drawn from a dataset with a true kappa of less than 0.65. A rho of less than 0.05 means that the kappa value observed on the sample was greater than 95 percent of the kappa values in the null hypothesis distribution. In other words, a rho of less than 0.05 allows us to reject the null hypothesis that the true rate of agreement between two raters is below the chosen threshold (in this case, 0.65), thus supporting the hypothesis that rate of agreement over the whole dataset is above the threshold.

Table 1
Qualitative codes, definitions, and examples.

Code	Definition	Examples
Detect/Identify	Talk about radar detection of a track or the identification of a track, (e.g., vessel type).	1) IR/EW NEW BEARING, BEARING 078 APQ120 CORRELATES TRACK 7036 POSSIBLE F-4 2) TIC/IDC TRACK 7023 NO MODES NO CODES.
Track Behavior	Talk about kinematic data about a track or a track's location	1) AIR/IDS TRACK NUMBER 7021 DROP IN ALTITUDE TO 18 THOUSAND FEET 2) TAC/AIR TRACK 7031 THE HELO THAT WAS TAKING OFF THE OIL PLATFORM IS TURNED EAST
Assessment/Prioritization	Talk about whether a track is friendly or hostile, the threat level of a track, or indicating tracks of interest	1) TRACKS OF INTEREST 7013 LEVEL 5 7037 LEVEL 5 7007 LEVEL 4 TRACK 7020 LEVEL 5 AND 7036 LEVEL 5 2) CO, AYE. LET'S UPGRADE OUR THREAT LEVEL TO 6.
Status Updates	Talk about procedural information, e.g., track responses to tactical actions, or talk about tactical actions taken by the team	1) TAO ID, STILL NO RESPONSE FROM TRACK 37, POSSIBLE PUMA HELO. 2) GOT HIM COVERED.
Seeking Information	Asking questions regarding track behavior, identification, or status.	1) TAO CO, WE'VE UPGRADED THEM TO LEVEL 7 RIGHT? 2) WHERE IS 23?
Recommendations	Recommending or requesting tactical actions	1) AIR/TIC RECOMMEND LEVEL THREE ON TRACK 7016 7022 2) GB, THIS IS GW, REQUEST PERMISSION TO TAKE TRACK 7022 WITH BIRDS HOSTILE AIR. THREAT LEVEL HIGH. RANGE 7NM.
Deterrent Orders	Giving orders meant to warn or deter tracks.	1) TIC AIR, CONDUCT LEVEL 2 WARNING ON 7037 2) AIR THIS IS TAC CONDUCT LEVEL ONE QUERY TRACK 7036
Defensive Orders	Giving orders to prepare ship defenses or engage hostile tracks	1) TAO/CO COVER 7016 WITH BIRDS 2) AIR KILL TRACK 7022 WITH BIRDS

Table 2
Interrater reliability statistics.

Code	Human 1 vs Human 2	Human 1 vs Automated	Human 2 vs Automated
Detect/Identify	0.94 (< .01)	0.94 (< .01)	0.89 (< .01)
Track Behavior	0.83 (< .01)	0.94 (< .01)	0.77 (0.03)
Assessment/Prioritization	1 (< .01)	0.94 (< .01)	0.94 (< .01)
Status Updates	0.93 (< .01)	0.94 (< .01)	0.87 (< .01)
Seeking Information	0.94 (< .01)	1 (< .01)	0.94 (< .01)
Recommendations	1 (< .01)	0.92 (0.01)	0.92 (< .01)
Preparation	1 (< .01)	0.83 (0.03)	0.93 (< .01)
Deterrent Orders	0.94 (< .01)	0.93 (< .01)	0.89 (< .01)
Defensive Orders	0.84 (0.04)	0.92 (0.01)	0.92 (0.01)

Note. Agreement thresholds were kappa > 0.65 with rho < 0.05.

compare the distributions of PCA scores for commanders in the two conditions. We conducted post-hoc power analyses on these tests using the statistical software G*Power.

We interpreted the meaning of the PCA dimensions using the principal component loadings, which show how much each variable—in this case each code—contributes to each dimension.

To confirm that PCA represented an appropriate multivariable linear model of the codes, we also tested whether rotating the principal components and allowing them to correlate would yield a solution that was easier to interpret. We applied two approaches using oblique rotations to produce correlated components: one in which we rotated the original PCA solution and one in which we used an exploratory factor analysis.

3.5. Epistemic Network Analysis

Finally, we used ENA (Shaffer, 2017; Eagan et al., 2016) to test whether a quantitative model that accounted for interactions between team members could capture the qualitative differences we saw between commanders in the two conditions.

To conduct this analysis, we used the ENA Web Tool (version 0.1.0) (Marquart, Hinojosa, Swiecki, & Shaffer, 2018). The ENA algorithm uses a sliding window to construct a network model for each turn of talk in the data, showing how codes in the current turn of talk are connected to codes that occur within the *recent temporal context* (Siebert-Evenstone et al., 2017), defined as a specified number of lines preceding the current turn of talk. The resulting networks are aggregated for all turns of talk for each subject in the model. In this way, ENA models the connections that each subject makes between concepts and actions *taking into account the actions of others on the team* (Shaffer, 2017). That is, ENA models the connections that an individual made between codes, whether those connections were within their own talk, or whether they were made to things other people said or did in the recent temporal context.

We used an ENA model based on the codes in Table 1. Although our analysis was focused on individuals in the command roles, in order to account for contributions commanders made in the context of their team, all turns of talk, whether they were spoken by commanders, supporting members, or external agents where included in the analysis. We defined the recent temporal context as being five lines, each line plus the four previous lines. We chose to define the recent temporal context in this way because our qualitative analysis of the data suggested that the majority of referents for a given line were contained within four lines.

Mathematically, the collection of networks for all subjects in the analysis is represented as a matrix of connection counts. In other words, the columns of this matrix correspond to code *pairs* and the rows correspond to a point in high-dimensional space for each subject. The ENA model normalized this matrix before subjecting it to a dimensional reduction. A normalized model accounts for the fact that different subjects may have been more or less talkative during the experiment. For the dimensional reduction, we used a technique that combines (1) a hyperplane projection of the high-dimensional points to a line that maximizes the difference between the means of two groups—in this case, commanders in the experimental condition versus those in the control condition—and (2) a singular value decomposition (SVD). SVD is the same dimensional reduction technique commonly applied in PCA. The resulting space highlights the differences between groups (if any) by constructing a dimensional reduction that places the means of the groups as close as possible to the X-axis of the space.

Networks were visualized in this space using two coordinated representations for each subject: (1) a *projected point,* which represents the location of that subject's network in the low-dimensional projected space, and (2) a weighted network graph network where nodes correspond to the codes, and edges reflect the relative frequency of connection between two codes. The positions of the network graph nodes are fixed, and those positions are determined by an optimization routine that minimizes the difference between the projected points and their corresponding network centroids. Thus, projected points toward the extremes of either dimension will have network graphs with strong

connections between nodes located on the extremes. In other words, dimensions in this space distinguish subjects in terms of connections between codes whose nodes are located at the extremes and the positions of these nodes can be used to interpret the dimensions of the space. This interpretation is similar to that of dimensions in a PCA model based on coded data, but using temporally localized co-occurrences of codes rather than counts of individual codes (see Shaffer et al., 2016 for a more detailed explanation of the mathematics of ENA; see Arastoopour et al., 2015 and Sullivan et al., 2018 for more examples of this type of analysis).

