1-16-2019

Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: Surveillance

Alec Ostrander
Iowa State University, alecglen@iastate.edu

Desmond Bonner
Iowa State University, dbonner@iastate.edu

Jamiahus Walton
Iowa State University, jwalton@iastate.edu

Anna Slavina
Iowa State University, aslavina@iastate.edu

Kaitlyn M. Ouverson
Iowa State University, kmo@iastate.edu

Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: Surveillance

Abstract
This paper describes the development and evaluation of an Intelligent Team Tutoring System (ITTS) for pairs of learners working collaboratively to monitor an area. In the Surveillance Team Tutor (STT), learners performed a surveillance task in a virtual environment, communicating to track hostile moving soldiers. This collaborative problem solving task required significant communication to achieve the common goal of perfect surveillance. In a pilot evaluation, 16 two-person teams performed the task within one of three feedback conditions (Individual, Team, or None) across four trials each. The STT used a unique approach to filtering feedback so that teams in both individual and team conditions received a similar amount of feedback. In one performance measure, Team condition participants made fewer errors in one task than those in other conditions, though at a potential cost of mental workload. Feedback condition also significantly affected participants' subjective rating of both their own performance and their teammate's. This ITTS is one of the first automated team tutoring systems that provided real-time feedback during task execution. Recommendations are offered for the design of the optimal team task for future ITTSs that offer tutoring for small teams performing collaborative problem solving.

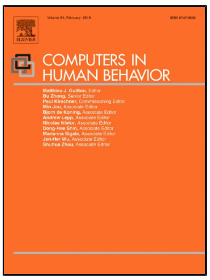
Keywords
intelligent tutoring systems, intelligent team tutoring systems, team training, small group dynamics

Disciplines
Industrial and Organizational Psychology | Industrial Engineering | Mechanical Engineering | Systems Engineering

Comments
This is a manuscript of an article published as Ostrander, Alec, Desmond Bonner, Jamiahus Walton, Anna Slavina, Kaitlyn Ouverson, Adam Kohl, Stephen Gilbert, Michael Dorneich, Anne Sinatra, and Eliot Winer. "Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: Surveillance." *Computers in Human Behavior* (2019). doi: 10.1016/j.chb.2019.01.006.

Authors
Alec Ostrander, Desmond Bonner, Jamiahus Walton, Anna Slavina, Kaitlyn M. Ouverson, Adam Kohl, Stephen Gilbert, Michael Dorneich, Anne Sinatra, and Eliot H. Winer

Accepted Manuscript



Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: Surveillance

Alec Ostrander, Desmond Bonner, Jamiahus Walton, Anna Slavina, Kaitlyn Ouverson, Adam Kohl, Stephen Gilbert, Michael Dorneich, Anne Sinatra, Eliot Winer

PII:	S0747-5632(19)30015-9
DOI:	10.1016/j.chb.2019.01.006
Reference:	CHB 5873
To appear in:	*Computers in Human Behavior*
Received Date:	10 May 2018
Accepted Date:	13 January 2019

Please cite this article as: Alec Ostrander, Desmond Bonner, Jamiahus Walton, Anna Slavina, Kaitlyn Ouverson, Adam Kohl, Stephen Gilbert, Michael Dorneich, Anne Sinatra, Eliot Winer, Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: Surveillance, *Computers in Human Behavior* (2019), doi: 10.1016/j.chb.2019.01.006

Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: Surveillance

Alec Ostrander[1], Desmond Bonner[1], Jamiahus Walton[1], Anna Slavina[1], Kaitlyn Ouverson[1], Adam Kohl[1], Stephen Gilbert[1], Michael Dorneich[1], Anne Sinatra[2], Eliot Winer[1]

[1]Iowa State University, [2]Army Research Laboratory

Author Note

This research is funded in part by a cooperative agreement from the U.S. Army Research Lab.

Correspondence concerning this article should be addressed to Stephen Gilbert, Iowa State University, 1620 Howe Hall, 537 Bissel Rd, Ames, IA 50011-2274. Contact: gilbert@iastate.edu

Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: The Surveillance Team Tutor

[Authors removed for blind review]

Abstract

This paper describes the development and evaluation of an Intelligent Team Tutoring System (ITTS) for pairs of learners working collaboratively to monitor an area. In the Surveillance Team Tutor (STT), learners performed a surveillance task in a virtual environment, communicating to track hostile moving soldiers. This collaborative problem solving task required significant communication to achieve the common goal of perfect surveillance. In a pilot evaluation, 16 two-person teams performed the task within one of three feedback conditions (Individual, Team, or None) across four trials each. The STT used a unique approach to filtering feedback so that teams in both individual and team conditions received a similar amount of feedback. In one performance measure, Team condition participants made fewer errors in one task than those in other conditions, though at a potential cost of mental workload. Feedback condition also significantly affected participants' subjective rating of both their own performance and their teammate's. This ITTS is one of the first automated team tutoring systems that provided real-time feedback during task execution. Recommendations are offered for the design of the optimal team task for future ITTSs that offer tutoring for small teams performing collaborative problem solving.

Keywords: intelligent tutoring systems, intelligent team tutoring systems, team training, small group dynamics

1 Introduction

Although the need for collaborative problem solving in modern work and social systems is pervasive, it is not always realized effectively. Recent research has sought to develop automated assessment methods to support training on collaborative problem solving (Flor, Yoon, Hao, Liu, & von Davier, 2016; Scoular, Care, & Hesse, 2017). Graesser et al. (2017) have suggested that Intelligent Tutoring Systems (ITSs) could provide further direction in this area to grant new perspectives on team assessment. ITSs have been quite successful in the instruction of individuals across a variety of domains. For example, VanLehn (2011) found that the effect of computer tutoring was nearly identical to human tutoring, while Kulik & Fletcher (2016) described a median increase in test scores from the 50th to 75th percentile when instructional objectives were aligned with tests. However, there are branches of training for which intelligent tutoring is still in its nascent state. One such area is training teams.

Several new challenges arise when attempting to tutor a collaborative problem solving team due to the complex and dynamic interactions among multiple users working together. Whereas an ITS for an individual need only perceive and respond to the user's direct input to the system, a successful Intelligent Team Tutoring System (ITTS) must interact with all users and monitor interactions between users. There are many different compositions of teams, team tasks, backgrounds of team members, and the approaches that are used to assess performance and provide feedback will be highly dependent on the team, the task, and the problem domain (Graesser et al., 2018). It is necessary to explicitly determine the different tasks that team members will be performing, how the team is structured, and how the system will communicate information to the team. In collaborative problem solving, team members tend to all have the same goal – to solve a problem together or to collaboratively complete a task (Fiore et al., 2017; Hao, Liu, von Davier, & Kyllonen, 2017). Whether the domain is solving a puzzle, engaging collaboratively in monitoring an area for threats, or anything in between, people on the team will need to be able to communicate with each other, and to be aware of the actions that their teammates have taken. The computer-based system with which the team is engaging must allow information to be passed back and forth between players and for them to perceive a shared state of the task.

Along with the complexity of assessing the cognitive and psychomotor aspects of a team's taskwork, an ITTS must also evaluate the team's teamwork, or collaborative skills (Cannon-Bowers, Tannenbaum, Salas, & Volpe, 1995). Sottilare et al. (2017) explored methods of mapping specific behavioral markers to teamwork constructs. Methods of operationalizing those principles for an ITTS are generally described by Bonner et al. (2017) and the next section describes the application of those principles for the current task. Using computer-based training for these complex team tasks is essentially required in order to facilitate multi-faceted assessment of both taskwork and teamwork.

