Team Training for Enemy Identification Using an Intelligent Tutoring System

Kaitlyn M. Ouverson, Alec Ostrander, Anastacia MacAllister, Adam Kohl,
Jamiahus Walton, Stephen B. Gilbert, Michael C. Dorneich, Eliot Winer
Iowa State University
Ames, IA
{kmo, alecglen, anastac, adamkohl, anastac, alecglen, adamkohl, jwalton}
@iastate.edu

Anne M. Sinatra
STTC
Orlando, FL
anne.m.sinatra.civ
@mail.mil

ABSTRACT

Team training has been identified as critical to the operations of the Department of Defense (DoD) due to the complex and frequent interactions required in military teams. Effective training is necessary to develop complete understandings of the task and to build cooperative teams. Intelligent Team Tutoring Systems (ITTSs) have the potential to reduce training costs, improve learning, and increase feedback consistency in comparison to traditional human tutors. Currently, ITTSs are underdeveloped due to the state of the technology and the complex nature of intelligent agents, which require a variety of considerations and many hours to create. In this paper, the authors explore the impact of automated tutor feedback and team composition on performance for participants tasked with identifying and tracking enemy combatants.

Thirty-seven three-person teams, each composed of two spotters and one sniper, were tutored on their surveillance task performance over four trials. The scenario was constructed using Virtual Battle Space 2.0 (VBS2) and a version of the Generalized Intelligent Framework for Tutoring (GIFT), which assessed learners and delivered real-time feedback. In 18 teams, members received private, individualized feedback, while in 19 teams, members received individualized public feedback (i.e., their teammates could observe). Additionally, all teams experienced a change in team composition as the sniper and one of the surveillance spotters traded roles for the fourth trial. Each team's performance in the task was assessed. Evidence of training effectiveness is observed in participants' subjective performance and task workload. While feedback privacy was not found to influence the subjective performance, an effect was found for objective performance. These results about the effectiveness of feedback in team settings will influence the future study and development of ITTSs for the military by adding to the literature on how automated feedback should be designed within team training settings.

ABOUT THE AUTHORS

Kaitlyn Ouverson is a Ph.D. student at Iowa State University studying Human-Computer Interaction. Her research interests focus on computer-supported cooperative work, including decision-supporting systems, interaction design, and groupware. She is currently working on a project which focuses on evaluating future training technologies for teams, and another project which aims to assist decision makers in the adoption of research-backed community-health interventions.

Alec Ostrander is a Ph.D. student in Industrial Engineering and Human-Computer Interaction at Iowa State University. He is interested in how intelligent systems can be designed to leverage principles and ideas from the interaction design and teamwork literature to create effective human-agent teams.

Anastacia MacAllister, M.S., is a Ph.D. candidate in Mechanical Engineering and Human-Computer Interaction at Iowa State University's Virtual Reality Applications Center. She is working on developing Augmented Reality work instructions for complex assembly and intelligent team tutoring systems.

Adam Kohl, M.S., is a PhD student in Mechanical Engineering and Human-Computer Interaction at Iowa State University's Virtual Reality Applications Center. He is working on developing pattern recognition techniques to enhance n-dimensional data visualization methods.

Jamiahus Walton, M.S., is a Ph.D. student in the Human-Computer Interaction program at Iowa State University. His research concentrates on developing Intelligent Team Tutoring Systems (ITTSs) that can effectively and efficiently tutor a team. His dissertation focuses on improving the feedback mechanisms of ITTSs.

Stephen B. Gilbert, Ph.D., is an associate director of the Virtual Reality Applications Center and associate professor of Industrial and Manufacturing Systems Engineering at Iowa State University. His research interests focus on technology to advance cognition, including interface design, intelligent tutoring systems, and cognitive engineering. He is a member of IEEE and ACM and works closely with industry and federal agencies on research contracts. He is currently lead on a project supporting the U.S. Army Research Laboratory STTC in future training technologies for teams.

Michael C. Dorneich, Ph.D. is an associate professor of Industrial and Manufacturing Systems Engineering and a faculty affiliate of the human computer interaction (HCI) graduate program at Iowa State University. Dr. Dorneich's research interests focus on creating joint human-machine systems that enable people to be effective in the complex and often stressful environments found in aviation, robotic, learning, and space applications. Dr. Dorneich has over 20 years' experience developing adaptive systems which can provide assistance tailored to the user's current cognitive state, situation, and environment.

Eliot Winer, Ph.D., is an associate director of the Virtual Reality Applications Center (VRAC), associate professor of mechanical engineering, and a faculty affiliate of the human computer interaction (HCI) graduate program at Iowa State University. He has integrated four virtual and three live environments in a simultaneous capability demonstration for the Air Force Office of Scientific Research and has co-led the development of a next-generation mixed-reality virtual and constructive training environment for ARL HRED. Dr. Winer has over 15 years' experience with virtual reality, computer graphics, and simulation technologies.

Anne M. Sinatra, Ph.D., is an Adaptive Tutoring Scientist at the Natick Soldier Research, Development, and Engineering Center (NSRDEC), Simulation Training and Technology Center (STTC) in Orlando, FL. Her background is in Cognitive and Human Factors Psychology. She conducts adaptive training research as a member of the Learning in Intelligent Tutoring Environments (LITE) Lab. She works on the Generalized Intelligent Framework for Tutoring (GIFT) project and is the Team Tutoring research vector lead.

Team Training for Enemy Identification Using an Intelligent Tutoring System

Kaitlyn M. Ouverson, Alec Ostrander, Anastacia MacAllister, Adam Kohl, Jamiahus Walton, Stephen B. Gilbert, Michael C. Dorneich, Eliot Winer
Iowa State University
Ames, IA
{kmo, alecglen, anastac, adamkohl, anastac, alecglen, adamkohl, jwalton}@iastate.edu

Anne M. Sinatra
STTC
Orlando, FL
anne.m.sinatra.civ@mail.mil

INTRODUCTION

Team training has been identified as critical to the operations of the Department of Defense (DoD) due to the complex and frequent interactions required in military teams (Shuffler, Pavlas, & Salas, 2012). Not only must today's teams prepare to work with their like-minded teammates, they must also be equipped to work with partners outside their in-group, whether they are military or civilian (Salas, Cooke, & Rosen, 2008).