To compare the individual contributions of the commanders between conditions, we analyzed the locations of their projected points in the ENA. Normality checks suggested that the distributions of projected points were normal. However, to maintain consistency between the statistical tests applied to code counts, PCA scores, and ENA points, we applied a two-sample Mann-Whitney *U* test to compare the distributions of the projected points for commanders in the control and experimental conditions, and conducted power analyses of these tests using G*Power. We interpreted the meaning of any differences by computing mean networks, averaging the connection weights across the networks in each condition. Finally, we also compared mean and individual networks using network difference graphs. These graphs are calculated by subtracting the weight of each connection in one network from the corresponding connections in another network.

4. Results

4.1. Qualitative results

4.1.1. The detect-to-engage sequence

The goal for each team during the training scenarios was to perform the ADW *detect-to-engage sequence*. This sequence begins as the team detects ships or aircraft on radar, referred to as *tracks*. Track detections are typically reported by one of the supporting members of the team.

After a track is detected, the team needs to identify it—for example, helicopter, jet, merchant ship, commercial airline—and its weapon capabilities in order to assess whether it is hostile. In addition to onboard computers and radar, the team uses *identification, friend or foe* (IFF) codes to understand track intent. IFF codes are broadcast by tracks and classified into five categories, or *modes*, that identify the track as civilian or military.

Depending on the track's behavior and identity, the team assigns it a threat level. These assignments are typically made by the commanding officers on the team, the CO and the TAO. Threat levels range from 1 to 7, with tracks at level 4 and above considered potentially hostile priorities. Based on the track's threat level, the team begins to take action toward it at the direction of the commanding officers. These actions range from deterrent messages sent to tracks, called warnings or queries, to defensive actions that prepare for and engage in combat with the track.

A typical sequence of actions would be to issue level 1 or 2 warnings to the track soon after it is detected and before it is within its weapons range. Low-level warnings declare the presence of the warship and request that the track identify itself and state its intentions. Soon after, the commanders would order the track *covered* with weapons in preparation for engagement, should it be necessary. These early actions can be critical in ADW situations because they allow the team gather information about track intent and prepare for combat before the ship is in immediate danger.

If the team receives no response to the low-level warnings and the track continues to demonstrate potentially hostile behavior, the commanders would order level 3 warnings, which warn the track that the ship is prepared to defend itself and request that it divert course. If the track does not respond to level three warnings and still fits a hostile profile, the commanders must decide whether to engage the track in combat.

Of course, in ADW scenarios, teams typically have to deal with multiple tracks simultaneously, and each track may be at different stages in the detect-to-engage sequence. As track behavior changes, teams may have to cycle through stages of the sequence multiple times for certain tracks. Moreover, in some circumstances, it may be appropriate to skip stages of the sequence in order to manage multiple tracks at once.

In addition to managing multiple tracks, teams have limited time to make decisions and take action. For example, a hostile Super Frelon helicopter can carry missiles with a range of approximately 30 nautical miles. This means that if a Super Frelon is initially detected 50 nautical miles from the warship and is traveling inbound at top speed, it could close the distance to its weapons range in under six minutes. Supersonic jets could close this distance in less than a minute. Thus, ADW teams must be able to detect, identify, assess, and act on tracks quickly in order to defend the ship against potential threats.

To make matters still more difficult, team members communicate complicated information about multiple tracks, give assessments, and issue orders over a single communication channel. Given the complex nature of the problem, time pressures, and the communication format, there are many opportunities for errors to arise. This is why the TADMUS project developed a decision support system to convey information about track behavior, track history, assessments, and the sequence of actions taken by the team more efficiently. The goal of this system was to support the decision making of the CO and TAO in these high stress situations by reducing their cognitive load.

In the examples below, we compare commanders without access to the decision support system (control condition), who received standard training, to commanders with access to the system (experimental condition), who received teamwork training in addition to the standard curriculum. Note that in the transcripts below, speakers sometimes began their turn of talk by addressing the member(s) of the team for whom the message was intended, followed by who was speaking. Thus, in the transcripts “TAO, CO” should be read as “TAO, this is CO.”

4.1.2. Example 1: control condition

In the following excerpt, supporting members of the team report different information about four new tracks, information that the commanders need to integrate in order to develop an understanding of the tactical situation.

Line	Speaker	Utterance
1	IDS	TAO TRACK 16 NO MODES AND CODES, TO THE NORTHWEST.
2	EWS	TAO TRACK 13, 14, 15. I AM GETTING PRIMUS 40 PUMA HELO.
3	IDS	TAO TRACK 13, 14, 15 NO MODES AND CODES.
4	TAO	TAO AYE.

In particular, both the IDS (line 1) and the EWS (line 2) identify new radar contacts. The IDS reports (lines 1 and 3) that none of the contacts have “modes and codes,” meaning that there is no information about whether the new tracks are friend or foe. As described previously, this designation, called *identification, friend or foe*, or IFF, is based on a set of numeric codes that identify a track as civilian or military aircraft. The EWS (line 2) reports the radar signature (“Primus 40“) and aircraft type (“Puma helo” or helicopter) for three of the contacts, tracks 13, 14, and 15. (Although it makes the situation potentially more confusing, track numbers are not assigned sequentially; tracks 13, 14, and 15 were identified *after* track 16.)

The IDS and EWS direct their information to the TAO, who acknowledges receipt of the information (line 4). However, notice that the TAO has to integrate these different pieces of information in order to understand the rapidly evolving situation.

As the situation progresses, this process only becomes more challenging. The demands of the task are such that the team cannot simply

report information, process that information, and move on without interruption; as we see below, new information can come at any time.

5	CO	TAO, CO, LET'S GO AHEAD AND ISSUE A THREAT LEVEL FOR THE PUMAS 13, 14, 15.
6	EWS	TAO, EWS TRACK 012 IDENTIFIED AS F-4.
7	TAO	LEVEL 4, AYE.
8	TAO	SAY AGAIN TRACK NUMBER F-4?
9	EWS	14. CORRECTION 12, BEARING 094
10	TAO	TAO, AYE.

In line 5, the CO steps in and asks the TAO to evaluate the threat posed by the Puma helicopters reported in lines 2 and 3. The TAO (line 7) classifies the tracks as “level 4” threats, meaning that they are potentially hostile tracks that the team should monitor.

Notice, however, that between when the CO asks for a threat assessment and the TAO replies, the EWS (line 6) reports another contact identified as an F-4 jet. The TAO then has to ask the EWS (line 8) to repeat the information because he or she was busy making the threat assessment. The EWS repeats the track number of the F-4, and adds additional information about its location (line 9). The TAO acknowledges this message in line 10.

As this excerpt shows, the members of this team are able to quickly distinguish similar sounding information (e.g., 14, level 4, F-4, 94). However, the commanders are often receiving new input while they are communicating decisions based on previous information. This means that they frequently have to request clarification from supporting members of the team to maintain an understanding of the tactical situation.

Over the next four minutes (omitted here), the team continues to detect, identify, and assess new tracks. However, despite identifying six potentially hostile tracks in the area, the TAO has yet to issue any warnings. Recall that warnings are a critical component of the detect-to-engage sequence issued at the direction of the commanders on the team.

In the excerpt below, the IDS recommends warnings on two tracks. However, the TAO is busy communicating with the EWS to clarify the identity and location of one of those tracks and misses this information.

49	IDS	TAO, ID, DO WE WANT THREAT LEVEL WARNING 3 ON 22 AND 16?
50	EWS	SAME BEARING IT'S A SUPER FRELON, THAT'S THE ONLY INFORMATION I HAVE RIGHT NOW.
51	IDS	NEGATIVE MODES AND CODES ON 16 AND 22. THEY ARE INSIDE 20.
52	TAO	EWS, TAO, UNDERSTAND SAME BEARING AS THE SUPER FRELON?
53	EWS	AFFIRM. 16, 22 ARE BASICALLY THE SAME BEARING.
54	TAO	TAO, AYE.