While collaborative problem solving occurs in homes, the workplace, and most domains of our lives (Fiore et al., 2017), a military-style surveillance task was chosen for this study based on the interests of the sponsoring research partner. The ITTS developed in this project, called the Surveillance Team Tutor (STT), was built using the Generalized Intelligent Framework for Tutoring (GIFT). GIFT is an open-source, domain-independent framework of computational tools for creating ITSs (Sottilare, Brawner, Goldberg, & Holden, 2012). The authoring tools within GIFT are designed to allow authors to create adaptive computer-based tutors in their topic area of expertise. GIFT's real-time assessment features and its domain-independence made it an appropriate platform for this collaborative problem-solving task. As part of the current work, the

team implemented novel technologies in GIFT so that it could tutor multiple individuals simultaneously and track their actions in a shared virtual environment. While previous publications have described the technical details of implementing this two-person surveillance ITTS (Authors, 2015, 2016, 2017c), this paper focuses on a pilot evaluation of the ITTS's instructional effectiveness and how the evaluation results can guide the design of more tutorable team tasks for collaborative problem solving.

2 Related Work

The STT offers guidance to two teammates participating in what is essentially a multiplayer video game-like surveillance task that requires significant communication and coordination under time-pressure. To understand the design choices made for the Surveillance Tutor and the evaluation approach, it is useful to briefly describe previous research related to ITTSs, team assessment, and feedback design.

2.1 Intelligent Team Tutoring Systems

When discussing ITTSs, it is important to consider several characteristics of the system: 1) the team composition (e.g., roles and background training of team members), 2) the role of the tutor (e.g., supervisory vs. a team member performing the tasks alongside humans), 3) unit of assessment (team, individual, or both), and 4) the type of feedback given, if any. These are just four of several characteristics that could be used to characterize a team tutor (Bonner et al., 2015; Fiore et al., 2017).

One of the first ITTS-like systems was the Advanced Embedded Training System (AETS) (Zachary et al., 1998). AETS was developed to facilitate air defense training in the Navy. It monitored the learners, keeping track of button presses, speech and eye movements so the human trainer had more time to assess and give feedback However, while individual team members received automated feedback both during and after their tasks, AETS did not provide any type of automated feedback at the team level; that feedback was usually given by instructors who collaborated with the AETS tutor by accumulating information for an after-action review.

In the Team Multiple Errands Task (TMET) (Walton, Bonner, et al., 2015; Walton, Gilbert, Winer, Dorneich, & Bonner, 2015), the tutor also played a supervisory role, giving real-time feedback to three people attempting to complete a task in a multiplayer virtual shopping mall in which the team must deduce the most efficient way to purchase a list of items under a variety of constraints designed to increase cognitive load (time limit, enter a store only once, etc.). In TMET, the tutor was not embodied or personified; the feedback was given as brief phrases based on individual's performance, though team assessment was accomplished via a team score on the screen. The teams of three were homogeneous, with no specific job roles or background training.

Avis, a socially capable mechanical engineering tutor (Kumar, Ai, Beuth, & Rosé, 2010), was able to give feedback to a team through conversational dialogue, acting as a guide for learning underlying concepts. While it could be considered an ITTS, Avis did not assess and provide feedback for the team as a whole; it attended to each learner individually and their statements about concepts. The tutor in this team setting was framed as an authority but less of a supervisor and more of a facilitator. The team here was also homogeneous, i.e., engineering undergraduates.

2.2 Team Assessment

Given the variations of teams as described above, it can be difficult to discern how best to measure teamwork and taskwork. Even when a measure is defined, it is difficult to generalize the

result to other categories of teams. In the team skill assessment systems developed for evaluating collaborative problem solving within the Programme for International Student Assessment (PISA) 2015 (Organisation for Economic Cooperation and Development, 2017), a simulated human agent engaged a learner as a conversational peer. Together, the student and agent collaborated to solve a problem. In some problems, the student collaborated with several tutoring agents with different skills. Students and tutoring agents could have different task roles, but always had the same social status, to remove cross-cultural differences in willingness to engage higher status colleagues. The PISA 2015 also varied dimensions such as whether all team members have the same or different information available, and the interdependency of the tasks (Organisation for Economic Cooperation and Development, 2017). Tasks all involved a single human team member, however, since it was designed to assess that person's collaborative problem solving skills.

While the PISA agents are not tutors, these examples hint at the large range of possible team assessment approaches. The STT built on this previous work, using a two-person independent task, with a non-embodied supervisory tutor that gives directive feedback in real-time, but with assessments at both the individual and team level.

The PISA 2015 systems are examples from a larger area of research on computer-supported collaborative learning (CSCL) using Intelligent Collaborative Learning Systems (ICLSs). These tools have often been used to support students who collaborate to solve a problem together, encourage each other to explore ideas, defend arguments, and reflect on their process. The STT, on the other hand, much like the AETS and TMET tasks, is designed for a fast-paced high-cognitive load psychomotor performance task, one of collaborative task types noted in McGrath's group task circumplex (1984). In these tasks, conversational dialogue beyond required task communications does not typically occur. Thus, while the STT requires an assessment of communication and coordination, tools that are valuable for analysis of collaborative learning communication, such as Soller's Collaborative Learning Conversational Skills Taxonomy (2001), do not apply as well.

The STT task design leverages the teamwork constructs proposed by Eduardo Salas, Shawn Burke, and colleagues, i.e., the "big five" teamwork components (2005) and the nine C's of teamwork (Salas, Shuffler, Thayer, Bedwell, & Lazzara, 2015). This big five of teamwork were inspired in name by the big five personality trait framework (Digman, 1990) and developed based on a meta-analysis of teamwork research. The five core constructs that support team effectiveness were: team leadership, mutual performance monitoring, backup behavior (doing a teammate's job when needed), adaptability, and team orientation. In this model of teamwork, these five constructs were coordinated via shared mental models, mutual trust, and closed loop communication. During the development of the STT, Salas' research team updated their model of teamwork to be described by nine C's: cooperation, conflict, coordination, communicating, coaching, cognition, composition, context, and culture. While both models guided the design of STT's team assessment, most directly helpful were the descriptions of behavioral markers that could serve as proxies for measuring actual constructs such as coordination or backup behavior. In a further meta-analysis, Sottilare and colleagues documented behavioral markers for a variety of the nine C's (2017). In the current experiment, an effort was made to operationalize a relevant subset of this thorough list of markers as a means of team assessment of communication in the STT, and this effort is described previously in more detail (Authors, 2017b).

2.3 Feedback Design in Team Tutors

In non-conversational ITTSs, communications with the tutor can be considered to be directed feedback, typically provided to the learner via text, audio, and/or visual indicators on screen. There

are many different characteristics to consider in the design of feedback in terms of timing, positive or negative tone, and content (Narciss & Huth, 2004), but there are at least three variables that become especially important in the context of a team task. The first is whether the target of the feedback is the individual or the team (DeShon, Kozlowski, Schmidt, Milner, & Wiechmann, 2004). Feedback focused on the individual might begin with the individual's name, e.g., "Maria, please remember to…", while feedback focused on the team might begin, "Team, everyone needs to…"

A related characteristic for this research study is whether the feedback is public or private, i.e., distributed to all team members vs. just to the individual(s) who need to hear it. While previous research suggests that public feedback can be more effective than private (Gabelica, Van den Bossche, Segers, & Gijselaers, 2012), more research is needed that explores the contexts in which this more public, potentially shaming feedback is appropriate (e.g., is positive feedback more appropriate in public?), as well as the interaction of public vs. private with the focus on team vs. individual.