Understandably, the DoD identified virtual training as a priority in the 2017 budget proposal (Office of the Under Secretary of Defense (Comptroller), 2016). Intelligent Team Tutoring Systems (ITTSs), which train teams in virtual environments, are a promising answer to this call. ITTSs have the potential to reduce costs associated with team training without sacrificing quality and offer training possibilities for rare or extreme events that require simulation, feats already demonstrated by ITTS's individual training counterparts (i.e., Intelligent Tutoring Systems, ITSs; Brawner, Sinatra, & Sottilare, 2015; Shute, 1991). Another improvement over traditional human-led team tutoring comes from the tutor's ability to incorporate a balance of praise and criticism. Smith-Jentsch, Cannon-Bowers, Tannenbaum, and Salas (2008) note that human instructors have a tendency to focus on either positive or negative feedback, while an inclusion of both is most effective at shaping performance. Additionally, Team Dimensional Training (TDT) research shows that feedback allowing individuals and teams to identify mistakes translates into fewer overall errors and greater task recall (Smith-Jentsch, 2015; Smith-Jentsch et al., 2008). However, this previous work used after-action review, while the present study utilizes just-in-time feedback.

While ITTSs are sometimes harder to evaluate (Ososky et al., 2017; Salas, Rosen, Held, & Weissmuller, 2009), proven effectiveness of ITSs in many subjects and for various reasons (Brawner, Sinatra, & Sottilare, 2015; Shute, 1991) sets the stage for the software as an efficacious team tutoring alternative. Due to the current state of intelligent tutoring technology and the complex nature of intelligent agents (Gilbert et al., 2017; Ososky et al., 2017), ITTSs are underdeveloped. Additionally, the effect of team composition on training strategy has not been thoroughly researched. In this paper, the authors explore the impact of automated tutor feedback and team composition on performance for three participants tasked with identifying and tracking enemy combatants. This research will move ITTS design one step forward, providing guidance for future ITTSs in terms of feedback design.

BACKGROUND

Intelligent Team Tutoring Systems (ITTSs) offer improvements in feedback consistency and training costs over traditional human tutoring. While a human tutor is most effective when instructing a student one-on-one (Bloom, 1984; Cohen, Kulik, & Kulik, 1982), an ITTS can simultaneously tutor as many students as its server will support. Furthermore, an ITTS can retain a lesson for years and teach it in an identical fashion to the next generation of students, spanning temporal and geographic gaps between tutees. With the growing importance of teamwork as humans move toward more-complex problems requiring multidisciplinary teams, tutors that are specially calibrated to improve a group's communication and cohesiveness (among other skills and abilities) will prove vital for enhancing team performance.

The present research was conducted using the Surveillance with Sniper (SwS) task and tutor. The specifics of the task are detailed below. Essentially, three participants are assigned to either one of two spotter roles or the sniper role. In the spotter role, a participant must watch for people running through the virtual environment and signal to his teammates when one crosses the boundary he oversees. The sniper is responsible for assessing the threat level of the

people after they have crossed either spotter's boundary. The spotters signal zone transfers and identify when a person has crossed their boundary, while the sniper assesses whether that person is a civilian, an OPFOR (opposing force) with a gun, or an OPFOR wearing a vest containing an improvised explosive device (IED). While this scenario is more simplistic than a military surveillance task, it accommodates the evaluation of the SwS tutor, which, as an early iteration of an ITTS, will inform the evolution of three-person team tutoring systems.

There is no standardized method for measuring team performance; metrics for performance must be methodically chosen for each individual component of team performance. An individual can be evaluated solely on her performance and completion of a given task. However, teams must be evaluated for their member-specific task completion, overall team task completion, and interpersonal team skills (Gilbert et al., 2017), with the team skills posing the biggest challenge. The good news is that attempts to measure team skills are well documented, including the use of behavioral markers (Rosen et al., 2011; Salas, Rosen, Burke, Nicholson, & Howse, 2007) and a framework of critical considerations for team success (Salas, Shuffler, Thayer, Bedwell, & Lazzara, 2015) among many other qualitative and quantitative methods, typically utilizing trained observers (i.e., the BARS or BOS method) or communication analysis (DeShon, Kozlowski, Schmidt, Milner, & Wiechmann, 2004; Salas et al., 2009; Shuffler et al., 2012).

Of particular importance to this research endeavor are the critical considerations delineated by Salas et al. (2015). Specifically, *cognition*, or the shared understanding of team member and whole team goals and abilities, and *composition*, or the individual factors relevant to role configuration, are directly examined. The present study introduces a manipulation of cognition via feedback privacy (Salas et al., 2015). When feedback is private, each individual is privy only to his or her feedback; however, when feedback is publicly displayed to all team members, individuals become aware of the goals of team members whose goals are different from their own. Teams' mental model accuracy has been shown to be correlated with mental model similarity within Navy teams (Smith-Jentsch et al., 2008). Mental model accuracy, but not similarity, was shown to significantly and positively impact performance and teamwork process outcomes over those with inaccurate mental models. However, the lack of significance for similarity does not necessarily signal no relationship, as supported by the correlation between these variables. While accurate mental models should be most impacted by the training and the support of the tutor, the similarity of mental models within each team should point more directly to how well the team works together. An accurate model signifies agreement with a domain expert, but consistency in mental models within a team signifies a common approach and understanding of team goals. Since team cognition is reliant on the team's shared mental models of the task and team goals, publicly displayed feedback would, understandably, increase the similarities of those models.

In this study, team composition, another of the Salas et al. (2015) critical considerations, is indirectly manipulated through a fourth trial role switch: in the fourth trial, the participant playing the sniper and one of the spotter participants switch roles. The manipulation can be thought of as a manipulation of role naïveté, as teams and members begin with little-to-no knowledge of the task and gain understanding as they experience the task and the tutor. Presumably, in Trial Four, the team members who switch roles would regain some level of naïveté, having only interactions with their new role to reference. In addition, in this task which includes a feedback privacy/publicity manipulation, public feedback should positively influence team cognition. If team cognition is higher (presumably when there is public feedback), then those participants in new roles during the fourth trial will perform better than the spotters and sniper did in their first trial. This effect may be hidden by strong individual differences among team members, however. While individual differences in knowledge, skills, and attitudes are not comprehensively examined in this preliminary paper, individual skill at adaptability is tested in the fourth trial role switch.