In line 49, the IDS addresses the TAO and suggests level 3 warnings on tracks 22 and 16. It is unusual to recommend a level 3 warning—-which is typically reserved for imminent threats—when the team had yet to issue any lower level warnings. The demands of the situation were thus impeding standard execution of the detect-to-engage sequence.

Soon after giving the recommendation, the IDS reports (line 51) that there is no identifying information for the track (“Negative modes and codes”), and emphasizes that it is nearing the weapons range for certain hostile aircraft (“inside 20 [nautical miles]“). This additional information supports the earlier recommendation (line 49) to issue a level 3 warning.

However, the TAO does not respond to this critical information, instead asking the EWS (line 52) to clarify the bearing of track 22. The TAO's need to clarify the tactical situation causes him or her to ignore time-sensitive information. As a result, despite suggestions from the IDS, the team does not take action toward two threatening tracks.

In the next few exchanges (omitted here), the team identifies track 22 as a possible helicopter flying in tandem with track 16. Having clarified the tactical situation, the TAO orders the IDS to issue the team's first warnings to their tracks of interest.

58	TAO	ID, TAO, LET'S GO OUT WITH LEVEL 1 WARNINGS ON 7013, 14, 15, 16, AND 22.
59	CO	TAO, CO, LET'S GO AHEAD AND MAKE 22 THE SAME THREAT LEVEL AS 16, LEVEL 6.
60	TAO	22 LEVEL 6, AYE.
61	IDS	SIR, THEY'RE INSIDE 20, DO YOU WANT A LEVEL 1 OR A LEVEL 3 WARNING?

In line 58, the TAO orders the IDS to send warnings to five tracks. Notice, however, that the TAO orders level 1 warnings, which are typically issued very early in the detect-to-engage sequence. This suggests that the TAO may have missed the IDS's previous recommendation for level 3 warnings (lines 49 and 51 above). In response, the CO upgrades the threat assessment for track 22 to the second highest threat level (line 59) and the TAO acknowledges this message (line 60).

At this point, the team has identified two tracks posing an imminent threat to the ship. In line 61, the IDS repeats the recommendation for a level three warning, again emphasizing the potential threat they pose: “Sir, they're inside 20 [nautical miles].” The TAO heeds the recommendation and finally (after the end of this excerpt) orders level 3 warnings to the tracks. This suggests that the TAO either missed information at critical times or was unable to quickly act on it because of the effort it took to understand the tactical situation. In particular, the TAO's efforts to understand the situation prevented the team from following the general trend of the detect-to-engage-sequence for air defense warfare, and lead to inappropriate orders given the situation.

Altogether, this example illustrates two key elements of team performance in the control condition. First, the activities of ADW teams are highly interactive. When commanders seek information and make assessments, they do so in relation to information about track detection, identification, and behavior. In such a context, it is impossible to understand or assess any one turn of talk without understanding who and what it is responding to.

Second, the example shows that the activities of ADW teams are time-sensitive and information dense, making them highly complex. At any given time, multiple conversations may be occurring and multiple tracks may be in different stages of detection, identification, assessment, or action. As a result, commanders often have to seek or clarify information in order to understand the tactical situation. This in turn means that in some circumstances, commanders are unable to take action in a timely manner.

4.1.3. Example 2: experimental condition

In this example, the structure of the team and its goals are the same as in the example above. Here, however, the two commanders have access to a decision support system that provides them information about track identity, assessment, current and past behavior, as well as a history of actions taken by the team.

At this point in the scenario, the team is managing two tracks. One of which, track 7, was just detected by the CO. Typically tracks are detected and reported by the supporting members of the team, but the availability of the decision support system enabled the CO to access this information directly, as the following excerpt illustrates.

Line	Speaker	Utterance
1	CO	OK 07 IS MOVING TOWARDS US SO WE'VE GOT TO COVER WITH GUNS OR BULLDOGS ON 07
2	TAO	NEGATIVE
3	TAO	TIC GO OUT WITH LEVEL ONE QUERY ON 07 AND COVER WITH BULLDOGS

The CO reports the detection of track 7, letting the team know it is inbound (line 1). In the same turn, the CO issues orders to "cover with guns or bulldogs [anti-ship missiles] on 07"—that is, assign weapons to the track in preparation for self-defense. After adding an order to issue a level 1 warning to the track, the TAO passes the CO's orders to the TIC (line 3).

Thus, commanders on this team are reacting to the developing tactical situation by contributing new information about the track (line 1) and immediately responding to it with early actions from the detect-to-engage sequence: warning the track and covering it with weapons (lines 1 and 3).

As the scenario evolves, the team manages another track of interest (track 36). However, the identity of Track 07 is still unknown, and it has yet to respond to level 1 warnings. The TAO has contacted the bridge of the ship for visual confirmation on the identity of track 07. As more information about the tracks comes in, the commanders respond to the developing situation.

27	ADWC	TAO, ADWC BEARING 064 TRACK 36 APPEARS TO BE INBOUND UNKNOWN MISSILE PLATFORM
28	TAO	INITIATE LEVEL ONE QUERY ON TRACK 36
29	IDS	AYE
30	BRIDGE	COMBAT THIS IS BRIDGE TRACK 7007 IS A MERCHANT SHIP

The ADWC reports (line 27) the detection of track 36, which is headed in their direction and capable of carrying missiles. The TAO follows up immediately by ordering a level 1 warning on the track (line 28). In the next line (30), the bridge reports to the team (hailing them as "combat") that they have visually confirmed track 7 as a merchant ship. Merchant ships are civilian vessels, making them unlikely threats.

31	TAO	AYE CO DID YOU COPY THAT
32	CO	YEAH I'VE GOT IT NOW WE'VE GOT 36 AND 20 COMING AT US
33	TAO	GOT THEM
34	TAO	INITIATING LEVEL ONE QUERY ON 36

Immediately after receiving the information from the bridge, the TAO acknowledges the information (line 31), and asks the CO: "did you copy that." The CO confirms receipt of the message (line 32), and then turns the team's attention to track 36, which was detected previously by the ADWC, and track 20, a new contact. The TAO confirms that they are aware of the two tracks (line 33) and reports that they are in the process of warning track 36 (line 34). Notice, however, that the TAO does not act on track 20.

In other words, there is a rapid exchange of information about multiple tracks; in response, the commanders adapt their priorities, make decisions, and take action toward potential threats. However, during this process, the TAO seems to have overlooked a potentially threatening track.

In the interim lines (omitted here), the team detects another new track and sends out the level 1 warning on track 36. After the warning goes out, the ADWC requests further action on track 36.

39	ADWC	TAO, ADWC REQUEST WE COVER TRACK 7036
40	TAO	COVER 36
41	IDS	COVER 36 AYE
42	CO	COVER 20 AS WELL

The TAO follows up on track 36 by ordering it covered with weapons (lines 39–41). Then, the CO returns the team's attention to track 20, which the TAO has still failed to address: "Cover 20 as well" (line 42).

These excerpts show that the commanders on this team responded to the tactical situation by passing information about tracks and taking appropriate actions toward them. Moreover, commanders made sure that important information was not lost. When it appeared that a potentially threatening track had been neglected, as we saw in line 34, it was the CO who made sure that team remained aware of the track.

This example illustrates two key elements of team performance in the experimental condition. First, as in the control condition, the activities of ADW teams in the experimental condition are highly interactive. However, the nature of the interactions was different because the commanders in this condition were trained in effective communication and had access to information that they did not need to acquire verbally or hold in their memory. While commanders still responded to other members of the team, there was less need to clarify communications, and thus more opportunities to contribute to the team's understanding of the tactical situation.