Finally, a subtler variable in ITTS feedback design is whether the feedback was generated by a team assessment or an individual assessment. While both sources of assessment might lead to similar feedback verbiage, recipients may feel differently about feedback generated by their individual actions vs. the team's performance overall, assuming that difference is transparent to the team member. Also, team assessments can offer feedback that's more specific to the relationships among team members.

To illustrate these three variables with an example, imagine a four-person team with members Alicia, Bob, Carlos, and Daya. An ITTS (or human coach) might assess individual performances and decide that three of the four team members need to work on communication. These individual assessments could lead to public team feedback ("Team, please work on your communication"), public individual feedback ("Alicia, Carlos, Daya, please work on your communication"), or private individual feedback to each relevant member ("Alicia, please work on your communication", "Carlos, please…" and "Daya, please…"). If the coach instead used a team assessment technique, perhaps analyzing the communications network created by the four members, the feedback can become more relational. Some examples include public team feedback ("Team, please give everyone a chance to talk"), public individual feedback ("Bob, Daya, you're talking more Alicia and Carlos; please give them a chance to talk"), or private individual feedback just to relevant members ("Bob, you're…" and "Daya, you're…").

A further important consideration for team tutor feedback is frequency. If an ITTS is constructed to give feedback whenever its individual and team assessment algorithms notice behavioral markers that merit feedback, it would be easy to quickly overwhelm the team members with too much feedback. As noted by Sottilare et al. (2011), a pedagogical module is needed within a team tutor to filter that feedback, gauging which statements should be prioritized. This issue arose in the STT.

Another critical factor in feedback design is whether it is actually helpful for improving task performance – feedback utility. While this requirement seems obvious, the design of optimally helpful feedback is often non-obvious. Feedback strategies range from simply indicating that something is incorrect to telling learners exactly what is incorrect, to telling them what to do to correct the situation. VanLehn (2006) noted the pedagogical importance of reminding students of their overall goal before revealing a specific action to take. Corbett & Anderson (2001) described different locus of control scenarios and found faster learning when the feedback was given to students immediately vs. upon request by a student. Walton et al. (Walton, Gilbert, et al., 2015)

highlighted the challenge of having a particularly engaging task and the corresponding difficulty of having learners not attend to feedback because it was not visually salient enough to notice, given the task workload. In the STT study, we focused on the following elements of feedback utility: for each task, 1) whether the feedback has the potential to improve task performance, and 2) whether the mental workload of the task allows the feedback to be received. If the learners' performance improves more when feedback is present than when it is not, both of these questions will be answered affirmatively.

3 The Surveillance Team Tutor

3.1 Team Tutor Architecture

The STT was built using a game engine called VBS2 for the learner environment as well as the GIFT framework, a modular framework for computer-based tutoring (Sottilare, Baker, Graesser, & Lester, 2018; Sottilare et al., 2011). GIFT was originally designed as modular framework to support intelligent tutoring systems for individuals, and for this project it was customized to support team tutoring. It is worth noting that for the two-player Surveillance Task, the STT contains a learner model for Player 1, a learner model for Player 2, and a learner model for the Team. Thus, individual player skills and team skills could be tracked separately. Also, the expert module, in charge of assessment and choosing feedback, contained assessment rules for both individual players and the team.

3.2 The Surveillance Task and Its Subtasks

A collaborative task was developed to serve as a testbed on which to study aspects of team performance. To make this exercise generally useful, the task needed to be scalable in terms of team size and difficulty, agnostic of tutoring approach, and based on communication between team members. A collaborative surveillance task was developed that required teams to engage in a collaborative scenario in which each team member held the same role and goals. A key technical challenge in developing a team tutor is the need to monitor not just the interactions between human and tutor, but also the interaction between team members. To simplify this challenge, the task was designed such that all relevant team interaction could be structured into three types of communication events. These events are termed Transfer, Acknowledge, and Identify throughout this paper.

The training exercise is based on a military surveillance scenario developed in Virtual Battlespace 2, a game engine. The team is stationed on top of a building in the virtual environment, and members are responsible for monitoring their respective 180-degree zones (Figure 1) and letting their teammate know if any OPFOR (Opposing Forces, or enemy characters) head toward the other member’s zone. If this happens, the first member must alert that other member of the transition. This is the *Transfer* event. To record this action for the tutoring system, a member types "1" or "2" to indicate which border is being crossed while saying a phrase like "Transfer at the 2-pole!" The second member should then acknowledge to the first member that the message was heard. This is the *Acknowledge* event, and the member would say, "Acknowledged" and type the E key. Once the OPFOR crosses into the new zone, the receiving team member should *Identify* the OPFOR by saying "Identified!" and typing the spacebar. These keystrokes were chosen to be easily typed with the left hand while the player's right hand used the mouse to scan back and forth across the 180-degree surveillance zone. The exercise supports a team of two participants in its basic configuration, and it can be scaled up to more team members, each with smaller individual zones of responsibility.

Feedback was provided as text that appeared in a panel left of the main surveillance screen. New feedback statements were highlighted briefly in blinking yellow highlight to draw attention to them. Eye tracking results demonstrated that learners were able to notice the appearance of feedback. More details about the creation of the task itself are described elsewhere (Authors, 2017c).

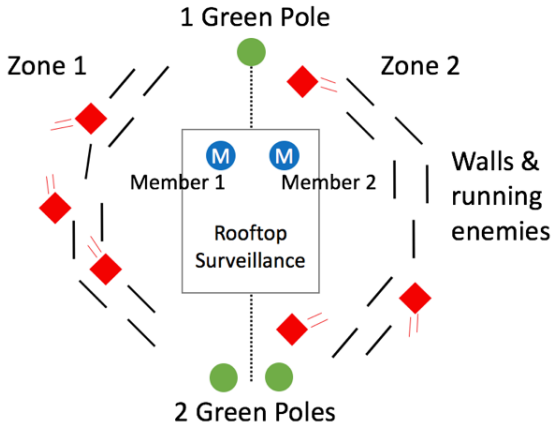


Figure 1. Overhead view of the Surveillance Task environment. Two sets of green poles serve to identify the zone borders, and enemies (red diamonds) move between zones.

This task could continue as long as our environment provided hidden OPFOR who appeared and then ran. In our study, each trial lasted five minutes long, and in that time 40 OPFOR altogether were programmed to appear from behind a wall and ran. Their start and end points were specified, but their exact running path was not. These began with one OPFOR running at a time, leading to successful exchanges such as:

- M1: (When seeing an OPFOR heading towards the 1-pole zone border) "Transfer at the 1 pole!"
- M2: "Acknowledged!"
- M2: (After seeing the OPFOR enter M2's zone): "Identified"

As time passed, more and more OPFOR began to run, sometimes with several crossing the border simultaneously in both directions, leading to more frantic exchanges such as:

- M1: "3 transferring at the 2 pole."
- M2: "Two transfers at your 1, and 1 at your 2."
- M1: "Acknowledge! Identify, Identify.
- M2: "Identify…. Identify."
- M1: "Identify!"

In this dialogue, M2 failed to acknowledge the first communication from M1, and M2 missed an Identify.

Table 1. Feedback for the Surveillance Task. In the Team condition, teams received only Acknowledge and Identify feedback to balance the overall amount of feedback between conditions.