Three Dimensions of Feedback

When designing appropriate feedback for an ITTS, several factors must be considered (Walton et al., 2018). For example, one must decide between using just-in-time feedback and after action reviews, both of which offer benefits and drawbacks (Gilbert et al., 2017), but one must also consider what actions or behaviors necessitate feedback. While many such aspects are common to any tutoring system, three considerations become important specifically in feedback for team-based training. The first is whether team members are assessed individually or as a team unit. Markers for performance and team behavior, as detailed by Salas, Rosen, Burke and colleagues (i.e., Rosen et al., 2011; Salas et al., 2007), can serve as triggers for feedback as well as markers for performance evaluations. The second is whether the feedback generated is addressed to an individual team member ("John, …") or to the team ("Alpha Team, …"). The third aspect is whether feedback should be presented privately to individuals or publicly to the whole team. The different combinations of these three design factors result in several possible team tutor feedback paradigms.

For some forms of training, evidence of a transfer effect has been observed, wherein subjects who are trained on one task show improvements in related, but non-trained tasks (Schubert, Strobach, & Karbach, 2014) and sometimes even show changes in more broad activities (Strobach & Schubert, 2001). As mentioned previously, public feedback is expected to increase team cognition, one of the key team constructs of Salas et al. (2015), specifically in relation to the goals of the two roles. Relatedly, the present study introduces partial-naïveté, which is enhanced through this public feedback. One assumes that a new Spotter in Trial 4, having watched a training video and interacted with his Sniper (and potentially her feedback, as in the public feedback condition) over the course of the first three trials, will have some idea of the goals and actions encapsulated by the Sniper role. When feedback is public, this knowledge will be further solidified, thereby enhancing team cognition (Salas et al., 2015). With varying levels of previous exposure and team cognition, there is potential for some participants to exhibit training transfer. Therefore, the researchers hypothesize that:

Hypothesis 1a. When feedback is public rather than private, subjective individual performance for new snipers in Trial 4 will be higher than the subjective individual performance of new snipers receiving private feedback.
Hypothesis 1b. When feedback is public rather than private, subjective individual performance for new spotters in Trial 4 will be higher than the subjective individual performance of new spotters receiving private feedback.

Similarly, a training transfer could be expected regardless of feedback condition, relying only on the existence of teammate interactions. Therefore, the researchers hypothesize that:

Hypothesis 2a. When feedback is public rather than private, subjective individual performance for the new sniper in Trial 4 will be higher than the subjective individual performance of old snipers in Trial 1.
Hypothesis 2b. When feedback is public rather than private, subjective individual performance for new spotters in Trial 4 will be higher than the subjective individual performance of old spotters in Trial 1.

While a certain amount of progress toward a desired outcome can be made without the presence of explicit feedback, properly calibrated feedback (i.e., feedback that orients the learner toward the actions needed to achieve a desired outcome and doesn't inhibit learning) expedites the process (Salomon & Globerson, 1989; Timperley & Hattie, 2007). However, feedback also necessitates considerations such as feedback content, delivery, and learner attention (DeShon et al., 2004; Geister, 2006; Gilbert et al., 2017). With regards to the two levels of performance, individual and team-level, different types of feedback have proven more successful. For example, public feedback has been shown to enhance team shared cognition and mutual trust, possibly because the feedback provides access to information that it is otherwise hidden, especially in a virtual team (Geister, 2006; Peñarroja, Orengo, Zornoza, Sánchez, & Ripoll, 2015). A related consequence of shared team cognition is backup behavior (Salas et al., 2015; Shuffler et al., 2012), which may be less influenced by individual-specific feedback. When individually-tailored, feedback has been shown to increase motivation in addition to task performance (Mumm & Mutlu, 2011). Initial work on this project showed that public feedback can have positive effects on the team's communication and perception of team competence. Therefore, the researchers hypothesize that:

Hypothesis 3. Objective team performance will be higher for teams that received public feedback than private feedback.
Hypothesis 4a. Subjective team performance will be higher for teams that received public feedback than private feedback.
Hypothesis 4b. Subjective individual performance will be higher for members of teams that received public feedback than private feedback.

Because teams are given three trials in their primary configuration, the researchers expected that participants would feel their individual and team performance improving as the trials progressed. However, subjective evaluations can be persuaded by individual differences in efficacy and reactions to feedback or criticism. A secondary measure of training is identified within the interpretation of the NASA taskload index (TLX), a well-documented measure of task workload (Hart & Staveland, 1988; Mohamed et al., 2014). As experience with a given task increases, it is expected that workload will decrease, signifying a mastery of some aspect of the task, thereby freeing up cognitive resources.

Hypothesis 5a. Subjective individual performance will improve for participants in Trials 1 through 3.
Hypothesis 5b. Subjective team performance will improve for participants in Trials 1 through 3.

Hypothesis 6. Workload will decrease as experience increases over Trials 1 through 3.

METHODS

Research Objectives

The researchers sought to evaluate the impact of an ITTS on teams of three. The training scenario was developed in Virtual Battlespace 2 (VBS2), a military virtual training environment, and utilized the Generalized Intelligent Framework for Tutoring (GIFT; Sottilare, Brawner, Goldberg, & Holden, 2012) to assess learners and present feedback.

Participants

Participants (*N*=111) were recruited using mailing lists at a large Midwestern University. Forty-five participants self-identified as female, 61 as male, and 5 persons chose not to identify. The average age of the recruited participants was just over 23 years of age, and 86.5% (*n*=96) were students from a variety of majors spanning all colleges at the university. Eighty-five percent (*n* = 94) of participants reported being native speakers of English.

Nearly every participant (89%, *n* = 99) reported working in teams at least once a month, and the majority (88%, *n* = 98) reported enjoying team work. Seventy-two participants (65%) reported playing videogames; just over half of those video games involving teams or cooperative play, on average (M=55.1%, SD =29.9%).

The participants completed the experiment in teams of three (*n* = 37). Teams were determined by experiment sign-up selection, which was mostly random, but did allow for groups of friends to participate together. As such, 36.2% of participants (*n* = 40) had met at least one person on their team before the experiment.

Task

When they arrived, participants were assigned to either of one of two Spotter roles or one Sniper role. Each role had individual goals that contributed to the overall team goal: tracking and determining the threat levels of potential OPFOR in the zone. Figure 1 shows the sequence of the task within the environment. The green circles represent the poles which served as reference points within the task. Gray rectangles are buildings and other structures, which were included to increase the complexity and realness of the task, since participants were required to act within a window of time to avoid losing a target behind a building and a surveillance team would be positioned at visual vantage-points. The Sniper in the team had to distinguish between three levels of target: (1) civilians, which pose no threat to the team, (2) low-level OPFOR, which carry guns and pose some threat, and (3) high-level OPFOR, which wear IED vests and may or may not carry guns. In Figure 1, these are represented by red triangles, rhombi, and pentagons, respectively.