Second, commanders in the experimental condition not only contributed to their teams' understanding of the emerging tactical situation by passing information about tracks; they also responded to these situations in a timely manner with appropriate decisions and actions. Commanders in the experimental condition were thus better able to manage complex situations, ensuring that potentially hostile tracks were not lost from the tactical picture. The decision support system and training allowed commanders to focus less on *understanding* the tactical situation and more on *contributing to and acting on* that situation.

4.2. Quantitative results

The qualitative results above suggest that the problem solving activities of ADW teams are highly interactive. In this context, individuals respond to and build upon the contributions of others as they pass information, make decisions, and take actions. In such a context, it is impossible to understand or assess any one contribution without understanding who and what it is responding to.

These results also suggest that commanders in the control and experimental conditions contributed to their teams in different ways. Those in the control condition often needed to clarify the tactical situation by asking questions. Consequently, they were less able to take appropriate actions toward tracks. Commanders in the experimental condition were able to focus less on understanding the tactical situation and more on contributing to and acting on that situation.

While qualitative methods are well suited to providing a thick description of a small number of cases, our goal here is to determine whether and how the differences we found constitute a pattern that generalizes *within* the larger set of data we have available. In the next sections, we report and compare the results of three quantitative approaches designed to capture the qualitative differences shown in the examples above.

4.2.1. Code comparisons

Our first approach to capturing the differences in commander contributions between the two conditions was to count the contributions of each individual—that is, count their coded talk—and compare those counts. We compared the distributions of the counts for each code using two-sample Mann-Whitney U tests. Count distributions for the codes related to understanding the tactical situation are represented as box plots in Fig. 1.

The Mann-Whitney U tests showed a significant difference between the control condition (Mdn = 34.0, N = 13) and the experimental condition (Mdn = 18.5, N = 16) for Seeking Information ($U = 16.5$, $p = 0.00002$, $d = 2.0$, power = 0.99). We found no significant differences for Track Behavior, Detect/Identify, Assessment/Prioritization, or Status Updates (Table 3). These results suggest that commanders in the control condition sought more information than those in the experimental condition.

We also compared the count distributions between the two conditions for the codes related to tactical decisions (Fig. 2).

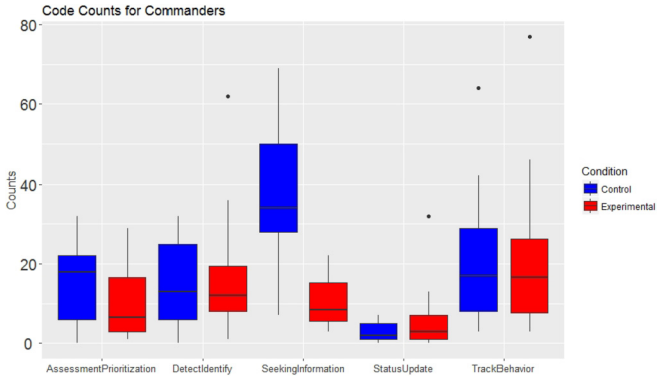

Fig. 1. Count distributions (box plots) for codes related to understanding the tactical situation, commanders in control versus experimental conditions.

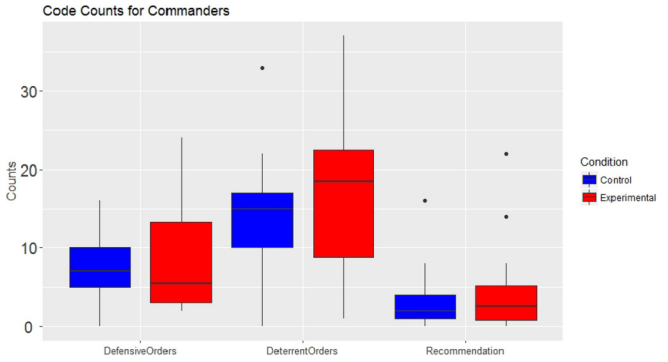

Fig. 2. Count distributions (box plots) for codes related to tactical decisions, commanders in control versus experimental conditions.

Mann-Whitney *U* tests showed no significant differences between conditions for Defensive Orders, Deterrent Orders, and Recommendation (Table 4). These results suggest that commanders in the two conditions did not differ in terms of decisions or actions they took toward tracks.

These results are unsurprising given the experimental design. If the decision support system was an effective means of providing commanders with information about the tactical situation, we would expect commanders in the experimental condition to seek less information overall.

In the qualitative examples, commanders in the control condition did indeed seek information. However, this seeking was in relation to information about the tactical situation, that is, tracking behavior, identification, and assessment. Moreover, we saw that the commanders in the experimental condition made decisions and took action in relation to changes in the tactical situation, and at times, inserted new information about tracks into the team's common ground of shared information. In this code-counting model, however, there are no significant differences in commander behavior between conditions in terms of tactical decisions or contributions to the tactical situation, only differences in terms of seeking information.

Thus, while a univariate approach is relatively easy to apply to the data, it provides a very thin description of how the commanders in the two conditions contributed differently to their respective teams; we learn only that some commanders asked more questions than others.

Table 4
Code count comparisons—tactical decisions.

Code	Control (n = 13)	Experimental (n = 16)	*U*	*p*	Cohen's *d*	Power
	Mdn	*Mdn*				
Defensive Orders	7.0	5.5	99.5	0.85	0.07	0.05
Deterrent Orders	15.0	18.5	116.5	0.60	0.21	0.08
Recommendation	2.0	2.5	109.5	0.82	0.09	0.06

Note. Results for Mann-Whitney *U* tests between commanders in the control and experimental conditions for codes related to tactical decisions. No significant results found at the $p < 0.05$ level.

Table 5
Principal component loadings.

Code	Total Talk (PC1)	Tactical Talk (PC2)	Command Behavior (PC3)
Seeking Information	−0.753	0.107	−0.207
Track Behavior	−0.491	−0.588	0.557
Status Update	−0.241	0.533	0.674
Assessment/Prioritization	−0.858	−0.089	−0.228
Deterrent Orders	−0.861	0.122	−0.079
Detect/Identify	−0.166	−0.843	0.415
Recommendation	−0.329	0.580	0.566
Defensive Orders	−0.810	−0.001	−0.336

4.2.2. Principal Components Analysis

Only the first three principal components had eigenvalues greater than 1, meaning that they accounted for more variance in the data than any original variable. Together, these dimensions account for more than 78% of the variance.

The loadings for the first three dimensions (Table 5) shows that the first dimensions (*Total Talk*) distinguishes individuals in terms of their overall amount of coded talk: all of the codes loaded in the same direction. The second dimension (*Tactical Talk*) distinguishes individuals who were engaged in passing information about tracks—Track Behavior and Detect/Identify loading most negatively—from those who were recommending action and describing the tactical situation—Recommendations and Status Updates loading most positively. Finally, the third dimension (*Command Behavior*) distinguishes individuals engaged in command behavior—Defensive Orders and Assessment/Prioritization loading most negatively—from individuals who recommended action and passed information—Track Behavior, Status Updates, Detect/Identify, and Recommendations loading positively.

Table 3
Code count comparisons—understanding the tactical situation.

Code	Control (n = 13)	Experimental (n = 16)	*U*	*p*	Cohen's *d*	Power
	Mdn	*Mdn*				
Assessment/Prioritization	18.0	6.5	72.5	0.17	0.53	0.27
Detect/Identify	13	12	105.5	0.96	0.02	0.05
Seeking Information	34.0	8.5	16.5	**0.00002***	2.0	0.99
Status Update	2	3	121.5	0.45	0.29	0.11
Track Behavior	17.0	16.5	102.5	0.96	0.02	0.05

Note. Results for two-sample Mann-Whitney *U* tests between commanders in the control and experimental conditions for codes related to understanding the tactical situation. $^{*}p < 0.05$.