Question	Assessment	Feedback Statement
Transfers Present?	Below (statement randomly chosen)	It is important to communicate crossings
		Report transferring OPFOR to team by pressing 1 or 2 key
		It is important to communicate crossings. Your communication needs work
Transfers Present?	At, Above	[no feedback]
Transfer Timing?	Above	Successful handoff
Transfer Timing?	At	Make sure you are not transferring too early
Transfer Timing?	Below	It is important to communicate when an OPFOR crosses into your partner's zone
Ack. Time?	Above	Successful confirmation
Ack. Time?	At	Acknowledge your communications as soon as you receive them
Ack. Time?	Below	It is important to confirm at appropriate times. *After repeated Below assessments:* Remember to acknowledge your teammate's communications.
ID Time?	Above	Excellent work identifying OPFOR
ID Time?	At	It's important to identify OPFOR as quickly as possible
ID Time?	Below	Identify OPFOR immediately
Condition	**Feedback Statements Preceded by "P1," "P2," or "Team"**	
Individual	[P1]: Remember to acknowledge your teammate's communications *(P1 = Player 1, P2 = Player 2)*	
Team	Team, remember to acknowledge your teammate's communications	

3.3 Surveillance Task Feedback

As noted earlier, an important consideration in the design of feedback for a team tutor is whether the feedback is addressed to an individual or the team, whether the feedback is presented to the individual or the team, and whether behavior assessment occurs at the individual or team level. For this experiment, two configurations were explored. In the Team condition, the tutor assessed the team as a single unit, addressed feedback to the whole team, and presented it to the whole team (both players simultaneously). In contrast, the Individual condition assessed each user individually and presented personal feedback just to the user who was assessed to need the feedback. The following examines the assessment and feedback process in greater detail and highlights the differences between conditions.

The GIFT learner module assesses each action a user takes into one of three levels (Above Expectations, At Expectations, or Below Expectations) based on a set of customizable conditions. For example, in the STT, an *Identify* action that occurred less than 5 seconds after the OPFOR has crossed zones was assessed as *Above Expectations*, while another *Identify* that took 5-10 seconds was assessed as *At Expectation,* and an *Identify* that took more than 10 seconds was assessed as *Below Expectations*. These assessments then trigger feedback statements depending on the action and the assessment.

From the learner dialogues above, one can see that the possible errors that can be made consist of omitting an action or performing an action with bad timing, e.g., too late. Feedback was designed accordingly (see Table 1). In a sense, omitting an action can be considered the ultimate in bad timing, so for Acknowledge and Identify tasks, which should be performed very quickly after the triggering event is presented, only a timing assessment was performed. For the Transfer task, timing was less critical; learners were deemed successful if the OPFOR were anywhere in a region near the zone border at the time of transfer. Therefore, omitting transfers and transfer timing were assessed separately. It is also worth noting that these feedback statements vary in their pedagogical strategy; some are simply reminders of the task, while others give advice. For this

initial prototype ITTS and pilot study, the research team did not focus on refining the instructional design of the feedback statements.

3.4 Feedback Filtering 1: Assess But Don't React

Given the continuous, high task load nature of the surveillance task, it was apparent from preliminary pilot testing that feedback would have to be given intermittently rather than for every action a user took. The STT took two different approaches to filtering. First, a custom feedback controller was implemented to assess learner performance continuously, but give feedback only periodically. Because this approach had not previously been used in GIFT, and because it may inspire other feedback filtering designs in the field, it is worth describing in some detail.

The controller consisted of three counts, which may be thought of as bins for the action assessments (see Figure 2). When the user performed an action, the tutor assessed that action as being Above, At, or Below Expectations according to the established performance criteria. That action was then added to an Above, At, or Below bin *according to the relationship of the most recent assessment to the learner's current state*. If the action assessment was above the current learner state, it went in the "Above" bin. If the action assessment was the same as the current learner state, it went in the "At" bin. If the action assessment was below the current learner assessment, it went in the "Below" bin. Whenever a bin was filled (five actions for "Above" and three for "At" and "Below"), the learner state would change to the corresponding bin, and the learner received feedback based on their new state.

The structure in Figure 2 was implemented for four assessments: Transfer presence, Transfer timing, Acknowledge timing, and Identify timing. Additional learner measures such Acknowledge omission rate (misses) were recorded for research purposes, but feedback was not given based on those measures.

Learners in the Individual feedback condition were assessed based only on their own actions. Learners in the Team feedback condition were assessed based on the collective actions of themselves and their partners. Per the bins arrangement in Figure 2, this difference can be described as each learner having his or her own set of bins in the Individual condition, while in the Team condition, there was only one set of bins, and assessments from both members contributed to them. These two approaches for feedback (Individual condition learners receive it for their own actions, but Team condition learners receive it for either teammate's actions) were grounded in the research described above on feedback, but pragmatically, they led to participants in the Team condition receiving twice as much feedback as those in the Individual condition. Indeed, using this approach, learners on a team with three members would receive roughly three times the feedback of learners in the Individual condition.

3.5 Feedback Filtering 2: Give Team Condition Learners Less Feedback

To balance the amount of feedback learners received in each condition, the authors implemented a second approach to feedback filtering. It was decided to give learners in the Team condition no feedback based on Transfer presence or Transfer timing, reducing the overall amount of feedback for Team condition learners to approximately the same as those in the Individual condition (see Table 1). This decision was based the rationale that the Transfer task was a required precursor to the Acknowledge and Identify tasks, and it was hoped that feedback on those latter two tasks would improve Transfer performance.

The need for this second approach to filtering illustrates one the difficulties in designing experimental conditions that isolate feedback design variables. While the authors had originally intended to design the Individual and Team feedback conditions to contain identical instructional

content and differ only by the three feedback design factors described above (team vs. individual feedback target, public vs. private feedback, team vs. individual assessment basis), it was difficult to reach this goal without affecting the amount of feedback given. The authors felt it was more important experimentally to hold the amount and frequency of feedback roughly constant rather than hold constant the content of the feedback. Thus, this change in content (not giving Team condition learners Transfer feedback) became a fourth experimental variable that differed across conditions.

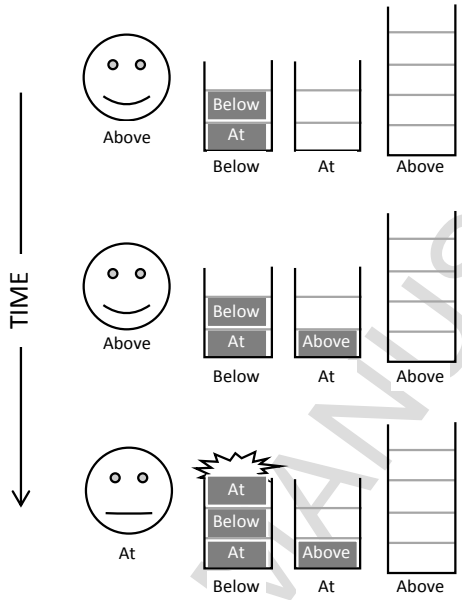


Figure 2. An example of the custom feedback controller that offers feedback in the STT only periodically while assessing performance at every action. In this example, a learner is at the Above assessment state. The learner has already taken an action assessed as At Expectation, but because this is below the current learner state (Above), it was counted in the Below bin. The learner's subsequent action was Below Expectation and also went in the Below bin. The learner's third action (second panel in figure) is Above Expectation, which goes in the At bin since the learner's state is currently Above. The fourth action (third panel) is assessed as At Expectation and goes in the Below bin for being below the learner state. At this time the Below bin is filled, so the learner state drops down to At Expectation (one level below Above), a feedback message is generated based on the new learner state, and the bins are emptied. This system requires five actions to fill the Above bin and trigger Above feedback, but only three actions for At or Below. Those thresholds could be adjusted according to the task and pedagogy chosen for feedback design.