Procedure

Participants gave informed consent and basic demographic information prior to signing up for a session. When they arrived at the lab, each participant completed a familiarity survey, which asked whether she knew her fellow participants. After watching a tutorial video, which introduced the task, the environment, and the controls for each role, participants began the first trial in their primary role. Each trial lasted for five minutes, and each experimental session consisted of four trials.

Participants received just-in-time feedback on their performance from an embedded ITTS. Feedback was supplied for each role's tasks and for both positive and negative performance with regards to a tolerance threshold, which served to reduce redundant feedback. Example feedback for each task is provided in Table 1. As discussed above, each team was randomly assigned to either a Spotter 1 – Sniper switch or a Spotter 2 – Sniper switch in the fourth trial to manipulate the team composition and the effect of prior role experience. Just before Trial 4, all players were given the chance to ask questions about their secondary role.

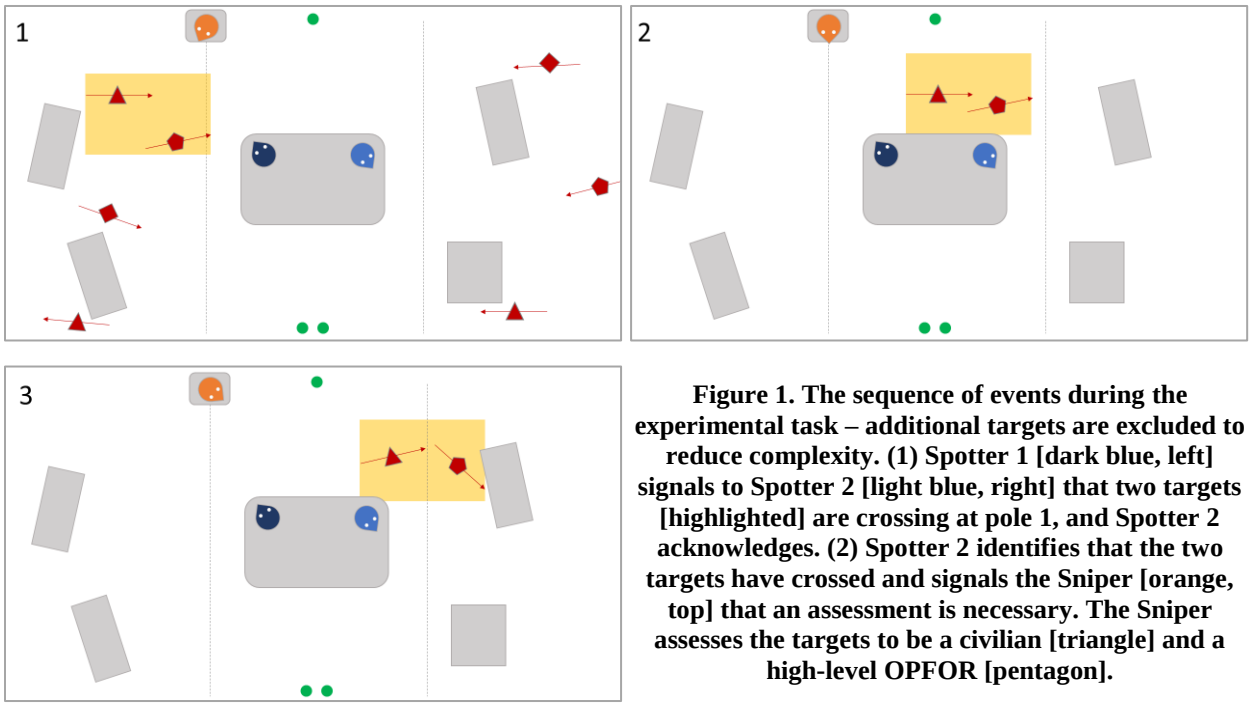

Figure 1. The sequence of events during the experimental task – additional targets are excluded to reduce complexity. (1) Spotter 1 [dark blue, left] signals to Spotter 2 [light blue, right] that two targets [highlighted] are crossing at pole 1, and Spotter 2 acknowledges. (2) Spotter 2 identifies that the two targets have crossed and signals the Sniper [orange, top] that an assessment is necessary. The Sniper assesses the targets to be a civilian [triangle] and a high-level OPFOR [pentagon].

After each trial, the participants were asked to complete a TLX survey and a post-trial survey. After the entire experimental session, the participants were asked to complete a post-experimental survey and participate with their teammates in an open forum discussion regarding the experimental environment and the feedback.

Table 1. A sample of feedback statements from the Surveillance with Sniper task.

Task	Praise Feedback	Constructive Feedback
Spotter Transfer	*"Successful handoff"*	*"Make sure you are not transferring too early"*
Spotter Acknowledge	*"Successful confirmation"*	*"Confirm transfer from teammate by pressing E key"*
Spotter Identify	*"Excellent work identifying targets"*	*"Scan your sector and identify all targets"*
Sniper Acknowledge	*"Successful confirmation"*	*"It is important to confirm at appropriate times"*
Sniper Assess	*"Excellent work assessing threat"*	*"It is important to assess threat in a timely manner"*

Independent Variables

Feedback Privacy

For each experimental session, feedback was either shown only to the person to whom it applied or publicly broadcast to everyone on the team. Imagine this as a coach individually conferencing each trainee on his or her performance (private feedback condition), or the coach telling everyone that an issue with someone's performance has surfaced, regardless of its relevance to the team (public feedback condition).

Trial

There were four trials per experimental session, but all longitudinal analyses were carried out on Trials 1 through 3. This is due to the change of the roles in the fourth trial, as the presence of secondary roles introduced confounding variables in the trial manipulation.

Role Naïveté

Role Naïveté is closely tied to the role-switch which occurs in Trial 4. In the first trial, all participants were new to their roles, having only watched a training video before jumping into the task (full-naïveté). In the second and third trials, the participants had gained some first-hand experience in their primary role (low-naïveté). In the fourth trial, either Spotter 1 or Spotter 2 (randomly assigned per experimental session) switched roles with the Sniper, thus returning to role naïveté, albeit with observations and the previous training video to guide them (partial-naïveté).

Dependent Variables and Metrics

This paper sought to understand the impact of feedback privacy, role naïveté, and trial on objective and subjective individual performance, objective and subjective team performance, pairwise performance for each role, and task workload. These dependent variables are operationalized in Table 2.

Table 2. The dependent variables and associated metrics used in this study.