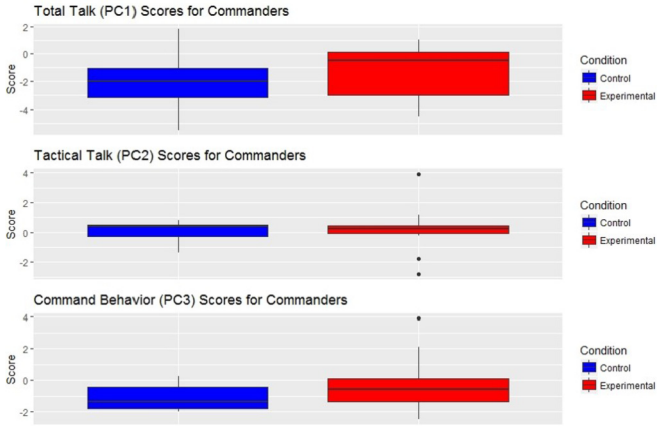

Fig. 3. PCA score distributions (box plots) for commanders by condition.

We also tested whether rotating the principal components and allowing them to correlate would yield a solution that was easier to interpret by (a) rotating the original PCA solution, and (b) using an exploratory factor analysis to extract and rotate three factors—the number suggested by the original PCA. Both approaches produced dimensions/factors with low correlations and similar interpretations to the original principal components. There were no significant differences between conditions on any of the rotated components/factors in any of the models. For the sake of brevity, we thus present only the specific outcomes of the PCA analysis.

Mann-Whitney *U* tests showed no significant differences between the PCA score distributions of commanders in the two conditions on any of the three dimensions (Fig. 3 and Table 6). These results suggest that commanders in the two conditions did not differ in their total talk, tactical talk, or command behavior as measured by PCA.

While PCA simultaneously accounts for multiple variables, we were not able to find meaningful differences between the contributions of commanders in the two conditions using this approach. Thus, a multivariate coding and counting approach was not sufficient for modeling the differences between commanders in the two conditions that we found qualitatively.

4.2.3. Epistemic Network Analysis

The first seven dimensions in the ENA model accounted for more variance in the data than any original variable. However, for the sake of clarity of interpretation, we only report results in terms of the first two dimensions, which account for more than 43% of the variance.

The network node locations (see Table 7) show that the first dimension (*Seeking Information/Tactical Decision Making)* distinguishes individuals in terms of the extent to which they connected Seeking Information versus tactical decisions with other relevant aspects of the discourse, that is, the other codes: Seeking Information is located on the negative side of the space, while Recommendations, Deterrent Orders, and Defensive Orders are located on the positive side. The second dimension (*Tactical Information/Action)* distinguishes individuals in terms of the extent to which they connect basic track information versus actions with other relevant aspects of the discourse: Track Behavior and Detect/Identify are located on the negative side of the space, while all of the other nodes are located on the positive side.

Table 6
PCA score comparisons.

Dimension	Control (n = 13)	Experimental (n = 16)	*U*	*p*	Cohen's *d*	Power
	Mdn	*Mdn*				
Total Talk (PC1)	−1.99	−0.50	129	0.29	0.42	0.18
Tactical Talk (PC2)	0.42	0.25	99	0.85	0.082	0.05
Command Behavior (PC3)	−1.35	−0.58	135	0.18	0.52	0.26

Note. Results for Mann-Whitney *U* tests between commanders in the control and experimental conditions on the three PCA dimensions. No significant results were found at the $p < 0.05$ level.

Table 7
ENA network node positions.

Code	ENA1	ENA2
Seeking Information	−2.28919	0.963926
Assessment/Prioritization	−0.11787	0.57612
Defensive Orders	0.202934	0.591342
Detect/Identify	0.425875	−1.55452
Recommendation	0.443784	1.006938
Track Behavior	0.457889	−0.86826
Status Update	1.042839	1.23255
Deterrent Orders	1.215247	1.331691

These interpretations mean that the projected points for commanders with relatively more connections to Seeking Information will be located toward the left-hand side of the space while those with relatively more connections to the tactical decision making codes will be located toward the right-hand side of the space. Similarly, the projected points for commanders with relatively more connections to Track Behavior and Detect/Identify will be located toward the bottom of the space, while those with relatively more connections among the other codes will be located toward the top of the space. Fig. 4 below shows the distribution of commanders in this space.

On the left-hand side of the space, we see almost exclusively projected points (circles) for commanders from the control condition (blue). On the right-hand side are only points from the commanders in the experimental condition (red). The mean location of commanders in the space are indicated by squares in the figure. The boxes around these squares are 95% confidence intervals around the means. We can also see that in the ENA model the TAOs from the two qualitative examples above are both representative of the commanders in their respective conditions (being close to the mean for each group), and also very distinct from one another, as the qualitative results showed. Finally, we see that one commander from the experimental condition, the red point farthest to the left, is an outlier: That commander's point is closer to the mean for the control condition than the mean for the experimental

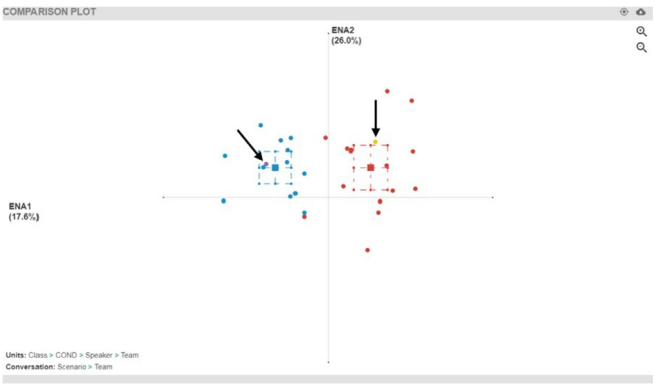

Fig. 4. ENA means and projected points for commanders in the control (blue, left) and experimental (red, right) conditions. Boxes are 95% confidence intervals on the first and second dimensions. The point in purple corresponds to the control condition TAO in the qualitative results described above. The point in yellow corresponds to the experimental condition TAO described above. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

condition on the first dimension.

The mean positions of commanders from the control and experimental conditions suggest that the two groups are different with respect to their positions on the first dimension, but not with respect to the second. To test whether these differences were statistically significant, we conducted a two-sample Mann-Whitney *U* test between distributions of the projected points in ENA space for commanders in the two conditions. We found a significant difference between the control (Mdn = −0.21, N = 13) and experimental (Mdn = 0.25, N = 16) point distributions on the first dimension ($U = 206$, $p < 0.01$, $d = 2.98$, power = 1.00). We found no significant difference between the control (Mdn = 0.25, N = 13) and experimental (Mdn = 0.19, N = 0.16) point distributions on the second dimension ($U = 106$, $p = 0.95$, d = 0.03, power = 0.05). These results suggest that commanders in the control condition made stronger connections to Seeking Information, while commanders in the experimental condition made stronger connections to codes related to tactical decision making.

To better understand these differences, we examined the mean networks for the commanders in both conditions.

The mean network for commanders in the control condition (blue, left, Fig. 5) shows that their strongest connections were from Seeking Information (indicated by thicker, more saturated lines) to Track Behavior and Detect/Identify, and less frequently to Assessment/Prioritization. In other words, the network graph shows the specific *kinds* of information that these commanders are seeking: consistent with the qualitative findings above, commanders in the control condition needed to seek information about tracks in order to *understand* the tactical situation.