4 Pilot Evaluation: Independent Variables, Hypotheses, and Predictions

A pilot study was conducted to evaluate the ability of the STT to provide feedback to positively influence team behavior and, by extension, improve team task performance. The first independent variable in this study focused on the impact of the Individual feedback condition versus Team feedback condition vs. a third No Feedback condition provided as a baseline. These feedback variables are unique to team tutoring contexts, and could affect the success of the collaborative learning experience. The second independent variable was trial number; each team performed the task for four consecutive trials. The following hypotheses reflect the effects the authors predicted seeing in three core areas: individual performance, team coordination, and self-assessment.

Because previous research suggests that participants who receive feedback would use it to improve their performance (DeShon et al., 2004): **H1 –** Teams who received feedback from the tutor (Team or Individual condition) would perform the surveillance task more quickly and with fewer errors than those in the No Feedback condition.

Because Team condition teammates were guided to refocus on the same behaviors together, this condition may aid in the development of a shared mental model of the task, an important driver of team coordination (Stout, Cannon-Bowers, Salas, & Milanovich, 1999), and improve performance. Thus: **H2 –** Teams receiving team feedback would perform better than teams who received individual or no feedback on team measures.

In addition to affecting the team's ability to coordinate, the feedback condition may also affect how users perceive their teammate. For a team to coordinate effectively, members must be able share relevant information freely (Simsarian Webber, 2002). Mutual trust can be developed if the team members believe that they share the same goals, which the Team feedback will promote. Thus: **H3 –** Participants in the Team feedback group would rate their teammates' performance higher than participants in the Individual and No Feedback groups.

An important secondary goal of feedback is to provide learners with an accurate understanding of their own performance. If learners believe they are doing a task well when they are not, they will not be able to rate their own performance accurately. The feedback used by the STT is designed to provide this assessment to the learners. Thus: **H4 –** Participants who received Individual or Team feedback would assess their own performance more accurately than those in the No Feedback condition.

5 Methods

5.1 Participants

Thirty-two participants (22 males, 10 females) participated in the study and made up a total of 16 two-person teams. Participant ages ranged from 18-35. All participants either attended or worked at a large university located in the Midwest and were not members of the military. Nine additional teams were also recruited, but their data were lost due to technical issues and was not included in the analysis. Six of the final teams were assigned to the No Feedback condition, four to the Individual condition, and six to the Team condition.

5.2 Dependent Variables and Metrics

Dependent variables were defined at both the individual and team levels. Individual task performance measures were derived directly from performance on the Transfer and Identify tasks because those tasks, when performed by an individual, did not directly affect the other team member. Team performance measures were derived from the Acknowledge task, acting as a behavioral marker for team coordination.

Participants were asked to complete self-assessment surveys and a NASA-TLX after each trial, which yielded self-assessment metrics. The post-session survey included the following five questions.

- Did you notice any feedback during the task? (Yes/No)
- Did you find the feedback helpful? (4-pt. Likert or N/A)
- My individual performance was… (5-pt. Likert)
- My team's performance was… (5-pt. Likert)
- My teammate performed poorly (6-pt Likert)

Table 2 provides an overview of the dependent variables measured in this study. Acquiring the data for these dependent variables from the team tutoring system architecture was a complex task beyond the scope of this paper, and is described elsewhere. Initial data analysis attempted to develop a single "roll-up" measure of team performance aggregated from multiple individual measurements (Authors, 2017a). However, after further consideration of concerns about the validity and consistency of roll-up measures (Cerully et al., 2017), the data were re-analyzed for this paper task-by-task based solely on a common metric of errors.

5.3 Experimental Design

This experiment was a 4 (Trials) X 3 (Feedback Groups) mixed experimental design. Each team dyad experienced four trials (within subjects, repeated measures). Teams were given Individual Feedback, Team Feedback, or No Feedback (between subjects).

5.4 Procedure

Each team member was placed in separate rooms with a computer and a speakerphone that was used to vocally communicate with their teammate. After a training video, the teams were asked to conduct four trials of the surveillance task. Feedback given was displayed on the left side of the screen. See Figure 3. Each trial lasted approximately five minutes.

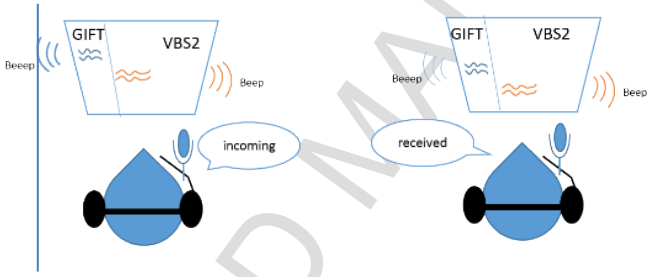


Figure 3. Illustration of the task and tutor setup.

5.5 Data Analysis Plan

Outliers were identified through assessment of studentized residuals. The Shapiro-Wilk test was used to check for normality of data. Levene's test of equality of error variances was used to test the homogeneity of variance. Box's test of equality of covariance matrices was used to check for similar covariances.

An attempt was made to analyze measured data with two-way mixed ANOVA tests; however, in most cases the dependent variables did not meet all the necessary assumptions of the two-way mixed ANOVA (likely due to small sample sizes), and they could not be transformed to do so. For consistency, all data were analyzed using nonparametric analyses. The main effect of the within-subjects factor, trial, was analyzed using Friedman's two-way analysis of variance by ranks. The main of effect of the between-subjects factor, feedback condition group, was analyzed using the Kruskal-Wallis H Test. The median and inter-quartile range (IQR) are reported in lieu of mean and standard variation to give a more accurate description of the non-normal data.

Post-hoc analysis used the Wilcoxon signed-rank test with a Bonferroni correction for multiple comparisons in order to distinguish trials that were significantly different from each other. The Mann-Whitney U test with a Bonferroni correction for multiple comparisons was used to distinguish Team vs. Individual feedback conditions. Adjusted p-values are presented in both

cases. All results are reported as significant for alpha < .05, and marginally significant for alpha < .10 (Gelman, 2013). Effect sizes *r* were calculated as the standardized test statistic *z* divided by the square root of the total sample size (Field, 2005). An effect size greater than 0.5 indicates a large effect; an effect size greater than 0.3 indicates a medium effect; an effect size greater than 0.1 indicates a small effect.

Pearson's correlation coefficient was computed to test the association between two variables: total errors made and self-assessment of task performance from the NASA-TLX. Error data were positively-skewed as assessed by histogram, but Spearman's rank order correlation yielded results similar to Pearson's, indicating the analysis was not affected.

Table 2: Dependent variables and descriptions

Dependent Variable	Metric	Units	Frequency
Individual Performance			
Missed Transfers	# of OPFOR without a corresponding Transfer	Count	Once per trial
Missed Acknowledges	# teammate Transfers without a corresponding Acknowledge	Count	Once per trial
Missed Identifies	# of OPFOR without a corresponding Identify	Count	Once per trial
Mean Cross-Identify Time	Mean time after OPFOR cross that participant Identified	Seconds	Once per trial
Team Coordination			
Transfer-Acknowledge Percent	# Successful Transfer-Acknowledge pairs / # OPFOR	%	Once per trial
Mean Transfer-Acknowledge Time	Mean time between teammate Transfer and Acknowledge	Seconds	Once per trial
"Triple" Percent	# Successful Transfer-Acknowledge-Identify Sequences / # OPFOR	%	Once per trial
Self-Assessment			
Self-Assessment	Survey Responses	Likert Scales	After each trial
Workload	NASA Task Load Index (Hart & Staveland, 1988)	Continuous scale 0-100	After each trial

5.6 Limitations

While the experiment did include audio recordings of verbal communication, the data were not able to be analyzed in real-time to inform feedback. Due to this constraint, the experiment relied on participant keypresses to log the actions taken, rather than the actual communications between team members. While this approach does introduce the opportunity for forgotten keypresses or accidental extra keypresses, post hoc data analysis demonstrated that these errors were sufficiently negligible.