Dependent Variable	Metric	Measure	Frequency
Objective Individual Performance	Event Reporting Tool (ERT) data	Count of missed transfers, missed acknowledges, missed identifications (IDs), and total errors	Every Trial (4x)
	ERT data	Transfer lead time and OPFOR-ID time in seconds	Every Trial (4x)
Objective Spotter Performance	ERT data	Percent successful transfer-acknowledge, and transfer-acknowledge-ID	Every Trial (4x)
	ERT data	Transfer-acknowledge time in seconds	Every Trial (4x)
Objective Pairwise Performance	ERT data	Percent ID-acknowledge and ID-assess	Every Trial (4x)
	ERT data	ID-acknowledge time in seconds	Every Trial (4x)
Objective Team Performance	ERT data	Percent transfer-acknowledge-ID-acknowledge-assess	Every Trial (4x)
Subjective Individual Performance	Post-Trial Survey	Score out of 7	Every Trial (4x)
Subjective Team Performance	Post-Trial Survey	Score out of 5, average of teammate ratings of performance	Every Trial (4x)
NASA TLX	Post-Trial Survey	Score out of 7	Every Trial (4x)

RESULTS

Each team's individual, pairwise, and team performance in the task were assessed, as well as the team members' task workload. As mentioned previously, researchers expected that:

(H1a) When feedback is public rather than private, subjective individual performance for new snipers in Trial 4 will be higher than the subjective performance of new snipers receiving private feedback, and (H1b) When feedback is public rather than private, subjective individual performance for new spotters in Trial 4 will be higher than the subjective performance of new spotters receiving private feedback;

(H2a) When feedback is public rather than private, subjective individual performance for new sniper in Trial 4 will be higher than the subjective performance of old snipers in Trial 1, and (H2b) When feedback is public rather than private, subjective individual performance for new spotters in Trial 4 will be higher than the subjective performance of old spotters in Trial 1;

(H3) Objective team performance will be higher for teams that received public feedback than private feedback;

(H4a) Subjective team performance will be higher for teams that received public feedback than private feedback, and (H4b) Subjective individual performance will be higher for members of teams that received public feedback than private feedback;

(H5a) Individual performance will improve for participants in Trials 1 through 3 and (H5b) Subjective team performance will improve for participants in Trials 1 through 3; and

(H6) Workload will decrease as experience increases over Trials 1 through 3.

Feedback Privacy's Effect on Perceived Role Performance

To evaluate Hypotheses 1a and 1b, a two-way analysis of variance (ANOVA) was run after Shapiro-Wilk's and Levene's tests revealed no evidence of non-normality or heterogeneity of variances, and an inspection of sample boxplots revealed no significant outliers. There was not a significant interaction between feedback privacy and role on subjective individual performance ($F(1,40) = 0.048$, $p = .83$), and the main effect of feedback privacy was also non-significant ($F(1,40) = 0.823$, $p = .37$). There was a marginally significant effect of role on subjective individual performance ($F(1,40) = 3.330$, $p = .08$, $d = -.598$), whereby the Spotters on average rated themselves as having performed significantly better (a lower score on the TLX question indicated higher performance, $M = 5.2$, SD = 9.3) than the Snipers ($M = 7.8$, SD = 5.9).

Role Naïveté's Effect on Perceived Role Performance

To evaluate Hypotheses 2a and 2b, a two-way mixed ANOVA was run for each hypothesis after Shapiro-Wilk's and Box's M tests revealed normal distribution in each cell and homogeneity of variance. Levene's Test revealed a violation of the assumption of homogeneity of variance for snipers ($p = .02$), which can decrease the power of the test if the groups are large (or the inverse if group sizes are small) and sample group sizes unequal. Because the group sizes are largely equal ($n = 19$ and $n = 18$) and the sample sufficiently large ($N = 74$, within subjects), the two-way mixed ANOVA was used despite the result of Levene's Test. All assumptions for the test were met for the spotter data. There was no significant interaction between feedback privacy and role naïveté on the difference between subjective individual performance for snipers ($F(1, 34) = 1.552$, $p = .22$, partial $\eta^2 = .044$). There was no significant interaction between feedback privacy and role naïveté on the difference between subjective individual performance for spotters ($F(1, 34) = .283$, $p = .598$, partial $\eta^2 = .008$).

Main effects of role naïveté and feedback privacy were investigated. For snipers, there was a significant difference in role naïveté ($F(1, 34) = 20.73$, $p < .001$, partial $\eta^2 = .379$) and no significant difference in feedback privacy ($F(1, 34) = .004$, $p = .950$, partial $\eta^2 = .000$). Given that the assumption of homogeneity of variance did not hold for the sniper data, nonparametric tests were performed to supplement and validate the above results. The Wilcoxon signed-rank test identified a significant effect of role naïveté for the sniper role ($z = -3.71$, $p < .001$, $Mdn_{t1} = 11$, $Mdn_{t4} = 5.5$). The Mann-Whitney U test did not identify a significant effect of feedback privacy in Trial 1 ($z = -1.05$, $p = .30$) or Trial 4 ($z = 0.72$, $p = .48$). For spotters, there was not a significant difference in role naïveté ($F(1, 34) = 1.73$, $p = .20$, partial $\eta^2 = .049$) and no significant difference in feedback privacy ($F(1, 34) = .341$, $p = .563$, partial $\eta^2 = .010$).

Feedback Privacy's Effect on Objective Team Performance

To evaluate Hypothesis 3, a two-way mixed ANOVA was run for each of three measures of team performance: 1) the percentage of teammate communications acknowledged (acknowledge percent), 2) the percentage of successful OPFOR hand-offs (transfer-acknowledge percent), and 3) the percentage of OPFOR that were adequately monitored (triple percent – named for the three actions the team must perform). Several assumptions of the mixed ANOVA were not able to be met for these measures. In each case, multiple transformations were attempted, but none were able to meet the necessary assumptions. The ANOVAs were completed, and nonparametric tests were performed to supplement and provide validation for the main effects of the analysis. The Friedman test was used to evaluate the repeated measures factor, role naïveté, while the Mann-Whitney U test was used to evaluate the between-subjects factor, feedback privacy.