The mean network for commanders in the experimental condition (red, right, Fig. 5), on the other hand, shows that their strongest connections were between Track Behavior and Detect/Identify: indicating, as we saw in the qualitative findings above, that they were *contributing* information about the tactical situation. Moreover, their strongest connections include connections among Deterrent Orders, Track Behavior, and Detect/Identify. In other words, the network graph shows that these commanders were making connections among three critical elements of the detect-to-engage sequence: consistent with the qualitative findings above, they were *using* information about the tactical situation to guide tactical decisions and actions.

To highlight the differences between commanders in the two conditions, we computed the difference between the mean networks for each condition. The resulting network is plotted in Fig. 6 below.

This mean network difference graph shows again that commanders in the control condition were *seeking information about the tactical situation*, and were doing so *more* than commanders in the experimental condition. It also shows that commanders in the experimental condition were *contributing* information about tracks and *linking information about the tactical situation to tactical actions*. And, that they were doing so *more* than commanders in the control condition.

Thus, the ENA network graphs provide both: (a) a visual

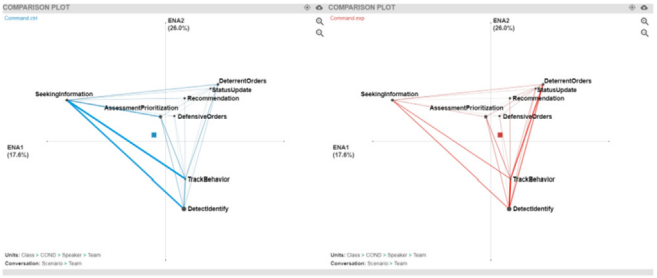

Fig. 5. Mean ENA networks for commanders in the control (blue, left) and experimental (red, right) conditions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

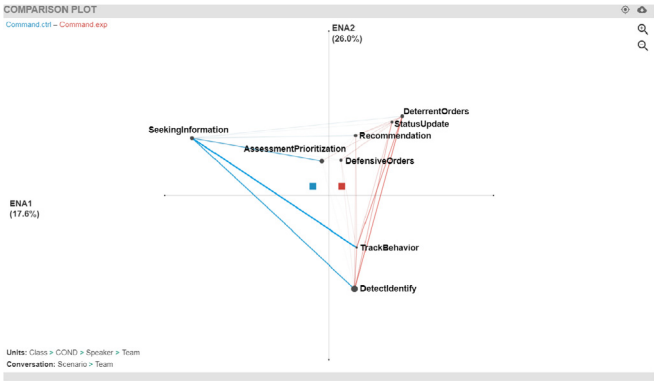

Fig. 6. ENA difference graph for commanders in the control (blue) and experimental (red) conditions.

representation of the key findings of the qualitative analysis above, and (b) a way to demonstrate that the key findings of the qualitative analysis are statistically significant between the two groups.

Moreover—and critically important for modeling the contributions of individuals in the context of the work of the team as a whole, we can also use ENA to examine the network graphs of individual commanders. For example, Fig. 7 below shows the network difference graph for the TAO in the control condition and the TAO in the experimental condition described in the qualitative analysis above.

This difference graph is quite similar to the difference graph for the mean networks of the two conditions. This suggests that, indeed, the differences observed in the qualitative analysis are a good representation of the overall condition differences.

Finally, ENA can identify individuals who are outliers in a group. Consider the network graph for one TAO from the experimental condition (the red point farthest to the left in Fig. 4 above) subtracted from the mean network for commanders in the experimental condition (See Fig. 8).

Here we can see that the performance of this TAO was more like the commanders in the control condition than the average commander in the experimental condition: This TAO makes stronger connections than average in the experimental condition to Seeking Information, and weaker connections than average between tactical information and tactical decisions and actions.

Altogether, these results suggest that ENA was able to model the individual contributions of commanders in the two conditions such that the models (a) aligned with the original qualitative analysis, (b) showed statistical differences between the two groups for findings from the original qualitative analysis, and (c) were able to distinguish differences

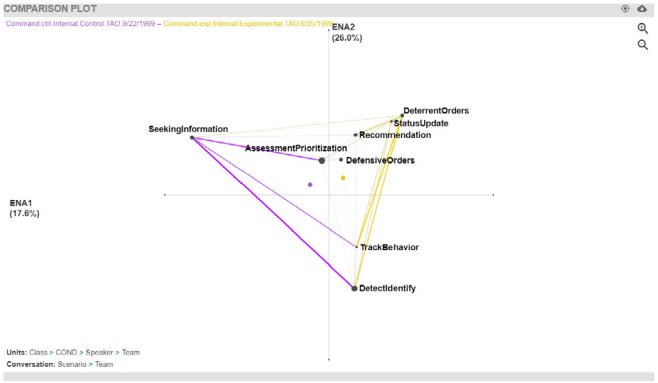

Fig. 7. ENA difference graph for the TAO in the control condition (purple, left) versus the TAO in the experimental condition (yellow, right) described in the qualitative results above. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 8. ENA difference graph for the outlier TAO in the experimental condition (red circle, green network, left) versus the mean network for commanders in the experimental condition (red square, red network, right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in individual performance.

5. Discussion

In this paper, we compared a qualitative analysis of CPS with three quantitative analyses: two using coding and counting and one using ENA. Our data was collected from ADW teams participating in simulated peacekeeping scenarios. Their CPS processes involved working together to detect, assess, and act on potential hostile vessels. While commanders on teams from the experimental condition were given teamwork training and access to a system designed to support their tactical decision making and reduce their cognitive load, commanders on teams from the control condition had access to standard ADW training and systems.

A qualitative analysis illustrated that the CPS processes of ADW commanders were highly interactive, interdependent, and temporal. However, this analysis also revealed the commanders in the two conditions reacted differently to the evolving tactical situations. Commanders in the control condition needed to seek information about potentially hostile aircraft to understand the tactical situation; commanders in the experimental condition were able to contribute information about the tactical situation for the team and use information about the tactical situation to guide tactical decisions and actions. These results suggested that the system used by commanders in the experimental condition successfully supported their decision making by providing them with important information about the evolving tactical situation, and that their training affected their interactions with the rest of the team.

Despite these rich qualitative findings, quantitative analysis based on coding and counting the different kinds of talk in the two conditions found only one statistically significant result: Commanders in the control condition sought more information than commanders in the experimental condition. In other words, a coding and counting approach did not model the nuanced differences revealed through qualitative analysis.

A PCA analysis, which accounted for the combinations of codes that individuals contributed, did not find *any* statistically significant differences between commanders in the two conditions. Thus, neither a univariate nor a multivariate coding and counting approach were able to model the differences we found qualitatively. An ENA analysis, on the other hand, was able to model the individual contributions of commanders in the two conditions such that the models (a) aligned with the original qualitative analysis, (b) showed statistical differences between the two groups for findings from the original qualitative analysis, and (c) were able to distinguish differences in individual performance.

Specifically, ENA was able to find a statistically significant difference between commanders in terms of the kinds of tactical information they sought to clarify. Moreover, the ENA analysis found that commanders in the experimental condition made statistically significantly more connections between different kinds of tactical information and between tactical information and tactical actions, compared to commanders in the control condition.

These findings from ENA are consistent with prior work by Morrison et al. (1996) who used a coding and counting approach to analyze communication data collected from a similar context. They found that ADW teams whose commanders had access to the decision support system (a) talked statistically significantly less about track information overall, (b) showed no difference in overall amount of actions ordered, (c) showed no difference in the overall amount of clarifications and (d) clarified less information about track behavior and identity in particular (not reported as statistically significant).