6 Results & Discussion

6.1 The results for each of the dependent measures defined in Table 2 are presented in Table 3 (effect of trial) and Trial-by-Trial Learning Curve

Although none of the research hypotheses predicted specific effects by trial, measures were reported trial-by-trial to illustrate any potential learning curve (Table 3). Broadly speaking, most

measures did not significantly differ by trial, but when they did, Trial 1 was significantly worse in performance. This general trend suggests that it is possible to gain familiarity with the Surveillance Task in one trial. Only the measure Transfer-Acknowledge Time continued to improve significantly after Trial 1, perhaps suggesting that team members required more than one trial to get enough rapport to optimize this coordinated exchange. This result is worth exploring with a larger sample size.

6.2 Hypotheses 1 & 2: Feedback Condition Did Not Affect Performance

The results of this study (Table 4) did not strongly support Hypothesis 1, in that for most measures, neither Individual nor Team feedback resulted in a significant decrease in errors committed by the participants compared with No Feedback. However, the one exception was Missed Acknowledges, of which there were significantly fewer in the Team feedback condition. Because the participants in that condition had received feedback only on Acknowledges and Identifies, as noted above, while the Individual condition teams also received feedback on Transfers, the authors suspect that the proportionally higher emphasis on Acknowledge with the team condition led to this result. Thus, this result does suggest that feedback given even in a high-workload task such as the STT can lead to behavioral change.

The authors had hoped to see a stronger impact of feedback condition on performance. That said, it is worth mentioning that this is one of the first times an ITTS has been attempted with real-time feedback, so there are still many unknowns. It is possible, given this high-workload attention task as reflected in the NASA-TLX scores, that the feedback given was simply not helpful to participants for completing these tasks more quickly or with fewer errors. Also, the added task load of reading feedback may have offset the benefits, or they may have ignored the feedback and removed any effect (Kulik & Kulik, 1988). Hypothesis 2 was not supported in the study, in that none of the team measures demonstrated that learners in the Team condition performed better at team coordination.

6.3 Hypothesis 3: Team Feedback Learners Rated Teammates Less Poorly

The results of this study did support Hypothesis 3, in that participants who received the Team feedback condition rated their team and their teammate significantly lower on the statement "My teammate performed poorly" than both those in the No Feedback and the Individual feedback conditions. These more favorable ratings may stem from some key drivers of team development: shared mental models (Stout et al., 1999) and the ability to freely share relevant information (Simsarian Webber, 2002). When team members share a mental model of the tasks at hand, they can empathize with other teammates' performance and will likely evaluate them more positively.

In the STT specifically, participants with no feedback focused on whichever subtask they deemed most important, while participants with Individual feedback were guided to focus on the subtasks they need to improve on, which may have been different from their teammate's. Neither of these conditions are conducive to the development of a shared mental model of the state of the team; instead these conditions lead a learner to focus on him- or herself. In contrast, participants who received the Team feedback condition received the same feedback at the same times, guiding them to focus on the same subtasks and to build a mental model of the performance and state of their teammates. It is also possible that the increased communication indicated by the increase in Acknowledge events contributed to a more favorable team dynamic. The mixture of four factors that distinguish the Team vs. Individual conditions (as described above in Feedback Filtering) made it difficult to infer exactly which of these factors caused this difference, but these results do

indicate that an effect was present, even with a small sample population, and worth exploring further.

Table 4 (effect of feedback condition). The first question of the post-trial self-assessment survey, which asked whether participants noticed the feedback, was used as a binary control variable to ensure that participants didn't ignore the feedback. All participants who received feedback indicated seeing it, though one learner did not notice it until Trial 2. That said, a subsample of trials that were run with eye tracking monitoring revealed that during times with many OPFOR crossing zones, learners did not attend to every new feedback statement that appeared. Feedback Helpfulness included responses only from the two groups who received feedback; thus the Mann-Whitney U test was used instead of the Kruskal-Wallis H test.

Table 3: Dependent measures by trial.

Dependent Variable	Median (Inter-Quartile Range)				$\chi^2(3)$	p	Post-Hoc Analysis
	Trial 1	Trial 2	Trial 3	Trial 4			
Individual Performance							
Missed Transfers	6 (6)	6 (4)	5 (4)	6 (4)	3.655	.301	
Missed Acknowledges	6 (6)	4 (4)	3.5 (5)	3 (3)	10.01	.018	T1 > T2 (z = 2.71, p = .04, r = .34)
Missed Identifies	2.5 (3)	2 (2)	2 (1)	2 (2)	5.708	.127	
Cross-Identify Time (s)	2.4 (2.1)	2.4 (2.2)	2.2 (2.2)	2.0 (2.0)	3.037	.386	
Team Coordination							
Trans.-Ack. Percent (%)	44 (25)	53 (24)	53 (25)	57 (23)	6.225	.101	
Trans.-Ack. Time (s)	1.7 (1.6)	1.4 (0.9)	1.3 (0.9)	1.1 (0.9)	19.871	< .001	T1 > T3 (z = 2.76, p = .035, r = .35) T1 > T4 (z = 4.31, p < .001, r = .54) T2 > T4 (z = 2.66, p = .047, r = 33)
Triple Percent (%)	39 (21)	43 (16)	50 (20)	47 (25)	13.158	.004	T1 < T3 (z = 3.49, p = .003, r = .44)
Self-Assessment Questionnaire							
Feedback Helpfulness	3 (1)	3 (1)	2 (1)	3 (1)	0.28	.964	
Individual Assessment	2 (1)	3 (0)	3 (0)	3 (0)	20.74	< .001	T4 > T1 (z=*3.85*, p=.001, r=*.58*)
Team Assessment	2 (1)	3 (0)	3 (0)	3 (1)	7.15	.067	T4 > T1 (z=3.56, p=.002, r=.54)
My Teammate (low is best)	2 (0)	2 (1)	1.5 (1)	1.5 (1)	2.30	.513	
NASA Task Load Index							
Mental Demand	55 (30)	52.5 (18)	50 (35)	45 (31.3)	1.18	.757	
Temporal Demand	55 (12.5)	50 (16)	50 (26)	40 (35)	6.57	.087	No differences found post-hoc.
Performance	30 (12.5)	22.5 (16)	22.5 (16)	10 (10)	16.34	< .001	T4 < T1 (z=.277, p=.034, r=.45)
Effort	52 (17.5)	40 (21.3)	47.5 (31)	40 (36.3)	1.40	.707	

Frustration	40 (41.3)	40 (36.3)	40 (35)	35 (31.3)	3.71	.295	
Task Load	45 (16.3)	45 (16.3)	45 (16.3)	40 (22.5)	4.50	.212	

6.4 Trial-by-Trial Learning Curve

Although none of the research hypotheses predicted specific effects by trial, measures were reported trial-by-trial to illustrate any potential learning curve (Table 3). Broadly speaking, most measures did not significantly differ by trial, but when they did, Trial 1 was significantly worse in performance. This general trend suggests that it is possible to gain familiarity with the Surveillance Task in one trial. Only the measure Transfer-Acknowledge Time continued to improve significantly after Trial 1, perhaps suggesting that team members required more than one trial to get enough rapport to optimize this coordinated exchange. This result is worth exploring with a larger sample size.