Acknowledge Percent

There was no interaction between feedback privacy and role naïveté ($F(2,124) = .084$, $p = .919$, partial $\eta^2 = .001$). There was a significant difference in role naïveté ($F(2, 124) = 19.3$, $p < .001$, partial $\eta^2 = .238$), with post-hoc tests identifying Trial 1 as having significantly fewer acknowledged communications, on average, ($M = 19\%$, $Mdn = 18\%$, $SD = 14\%$) than both Trial 2 ($M = 33\%$, $Mdn = 29\%$, $SD = 25\%$, $p < .001$, $d = -.604$) and Trial 3 ($M = 34\%$, $Mdn = 30\%$, $SD = 24\%$, $p < .001$, $d = -.76$), but not between Trials 2 and 3 ($p = 1$). The result of the Friedman test matched the result of the mixed ANOVA ($p < .001$). The mixed ANOVA did not identify a significant difference between levels of feedback privacy ($F(1, 62) = 2.67$, $p = .11$, partial $\eta^2 = .041$); however, the Mann-Whitney U test did point to a significant difference in the median percentage ($U = 4183$, $z = -2.29$, $p = .022$, $r = .16$). The median percentage for teams with private feedback was 28% while for public feedback it was 21%.

Transfer-Acknowledge Percent

There was no interaction between feedback privacy and role naïveté ($F(2,128) = .499$, $p = .608$, partial $\eta^2 = .008$). There was a significant difference in role naïveté ($F(2, 128) = 14.015$, $p < .001$, partial $\eta^2 = .18$), with post-hoc tests identifying Trial 1 as having significantly fewer successful OPFOR hand-offs, on average, ($M = 10\%$, $Mdn = 10\%$, $SD = 10\%$) than both Trial 2 ($M = 19\%$, $Mdn = 11\%$, $SD = 21\%$, $p = .001$, $d = -.55$) and Trial 3 ($M = 20\%$, $Mdn = 15\%$, $SD = 19\%$, $p < .001$, $d = -.678$), but not between Trials 2 and 3 ($p = 1$). The result of the Friedman test matched the result of the mixed ANOVA ($p = .001$). The mixed ANOVA identified a significant difference between levels of feedback privacy ($F(1, 64) = 5.67$, $p = .02$, partial $\eta^2 = .081$). The result of the Mann-Whitney U test matched the result of the mixed ANOVA ($U = 4100$, $z = -2.83$, $p = .005$, $r = .197$). The median percentage for teams with private feedback was 15% while for public feedback it was 10%.

Triple Percent

There was no interaction between feedback privacy and role naïveté ($F(2,128) = 2.064$, $p = .131$, partial $\eta^2 = .031$). There was a significant difference in role naïveté ($F(2, 128) = 13.920$, $p < .001$, partial $\eta^2 = .179$), with post-hoc tests identifying Trial 1 as having significantly fewer adequately monitored OPFOR, on average, ($M = 5\%$, $Mdn = 0\%$, $SD = 7\%$) than both Trial 2 ($M = 12\%$, $Mdn = 0\%$, $SD = 18\%$, $p < .001$, $d = -.48$) and Trial 3 ($M = 11\%$, $Mdn = 5\%$, $SD = 14\%$, $p < .001$, $d = -.527$), but not between Trials 2 and 3 ($p = 1$). The result of the Friedman test matched the result of the mixed ANOVA ($p < .001$). The mixed ANOVA identified a marginally significant difference between levels of feedback privacy ($F(1, 64) = 3.34$, $p = .072$, partial $\eta^2 = .05$). The result of the Mann-Whitney U test matched the result of the mixed ANOVA, also yielding a marginally significant result ($U = 4510$, $z = -1.962$, $p = .05$, $r = .137$). The median percentage for teams with private feedback was 5% while for public feedback it was 0%.

Feedback Privacy's Effect on Subjective Performance

To evaluate Hypotheses 4a and 4b, a two-way mixed ANOVA was run for each hypothesis after Levene's, and Box's M tests revealed homogeneity of variance and covariance, and a visual inspection of sample Q-Q plots revealed a relatively normal distribution. There were no significant outliers, as assessed by examination of studentized residuals for values greater than ± 3. There was no significant interaction between feedback privacy and trial on subjective team performance ($F(2.20, 74.75) = .306$, $p = .76$, partial $\eta^2 = .009$). Additionally, there was no significant interaction between feedback privacy and trial on subjective individual performance ($F(2.43, 257.96) = .327$, $p = .76$, partial $\eta^2 = .003$).

Main effects of trial and feedback privacy were investigated, with feedback privacy main effects described here and trial main effects described in the next section. For subjective team performance, there was no significant difference in feedback privacy ($F(1, 34) = 1.22$, $p = .277$, partial $\eta^2 = .035$). For subjective individual performance, there was no significant difference in feedback privacy ($F(1, 106) = .206$, $p = .65$, partial $\eta^2 = .002$).

Learning Effects

Hypotheses 5a and 5b were evaluated using the same two-way mixed ANOVAs run for Hypotheses 4a and 4b after the assumptions were determined to be met. Main effects of trial on subjective performance were investigated. For subjective individual performance, there was a significant difference in trial ($F(2.43, 257.96) = 24.88$, $p < .001$, partial $\eta^2 = .190$). For subjective team performance, there was a significant difference in trial ($F(2.20, 74.75) = 12.49$, $p < .001$, partial $\eta^2 = .269$). Bonferroni-adjusted mean differences are presented in Table 3.

Table 3. Bonferroni-adjusted mean differences in subjective individual and team performance by trial

Subjective performance level	Trial 1 – Trial 2	Trial 2 – Trial 3	Trial 1 – Trial 3
Individual	2.63*	1.204*	3.83*
Team	-.253*	-.151	-.404*

* = The mean difference is significant at the .05 level
Scales were different for the evaluation of team and individual performance

To evaluate Hypothesis 6, a one-way repeated-measures ANOVA was run after Shapiro-Wilk's and Levene's tests revealed normal distribution in each cell and homogeneity of variance. An inspection of sample boxplots revealed no outliers in the data. The relationship between trial and task workload was statistically significant ($F(1.62,178.00) = 41.44$, $p < .001$, partial $\eta^2 = .274$).

Contrast effects were examined to determine the directions of the relationship. Trial 1 task workload ($M = 10.4$, $SD = 5.7$) was higher than that of Trial 2 ($M = 7.8$, $SD = 4.7$, $F(1, 110) = 32.38$, $p < .001$, partial $\eta^2 = .227$) and of Trial 3 ($M = 6.6$, $SD = 4.7$, $F(1, 110) = 59.81$, $p < .001$, partial $\eta^2 = .352$). Trial 2 task workload was also significantly higher than that of Trial 3 ($F(1, 110) = 14.92$, $p < .001$, partial $\eta^2 = .119$).