There are, however, some important differences between this previous work and the ENA analysis presented here. First, this previous work examined the effect of the decision support system on CPS processes at the *team* level. ENA was able to test claims about commanders at the *individual* level. An ENA analysis was able to model the unique contributions of commanders distinct from, but taking into account the actions of, the other members of the team. In addition, it was able to find statistically significant differences between commanders in the two conditions that corroborated qualitative findings, and identify the performance patterns of individual commanders and compare them.

The differences between results from coding and counting and results from ENA analysis can be explained by the way each method models team discourse. Univariate coding and counting assumes that team discourse can be modeled by looking individually at the frequency with which team members use—or talk about—concepts from the domain, without reference to other concepts used by an individual or to the concepts used by other team members. PCA assumes that team discourse can be modeled by finding latent constructs in the discourse that explain how individuals use concepts from the domain. However, PCA still works without reference to how other team members use concepts, and – more importantly – without reference to the temporal relationships between one concept and another in the data.

ENA, on the other hand, assumes that meaning –and therefore both understanding and action – is constructed through the local relationships among concepts. ENA models these relationships by accounting for the concepts used by an individual, while taking into account the concepts other team members are using within the recent temporal context. Thus, the critical difference between the methods is that the theory of discourse on which counting and coding is based, and the way it is operationalized in both univariate and multivariate techniques, fundamentally ignores *interactivity* (that a key property of the discourse is the relationships between actions), *interdependency* (the actions of one team member depends on the actions of others) and *temporality* (there is a span of activity in which actions depend on one another).

As we saw qualitatively, the individual contributions of commanders were in relation to multiple events. Commanders in the control condition sought information in relation to talk about track behavior, identification, and assessment; commanders in the experimental condition made decisions and took action in relation to similar information. However, these relationships had a particular structure that our univariate and PCA approaches ignored. Namely, the contributions of the commanders were in response to, and thus in some way dependent on, the recent contributions of the team. And as we saw, those contributions came from several sources: the supporting members of the team, agents external to the team, and the commanders themselves. Because ENA modeled connections between the contributions of individual commanders and the contributions of the rest of the team within the recent temporal context, the method was able to detect the qualitative differences we found in the two conditions.

While our results suggest that ENA has important advantages over

coding and counting, our findings have several limitations. First, we focused on CPS in this study because of its particular importance in educational and professional settings. There are certainly other socio-cognitive processes that are also interactive, interdependent, and temporal, and the results we present here might generalize to those settings. However, addressing the full range of socio-cognitive settings that involve interaction, interdependence, and temporality is beyond the scope of this paper.

Second, our power analyses suggest that the non-significant differences we found for coding and counting could be due to under-powered tests, and that larger sample sizes are needed to strengthen our claims. In this case, it was not possible to collect and analyze more data given that the original data collection was completed over 20 years ago by researchers with different analytic goals. More importantly, however, investigations of CPS phenomena using small sample sizes are commonplace in the field, particularly when qualitative methods are used. Thus, methods that can detect statistically significant differences with small samples, such as ENA, are valuable in the context of CPS assessment. In addition, had we had sample sizes large enough to detect significant differences with coding and counting, our results suggest that ENA also has an *interpretive* advantage over univariate and multivariate coding and counting approaches. Compared to univariate approaches, ENA (a) provides a model of CPS that accounts for key properties of CPS that are not modeled by coding and counting; and (b) provides a visual representation that summarizes multivariate relationships and eases the interpretation of complex dimensions. Finally, we note that while more data might allow coding and counting approaches to find statistical differences, the additional data could also provide further interpretive power to ENA. Thus, future work should include testing the generalizability of our claims using data with larger sample sizes.

Third, the goal of our study was to compare methods of measuring CPS processes, not to find novel differences between ADW commanders who participated in the experiment described above. Other studies have investigated the effect of training and the decision support system on team communication and team performance in some detail. Our study builds on this prior work by focusing on the communications and performance of individuals in the context of their teams, applying methods new to this data—namely, PCA, ENA, and a qualitative analysis—and by finding statistically significant results where others have not. In the context of CPS, however, the finding we want to emphasize is that our results suggest that ENA is a both a more valid and more effective approach to measuring CPS processes compared to coding and counting. That having been said, given the emphasis placed on assessment in ADW, the novel methods we have applied to this data, and significant results that build upon prior work with this data, we argue that our findings will also be of interest to the ADW community.

Fourth, while the supporting members of the team were included in our analyses, we only reported findings for individuals in the command roles. Our future work will explore how the teamwork training and presence of the decision support system for commanders affected the CPS processes of other ADW team members.

Fifth, as our qualitative analysis suggests, specific *sequences* of team interaction may be important. For example, it is often appropriate for warnings to come before defensive actions. The ENA model we used was *sensitive* to the order of team interactions because it modeled interactions within a given recent temporal context; however, it did not explicitly represent that ordering in the network models. Thus, important future work will be to investigate the additional value of applications of ENA that do represent order, as well as approaches such as Lag Sequential Analysis or Sequential Pattern Mining, which explicitly model sequences of action.

Finally, our results are of course limited to the particular air defense warfare teams examined in this study.

6. Conclusion

Our findings suggest several implications for CPS assessments. While coding and counting methods can be useful and easy to apply, they are limited when used as a sole method of assessment. Because these methods ignore critical features of CPS, they can lead to missed findings, missed opportunities for feedback, or invalid inferences about CPS performance.

Moreover, our results suggest that ENA can model the contributions of individuals to CPS while taking into account the recent contributions of other team members. The network models produced by ENA can also show the *particular kinds* of contributions individuals make and how they are related to the contributions of others. This means that ENA models could be used to give targeted and actionable feedback on *individual* performance during CPS activities. If automated codes are used in the ENA analysis (as they were here), these models could be integrated into real-time assessments that classify individuals based on their CPS performance.

Together, these results suggest that ENA is a powerful tool for CPS assessment, and that coding and counting approaches are limited in their ability to model important aspects of CPS in a team setting.

Acknowledgments

This work was funded in part by the National Science Foundation (DRL-1661036, DRL-1713110), the U.S. Army Research Laboratory (W911NF-18-2-0039), the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

References

Arastoopour, G., Swiecki, Z., Chesler, N. C., & Shaffer, D. W. (2015). Epistemic Network Analysis as a tool for engineering design assessment. *Presented at the American society for engineering education, Seattle, WA*.

Bagley, E. A., & Shaffer, D. W. (2015). Stop talking and type: Comparing virtual and face-to-face mentoring in an epistemic game. *Journal of Computer Assisted Learning, 26*(4), 369–393.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Bartholomew, D. J., Steele, F., Galbraith, J., & Moustaki, I. (2008). *Analysis of multivariate social science data.* Chapman and Hall/CRC.

Bowers, C. A., Braun, C. C., & Morgan, B. B. (1997). Team workload: Its meaning and measurement. In M. T. Brannick, E. Salas, & C. W. Prince (Eds.). *Team performance assessment and measurement: Theory, methods, and applications* (pp. 85–108). Psychology Press.

Cannon-Bowers, J. A., & Salas, E. E. (1998). *Making decisions under stress: Implications for individual and team training.* Washington, D.C.: American Psychological Association.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271–315.

Clark, H. H. (1996). *Using language.* Cambridge university press.

Dillenbourg, P. (1999). Collaborative learning: Cognitive and computational approaches. advances in learning and instruction series. *ERIC*. Retrieved from https://eric.ed.gov/?id=ED437928.

Eagan, B. R., Rogers, B., Pozen, R., Marquart, C., & Shaffer, D. W. (2016). *rhoR: Rho for inter rater reliability.* Retrieved from https://cran.r-project.org/web/packages/rhoR/index.html, 1.1.0.