6.5 Hypotheses 1 & 2: Feedback Condition Did Not Affect Performance

The results of this study (Table 4) did not strongly support Hypothesis 1, in that for most measures, neither Individual nor Team feedback resulted in a significant decrease in errors committed by the participants compared with No Feedback. However, the one exception was Missed Acknowledges, of which there were significantly fewer in the Team feedback condition. Because the participants in that condition had received feedback only on Acknowledges and Identifies, as noted above, while the Individual condition teams also received feedback on Transfers, the authors suspect that the proportionally higher emphasis on Acknowledge with the team condition led to this result. Thus, this result does suggest that feedback given even in a high-workload task such as the STT can lead to behavioral change.

The authors had hoped to see a stronger impact of feedback condition on performance. That said, it is worth mentioning that this is one of the first times an ITTS has been attempted with real-time feedback, so there are still many unknowns. It is possible, given this high-workload attention task as reflected in the NASA-TLX scores, that the feedback given was simply not helpful to participants for completing these tasks more quickly or with fewer errors. Also, the added task load of reading feedback may have offset the benefits, or they may have ignored the feedback and removed any effect (Kulik & Kulik, 1988). Hypothesis 2 was not supported in the study, in that none of the team measures demonstrated that learners in the Team condition performed better at team coordination.

6.6 Hypothesis 3: Team Feedback Learners Rated Teammates Less Poorly

The results of this study did support Hypothesis 3, in that participants who received the Team feedback condition rated their team and their teammate significantly lower on the statement "My teammate performed poorly" than both those in the No Feedback and the Individual feedback conditions. These more favorable ratings may stem from some key drivers of team development: shared mental models (Stout et al., 1999) and the ability to freely share relevant information (Simsarian Webber, 2002). When team members share a mental model of the tasks at hand, they can empathize with other teammates' performance and will likely evaluate them more positively.

In the STT specifically, participants with no feedback focused on whichever subtask they deemed most important, while participants with Individual feedback were guided to focus on the subtasks they need to improve on, which may have been different from their teammate's. Neither

of these conditions are conducive to the development of a shared mental model of the state of the team; instead these conditions lead a learner to focus on him- or herself. In contrast, participants who received the Team feedback condition received the same feedback at the same times, guiding them to focus on the same subtasks and to build a mental model of the performance and state of their teammates. It is also possible that the increased communication indicated by the increase in Acknowledge events contributed to a more favorable team dynamic. The mixture of four factors that distinguish the Team vs. Individual conditions (as described above in Feedback Filtering) made it difficult to infer exactly which of these factors caused this difference, but these results do indicate that an effect was present, even with a small sample population, and worth exploring further.

Table 4: Dependent measures by feedback condition.

Dependent Variable	Median (Inter-Quartile Range)			$\chi^2(2)$	p	Post-Hoc Analysis
	No Feedback	Individual Feedback	Team Feedback			
Individual Performance						
Missed Transfers	6 (6)	5 (2)	5.5 (5)	0.344	.842	
Missed Acknowledges	4 (3)	5 (6)	3 (3)	9.850	.007	Team < Ind. (z = 2.89, p = .012, r = .32) Team < No F. (z = 2.55, p = .032, r = .27)
Missed Identifies	2 (2)	2 (3)	2 (2)	1.240	.538	
Cross-Identify Time (s)	1.9 (2.0)	2.4 (2.6)	2.5 (2.5)	5.468	.065	No differences found post-hoc.
Team Coordination						
Trans.-Ack. Percent (%)	51 (20)	51 (30)	57 (23)	1.800	.407	
Trans.-Ack. Time (s)	1.2 (1.3)	1.4 (1.0)	1.3 (0.7)	0.363	.834	
Triple Percent (%)	42 (21)	42 (25)	47 (17)	3.034	.219	
Self-Assessment Questionnaire						
Feedback Helpfulness	n/a	3 (1.25)	2 (1)	n/a	.127	
Individual Assessment	3 (0)	3 (1)	3 (0)	4.58	.102	
Team Assessment	3 (0)	3 (1)	3 (0)	9.21	.100	No F. > Ind. (z=2.905, p=.011, r=.44) No F. > Team (z=.2123, p=.1, r=.32)
My Teammate (low is best)	2 (.25)	2 (1)	1 (1)	7.76	.021	Team < Ind. (z=2.666, p=.023, r=.4) Team < No F. (z=2.87, p=.086, r=.33)
NASA Task Load Index						
Mental Demand	47.5 (35)	45 (42.5)	55 (6)	2.37	.306	
Temporal Demand	50 (26)	50 (30)	53 (7.5)	.73	.692	
Performance	15 (20)	30 (31)	22.5 (21)	12.24	.002	Ind. > No F. (z=-3.41, p=.002, r=.43)
Effort	45 (36)	37.5 (16)	50 (6)	5.87	.053	Ind. < Team (z=-3.12, p=.005, r=.45)
Frustration	22.5 (51)	32.5 (25)	50 (5)	12.54	.002	No F. < Team (z=-2.84, p=.004, r=.38) Ind. < Team (z=-3.52, p < .001, r=.56)
Task Load	40 (20)	42.5 (25)	50 (6)	6.03	.049	No F. < Team (z=-2.56, p=.031, r=.34)

6.7 Hypothesis 4: Team Feedback Learners Assessed Own Performance More Accurately

Hypothesis 4 was partially supported in the study. Participants who received the Team feedback condition were able to assess their own performance significantly more accurately than other participants. Participants who received the Individual feedback condition did not share this

benefit of higher accuracy; however, the wide range of self-assessments indicate that some participants in this group did rate themselves lower than those who did not receive feedback. This lower self-rating can be viewed positively, because overestimation in self-assessments is a common and well-documented problem in many disciplines (Dunning, Heath, & Suls, 2004; Gregersen, 1996; Kruger & Dunning, 1999), and such an overestimation trend is seen amongst participants who received no feedback. Thus, the feedback received in the Individual condition did not yield more accuracy, but possibly contributed to removing the overestimation effect seen in the No Feedback participants.

To address Hypothesis 4, Pearson's correlation coefficient was computed for each feedback condition to characterize the relationship between each participant's self-assessment of performance on the NASA-TLX (averaged across trials), and the total number of errors committed (averaged across trials). An accurate self-assessment of performance would negatively correlate with errors; the more errors made, the lower one's self-assessment should be. Participants with No Feedback assessed their performance highly overall, no matter how many errors they made, and there was no significant correlation (r = -.17, p = .631). Participants with Individual Feedback assessed themselves with values that ranges from very low to very high, seemingly randomly, independent of the errors they made. There was no significant correlation (r = -.08, p = .880). However, participants who received Team feedback had self-assessment ratings strongly correlating with errors made, decreasing linearly with errors made (r = -.96, p = .003).

6.8 Team Feedback Increased Workload

One factor to consider is that of workload. The results of this study revealed a possible unintended consequence of the feedback conditions. While the Team feedback condition appeared to have had the most beneficial effects on learners, it also appeared to have significantly increased the workload of the participants who received it, especially their self-assessments of effort and frustration. This reinforces the concept that care should be taken to ensure the benefits of feedback are not outweighed by the increased attentional costs.

7 Conclusions: Towards the Perfect Team Task

The challenges of developing and analyzing this team task point us to the potential value of a framework for evaluating a team task's fit for automated training by an ITTS. While these results do not yet provide sufficient evidence for a complete framework, the following requirement recommendations emerged from the experiences with this team task design which could serve as initial framework components. While some recommendations apply to any ITS, longer explanations are offered for requirements related specifically for ITTSs.