DISCUSSION AND CONCLUSION

Feedback privacy was shown to have limited (no significant) effect on subjective individual performance when participants were partially-naïve to their roles for both Spotters and Snipers. Therefore, hypotheses 1a and 1b (that secondary snipers/spotters who had received public feedback will rate their Trial 4 performance higher than those secondary snipers/spotters who had received private feedback) were not supported. Interestingly, there was a marginally significant effect of role on subjective performance, meaning that secondary Spotters rated themselves as having performed better than secondary Snipers. This likely points to a difference in workload demands between the sniper and spotter role.

Role naïveté and feedback privacy were shown to have a significant and marginally significant effect on subjective individual performance, respectively; however, no significant interaction between the variables was observed. This lack of a significant interaction could be due to the violated homogeneity of variance assumption, which lowers the power of the ANOVA test. Further analysis with greater power could reveal an interaction. Regardless, hypothesis 2a (that secondary snipers who had received public feedback will rate their Trial 4 performance higher than the primary snipers' Trial 1 performance, compared to those who had received private feedback) was partially supported. Hypothesis 2b (that secondary spotters who had received public feedback will rate their Trial 4 performance higher than the primary spotters' Trial 1 performance, compared to those who had received private feedback) was not supported. The discrepancy of statistical support could be due, in part, to the difference in subjective performance between secondary snipers and spotters revealed above. More work is needed to understand the finer details of this result.

Role naïveté and feedback privacy both demonstrated significant effects on the objective team performance of the team. Only data from Trials 1-3 was analyzed, as the manipulation introduced by the switch introduced additional sources of error that were best controlled via exclusion of the final trial. As would be expected, team performance improved across the three trials as participants gained experience with their role and team. However, the results of the analyses of feedback privacy on the various performance metrics indicated that teams that received private feedback performed more favorably as a team than those that received public feedback. Therefore, hypothesis 3 (that teams receiving public feedback would perform objectively better as a team) is not supported; in fact, there is evidence the opposite may be true.

Feedback privacy was not shown to influence subjective individual or subjective team performance; therefore, hypotheses 4a and 4b were not supported. Learning effects were observed in subjective team and subjective individual performance and in task workload. As teams gained experience together in the trial, they rated themselves as performing better individually and as a team, and their task workloads were lightened. Therefore, hypotheses 5a and 5b (that subjective individual/team performance will improve from trials 1 through three) and hypothesis 6 (that workload will decrease over trial) were all at least partially supported. Bonferroni-adjusted means and contrast analyses solidified the support for these hypotheses, with partial support uncovered only for hypothesis 5b. This dip

in workload and climb in subjective self- and team-ratings points to learning for all participants in the experiment, regardless of whether participants received public or private feedback.

Although the preliminary results of this study were unexpected, there is strong theoretical backing for why private feedback may be preferable in this scenario. As described above, three aspects of feedback design that are uniquely relevant for team contexts: the subject of the assessment (individual or team), the subject of the feedback (individual or team), and the feedback's level of privacy (public or private). A previous study (results in review) investigated the extremes of these dimensions, comparing feedback derived from a team assessment, addressed to the team and presented to the team – a "TTT" configuration – to feedback derived from each individual's assessment, addressed to that individual and presented only to that individual – an "III" configuration. The study showed that the team-feedback condition outperformed the individual-feedback condition in several aspects. Therefore, hypothesis 3 was developed on the simplistic conjecture that each of the 3 considerations contributed a part to the sum benefit seen in the results of that study. Instead, the results of these analyses, which compared an "III" configuration to a "IIT" configuration, suggest a more complex interaction between the feedback dimensions.

This result can also be understood from a more applied perspective. While it was beneficial for teams in the previous study to receive feedback based on the whole team's performance and addressed to the team, it is possible that feedback presented to the team without an explicitly named subject (e.g., "John…" or "Team…") could cause confusion or other detrimental effects. Finally, it is worth reiterating that the current study incorporated team members with different roles. Thus, team feedback that is primarily relevant to the spotter role is no longer directly relevant to all team members, and vice versa for the sniper role. It is possible that a higher occurrence of feedback statements deemed non-valuable conditioned participants to pay less attention to team feedback than to individual feedback. This is consistent with previous research, which identifies that too much feedback can negatively impact emotion valence and disrupt learning (Price, Mudrick, Taub, & Azevedo, 2018). In any case, more work is needed to understand the relationship between these three aspects of team feedback design and their effect on team behavior in different team tasks.

Limitations and Future Directions

Each team was assigned to receive either public feedback or private feedback (described above). However, the teams were not explicitly told the condition to which they were assigned. While the feedback was designed with the intention of cueing them into their condition, the lack of explicit direction about condition introduced excess error into the study. Regardless, the researchers are assuming that the manipulation was successful, since 93% (n = 103) of participants noticed the feedback during at least one of the four experimental trials.

Noticeably, this experiment does not feature a feedback-free control condition. A control condition was not included in an effort to decrease the number of required participants. Because a control condition was not included in this iteration of testing, the tutor effectiveness (i.e., whether the tutor improved performance better than regular practice within the scenario) cannot be fully accurately evaluated. However, pilot studies indicated that the presence of the ITTS positively impacted a number of performance metrics compared to when the task was undertaken with no feedback.

Further study is needed to explore other possible configurations of team-based feedback. One area of interest is a tutor that assesses and generates feedback addressed to individuals but presents these statements to the team. Prior work with human training has shown that such a feedback design can use team members to improve individual performance through peer monitoring (Loughry & Tosi, 2008). Another key area is the implementation and evaluation of mixed tutoring strategies. The tutor that could present team-based feedback to team members who share common goals, such as the two spotters, while presenting separate feedback to other members may receive the benefits that team feedback can bring while avoiding some of the pitfalls this early evaluation has identified. Finally, it should be noted that the scenarios developed for ITTS evaluation were designed to be simple, with shared roles intended to inherently promote the development of shared mental models. The results of this evaluation show the importance of developing shared mental models of the task at hand among team members, even if they have differing roles and knowledge. Doing so is an open and exciting next step for ITTS research.

ACKNOWLEDGMENTS

The research described herein has been sponsored by a collaborative agreement with the U.S. Army Research Laboratory (ARL). The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

REFERENCES

Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Brawner, K., Sinatra, A. M., & Sottilare, R. (2015). Architectural Research in Intelligent Tutoring Technologies. In A. Bruzzone, F. Longo, & R. Sottilare (Eds.), *Proceedings of the Fifth International Defense and Homeland Security Simulation Workshop* (pp. 48–55).