Entin, E. E., & Serfaty, D. (1999). Adaptive team coordination. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 41*(2), 312–325. https://doi.org/10.1518/001872099779591196.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Foltz, P. W., & Martin, M. J. (2008). Automated communication analysis of teams. In E. Salas, G. F. Goodwin, & S. Burke (Eds.). *Team effectiveness in complex organizations and systems: Cross-disciplinary perspectives and approaches*. New York: Routledge.

Foltz, P. W., Martin, M. J., Abdelali, A., Rosenstein, M., & Oberbreckling, R. J. (2006). Automated team discourse modeling: Test of performance and generalization. *Proceedings of the 28th annual cognitive science conference* (Vancouver, CA).

Gašević, D., Joksimović, S., Eagan, B., & Shaffer, D. W. (2019). SENS: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior, 92*, 562–577.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for*

qualitative research. Aldine de Gruyter.
Griffin, P., & Care, E. (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Springer. Retrieved from http://link.springer.com/content/pdf/10.1007/978-94-017-9395-7.pdf.
Hall, J. K., Dwyer, D. J., Cannon-Bowers, J. A., Salas, E., & Volpe, C. E. (1993). Toward assessing team tactical decision making under stress: The development of a methodology for structuring team training scenarios. *Proceedings of the 15th annual intersective/industry training systems conference* (pp. 357–363). Washington, D.C.: American Defense Preparedness Association.
Halpin, P. F., & von Davier, A. A. (2017). Modeling collaboration using point process. *Innovative assessment of collaboration* (pp. 233–247). Springer.
Hao, J., Liu, L., von Davier, A. A., Kyllonen, P., & Kitchen, C. (2016). Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining. In T. Barnes, M. Chi, & M. Feng (Eds.). *Proceedings of the 9th international conferences on educational data mining* (pp. 382–387). (Raleigh, North Carolina, USA).
Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research, 18*(4), 323–342. https://doi.org/10.1016/S0740-8188(96)90003-1.
Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin, & E. Care (Eds.). *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Dordrecht, the Netherlands: Springer.
Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
Hutchins, S. G., & Kowalski, J. T. (1993). Tactical decision making under stress: Preliminary results and lessons learned. *Proceedings of the 10th annual conference on command and control decision aids*. Washington, DC: National Defense University.
Hutchins, S. G., Morrison, J. G., & Kelly, R. T. (1996). Principles for aiding complex military decision making. *Proceedings of the second international command and control research and technology symposium*. Monterey, CA: National Defense University.
Johnston, J. H., Fiore, S. M., Paris, C., & Smith, C. A. P. (2013). Application of cognitive load theory to develop a measure of team cognitive efficiency. *Military Psychology, 25*(3), 252–265. https://doi.org/10.1037/h0094967.
Johnston, J. H., Poirier, J., & Smith-Jentsch, K. A. (1998). Decision making under stress: Creating a research methodology. In J. A. Cannon-Bowers, & E. Salas (Eds.). *Making decisions under stress: Implications for individual and team training* (pp. 39–59). Washington, D.C.: American Psychological Association.
Johnston, J. H., Smith-Jentsch, K. A., & Cannon-Bowers, J. A. (1997). Performance measurement tools for enhancing team decision making. In M. T. Brannick, E. Salas, & C. Prince (Eds.). *Team performance assessment and measurement: Theory, method, and application*Hillsdale, NJ: Erlbaum 331–327.
Kapur, M. (2011). Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning, 6*(1), 39–56. https://doi.org/10.1007/s11412-011-9109-9.
Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7*(3), 77–124.
Kyllonen, P. C., Zhu, M., & von Davier, A. A. (2017). Introduction: Innovative assessment of collaboration. *Innovative assessment of collaboration* (pp. 1–18). Springer.
Maldano, R. M., Kay, J., Al-Qaraghuli, A., & Kharrufa, A. (2011). Analysing frequent sequential patters of collaborative learning activity around an interactive tabletop. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.). *Proceedings of the 4th international conference on educational data mining* (pp. 111–120). (Eindhoven, Netherlands).
Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *The Academy of Management Review, 26*(3), 356. https://doi.org/10.2307/259182.
Marquart, C. L., Hinojosa, C., Swiecki, Z., & Shaffer, D. W. (2018a). *Epistemic Network Analysis*. Retrieved from http://app.epistemicnetwork.org/login.html, 1.0.
Marquart, C. L., Swiecki, Z., Eagan, B., & Shaffer, D. W. (2018b). *ncodeR*. Retrieved from https://cran.r-project.org/web/packages/ncodeR/ncodeR.pdf, 0.1.2.
Mislevy, R. (1996). *Evidence and inference in educational assessment*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
Miyake, N., & Kirschner, P. A. (2014). The social and interactive dimensions of collaborative learning. In R. K. Sawyer (Ed.). *The Cambridge handbook of the learning sciences* (pp. 418–438). New York, NY: Cambridge University Press.
Morrison, J. G., Kelly, R. T., Moore, R. A., & Hutchins, S. G. (1996). *Tactical decision making under stress (TADMUS) decision support system, Vol. 13*.
OECD. (2017). PISA 2015 collaborative problem-solving framework. *PISA* (pp. 131–188). Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/chapter/9789264281820-8-en.
Paris, C., Johnston, J. H., & Reeves, D. (2000). A schema-based approach to measuring team decision making in a Navy combat information center. *The human in command* (pp. 263–278). Boston, MA: Springer. https://doi.org/10.1007/978-1-4615-4229-2_18.
Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaïane, O. R. (2009). Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 759–772.
Prokofieva, M. (2013). Evaluating types of students' interactions in a wiki-based collaborative learning project. *Australasian Journal of Educational Technology, 29*(4) https://doi.org/10.14742/ajet.239.
Reimann, P., Yacef, K., & Kay, J. (2011). Analyzing collaborative interactions with data mining methods for the benefit of learning. In S. Puntambekar, G. Erkens, & C. E. Hmelo-Silver (Eds.). *Analyzing interactions in CSCL* (pp. 161–185). Boston, MA: Springer.
Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.). *Computer supported collaborative learning* (pp. 69–97). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-85098-1_5.
Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal of Artificial Intelligence in Education, 25*, 380–406.
Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey, & E. Salas (Eds.). *Teams: Their training and performance* (pp. 3–29). Westport, CT, US: Ablex Publishing.
Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.
Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics, 3*(3), 9–45.
Siebert-Evenstone, A., Arastoopour Irgens, G., Collier, W., Swiecki, Z., Ruis, A. R., & Williamson Shaffer, D. (2017). In search of conversational grain size: Modelling semantic structure using moving stanza windows. *Journal of Learning Analytics, 4*(3), 123–139. https://doi.org/10.18608/jla.2017.43.7.
Smith, C. A. P., Johnston, J., & Paris, C. (2004). Decision support for air warfare: Detection of deceptive threats. *Group Decision and Negotiation, 13*(2), 129–148.
Srikant, R., & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology—EDBT, 1–17*.
Sullivan, S. A., Warner-Hillard, C., Eagan, B. R., Thompson, R., Ruis, A. R., Haines, K., ... Jung, H. S. (2018). Using epistemic network analysis to identify targets for educational interventions in trauma team communication. *Surgery, 163*(4), 938–943.
Suthers, D. D. (2006). Technology affordances for intersubjective meaning making: A research agenda for CSCL. *International Journal of Computer-Supported Collaborative Learning, 1*(3), 315–337.
Suthers, D. D., & Desiato, C. (2012). Exposing chat features through analysis of uptake between contributions. *System science (HICSS), 2012 45th Hawaii international conference* (pp. 3368–3377). IEEE.
Sweet, T. M. (2017). Modeling collaboration with social network models. *Innovative assessment of collaboration* (pp. 287–302). Springer.