Carpenter, S., Fortune, J. L., Delugach, H. S., Etzkorn, L. H., Utley, D. R., Farrington, P. A., & Virani, S. (2008). Studying team shared mental models. *Proceedings of the 3rd International Conference on the Pragmatic Web: Innovating the Interactive Society*, 41–48. https://doi.org/10.1145/1479190.1479197

Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational Outcomes of Tutoring: A Meta-analysis of Findings. *American Educational Research Journal*, *19*(2), 237–248. https://doi.org/10.3102/00028312019002237

DeShon, R. P., Kozlowski, S. W. J., Schmidt, A. M., Milner, K. R., & Wiechmann, D. (2004). A Multiple-Goal, Multilevel Model of Feedback Effects on the Regulation of Individual and Team Performance. *Journal of Applied Psychology*, *89*(6), 1035–1056. https://doi.org/10.1037/0021-9010.89.6.1035

Geister, S. (2006). Effects of Process Feedback on Motivation, Satisfaction, and Performance in Virtual Teams. *Small Group Research*, *37*(5), 459–489. https://doi.org/10.1177/1046496406292337

Gilbert, S. B., Slavina, A., Dorneich, M. C., Sinatra, A. M., Bonner, D., Johnston, J., … Winer, E. (2017). Creating a Team Tutor Using GIFT. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-017-0151-2

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Loughry, M. L., & Tosi, H. L. (2008). Performance implications of peer monitoring. Organization Science, 19(6), 876–890. https://doi.org/10.1287/orsc.1080.0356

Mohamed, R., Raman, M., Anderson, J., McLaughlin, K., Rostom, A., & Coderre, S. (2014). Validation of the National Aeronautics and Space Administration Task Load Index as a tool to evaluate the learning curve for endoscopy training. *Canadian Journal of Gastroenterology & Hepatology*, *28*(3), 155–159. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24619638

Mumm, J., & Mutlu, B. (2011). Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior*, *27*(5), 1643–1650. https://doi.org/10.1016/j.chb.2011.02.002

Office of the Under Secretary of Defense (Comptroller). (2016). Defense Budget Overview: US DoD Fiscal Year 2017 Budget Request. Retrieved from http://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2017/FY2017_Budget_Request_Overview_Book.pdf

Ososky, S., Dorneich, M., Gilbert, S. B., Goldberg, B., Johnson, C. I., & Sinatra, A. M. (2017). The future of adaptive tutoring: Wrangling complexity across domains, applications, and platforms chair. *Proceedings of the Human Factors and Ergonomics Society*, *2017–Octob*, 1985–1989. https://doi.org/10.1177/1541931213601992

Peñarroja, V., Orengo, V., Zornoza, A., Sánchez, J., & Ripoll, P. (2015). How team feedback and team trust influence information processing and learning in virtual teams: A moderated mediation model. *Computers in Human Behavior*, *48*, 9–16. https://doi.org/10.1016/j.chb.2015.01.034

Price, M., Mudrick, N., Taub, M., & Azevedo, R. (2018). The Role of Negative Emotion Regulation on Self-Regulated Learning with MetaTutor. In 14th International Conference on Intelligent Tutoring Systems. Montreal.

Rosen, M. A., Bedwell, W. L., Wildman, J. L., Fritzsche, B. A., Salas, E., & Burke, C. S. (2011). Managing adaptive performance in teams: Guiding principles and behavioral markers for measurement. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2010.09.003

Salas, E., Cooke, N., & Rosen, M. (2008). On teams, teamwork, and team performance: discoveries and developments. *Human Factors*, *50*(3), 540–547. https://doi.org/10.1518/001872008X288457.

Salas, E., Rosen, M. A., Burke, C. S., Nicholson, D., & Howse, W. R. (2007). Markers for enhancing team cognition in complex environments: The power of team performance diagnosis. *Aviation Space and Environmental Medicine*, *78*(5), B77-85.

Salas, E., Rosen, M. A., Held, J. D., & Weissmuller, J. J. (2009). Performance Measurement in Simulation-Based Training. *Simulation & Gaming*, *40*(3), 328–376. https://doi.org/10.1177/1046878108326734

Salas, E., Shuffler, M. L., Thayer, A. L., Bedwell, W. L., & Lazzara, E. H. (2015). Understanding and Improving Teamwork in Organizations: A Scientifically Based Practical Guide. *Human Resource Management*, *54*(4), 599–622. https://doi.org/10.1002/hrm.21628

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research*, *13*(1), 89–99. https://doi.org/10.1016/0883-0355(89)90018-9

Schubert, T., Strobach, T., & Karbach, J. (2014). New directions in cognitive training: on methods, transfer, and application. *Psychological Research*, *78*(6), 749–755. https://doi.org/10.1007/s00426-014-0619-8

Shuffler, M. L., Pavlas, D., & Salas, E. (2012). Teams in the Military: A Review and Emerging Challenges. In *The Oxford Handbook of Military Psychology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399325.013.0106

Shute, V. J. (1991). Rose garden promises of intelligent tutoring systems: Blossom or thorn. In *Fourth Annual Workshop on Space Operations Applications and Research (SOAR 90)* (pp. 431–438). NASA, Lyndon B. Johnson Space Center. Retrieved from https://ntrs.nasa.gov/search.jsp?R=19910011382 2018-05-16T19:33:42+00:00Z

Smith-Jentsch, K. A. (2015). Team Self-Correction to Enhance Performance.

Smith-Jentsch, K. A., Cannon-Bowers, J. A., Tannenbaum, S. I., & Salas, E. (2008). Guided team self-correction: Impacts on team mental models, processes, and effectiveness. *Small Group Research*, *39*(3), 303–327. https://doi.org/10.1177/1046496408317794

Sottilare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). *The Generalized Intelligent Framework for Tutoring (GIFT)*. https://doi.org/10.13140/2.1.1629.6003

Strobach, T., & Schubert, T. (2001). Positive consequences of action-video game experience on human cognition: Potential benefits on a societal level, *49*(0), 1–6.

Timperley, H., & Hattie, J. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Walton, J., Ostrander, A., Ouverson, K., Gilbert, S., Dorneich, M., Winer, E., & Sinatra, A. (2018). Feedback Design Considerations for Intelligent Team Tutoring Systems. In *Human Factors and Ergonomics Society Annual Meeting*. Philadelphia, PA.