Team performance variability should be constrained as much as possible. Since any team task will likely yield high performance variability due to individual differences in both the team members' team skills and their task skills, it is useful in a research context to reduce variability in any way possible. While this conclusion may seem obvious, since researchers typically attempt to reduce sources of variability to the independent variable, it is worth emphasizing in the study of teams because the highly variable noise from individual differences can easily overwhelm the signal created by an independent variable. Variation in team performance should not be driven by task familiarity, for example. Ideally, tasks are novel for learners or equally familiar to all learners. Variation due to team member familiarity should be removed as well, either by ensuring that all team members are strangers or that all teams have members with a similar level of rapport. In the

STT, high variability in team performance made it difficult to discern the impact of the independent variables.

Carefully consider interdependencies between team members. Interdependency is higher if one team member's performance on a task depends strongly on the success of another member's performance. In the STT, a learner could not be successful at the Acknowledge task if the other teammate did not succeed at communicating Transfers. If the overall task includes no task dependencies, it is difficult to argue that learners are participating in a task that requires team skills. While there are "team tasks" with low interdependency such as swim team championships, in which the final team score is a function of each swimmer's individual swim times, tasks for which ITTSs would be constructed would likely have more task dependence. However, if task dependencies are too strong, e.g., in which only Person A can do Subtask 1, and only Person B can do Subtask 2, etc., then the overall task will not allow backup behavior (one member helping out the other in a pinch), which is a key measure of team trust and rapport. These interdependencies can be complex because they emerge both from the relationships of the subtasks themselves and from the job roles given to different team members. The PISA framework (Organisation for Economic Cooperation and Development, 2016), for example, carefully considers interdependencies in terms of the different information that team members may have, e.g., in a jigsaw or hidden profile task.

Task performance should be noticeably and steadily improvable with tutoring. If learners can master the tasks solely through practice (without tutoring), or if learners' performance is not likely to significantly improve no matter how much practice and tutoring they receive, a tutor is not helpful. The STT results, showing improvement from Trial 1 to Trial 2 but not much change after that, even with tutoring, suggest that tutoring on this task may not be tightly tied to performance improvement.

Feedback should be perceivable, useful, and transparently triggered. If the task required too high a cognitive load and feedback cannot be internalized without a task performance decrement, the feedback may not be perceivable. The STT workload results suggest that while the Team condition was useful, it cost learners mental workload to consider the team. Also, if feedback is consistently overlooked because of task load, the feedback may need to be moved to an after-action review. If attending to the feedback does not result in a performance boost experienced by the learner, it will be likely perceived as not useful. Lastly, the triggering mechanisms of the feedback should be transparent. In the STT, while the bin-based feedback filtering method was successful at not overwhelming learners with feedback, it decreased the system transparency, so that it was not always clear to learners which of their actions triggered the feedback they received. The team-bin approach reduced this transparency further.

There should be measurable actions that serve as necessary and sufficient evidence of task completion. In complex real-world tasks, the constructs that one wants to measure are often not possible to measure directly and objectively, e.g., mental workload or team trust. Instead, researchers choose observable behavioral markers as proxies for these constructs (Sottilare, Burke, et al., 2018). This recommendation suggests that researchers adopt a similar approach to designing team tasks so that there is always a measurable indicator of task completion, and ideally, task progress.

For example, imagine a context in which a learner's goal is to move to a certain location in a video game and the learner's actual location in the game cannot easily be identified or recorded. As long as the location can be inferred from the sequence of keystrokes and game controller actions taken by the learner, and those actions can be recorded, then this requirement is met. As another

example, imagine that a team's goal is to make a cake following a specific recipe that includes milk and eggs, but the only observable measure is viewing the final cake. In this case, this recommendation would not be met because the inclusion of milk and eggs could not be easily assessed; there would be no measurable actions that could verify that the recipe was followed.

In the STT, this type of issue arose because it was not feasible to map Transfer and Identify actions onto specific OPFOR. Because multiple OPFOR could cross a zone border simultaneously, when a player said, "Transfer," it was not always clear which OPFOR the player intended. This situation led to analyzing performance by aggregating, e.g., counting whether three transfers were noted after three OPFOR crossed the boundary, rather than assessing the specific task that was desired: whether a specific OPFOR was successfully transferred.

This problem also rose with Acknowledge, an STT team task. Because every Transfer should trigger an Acknowledge, it was unclear after a sequence of multiple Transfers which Acknowledge mapped to which Transfer. Also, if there were three or more team members, it would be unclear which learner was being acknowledged. Finally, when learners saw two OPFOR crossing at the same time, it was allowed for the players to say something like, "2 transferring at the 2-pole." Although players typed the 2-key twice during this communication to communicate two transfers, in customary dialogue, this verbal statement would merit only one Acknowledgment, not two. This inconsistency of action-to-task mapping meant this necessary-and-sufficient-evidence recommendation was not satisfied.

8 Summary & Future Work

Despite the breadth of questions yet to be answered in the domain of intelligent team tutoring systems, the STT and this study represent a step forward for the field of evaluating collaborative problem solving. The Surveillance Team Tutor is one of the first ITTSs and the first built in GIFT. Results showed that feedback within computerized team tutors can affect behavior. However, even for a simple team task, the development of an effective team tutor is complex. Despite significant advances in computer technology, it is often difficult to objectively measure all the behaviors of collaborating team members, e.g., communication intent (discerning the message behind the words and gestures) and why a team member is taking a specific action (learner intent). These are area for future work. More specifically, the results of this pilot study offer recommendations to inform the design of future team tasks.

One such area is feedback timing. The authors suggest that further work is warranted to investigate the effects of feedback timing and content. In cases of high learner workload, an after-action review may be more appropriate for learning than real-time feedback. Alternatively, in tasks where task load varies over time, a system could be designed to time feedback with periods of lower workload to maximize learners' available attention.

Finally, the Surveillance Task is a very simplified scenario; there were only two members on the team and they shared the same role. This was intentional to promote the development of shared task understanding, but future work should expand on this work to engage with larger teams with multiple roles.

9 References

Authors. (2015). *Conference paper.*
Authors. (2016). *Conference paper*.
Authors. (2017a). *Conference paper*.

Authors. (2017b). *Conference paper*.

Authors. (2017c). Journal Article.

Bonner, D., Gilbert, S., Dorneich, M. C., Burke, S., Walton, J., Ray, C., & Winer, E. (2015). Taxonomy of Teams, Team Tasks, and Tutors. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2)* (pp. 189-198).

Bonner, D., Ouverson, K., Gilbert, S., Sinatra, A. M., Dorneich, M. C., Winer, E., Slavina, A., MacAllister, A., & Kohl, A. (2017). Operationalizing the C's of Teamwork in an Intelligent Tutoring System. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, pp. 745-749). Los Angeles, CA: SAGE Publications.

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., & Volpe, C. E. (1995). Defining competencies and establishing team training requirements. In R. A. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations* (pp. 333-381). San Francisco: Jossey-Bass.

Cerully, J. L., Martino, S. C., Rybowski, L., Finucane, M. L., Grob, R., Parker, A. M., Schlesinger, M., Shaller, D., & Martsolf, G. (2017). Using "roll-up" measures in healthcare quality reports: perspectives of report sponsors and national alliances. *Am J Manag Care, 23*(6), e202-e207.

Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 245-252). Seattle, Washington: ACM.

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. *The Journal of applied psychology, 89*(6), 1035-1056. doi: 10.1037/0021-9010.89.6.1035

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology, 41*(1), 417-440.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest, 5*(3), 69-106.

Field, A. (2005). *Discovering statistics using SPSS, 2nd ed*. Thousand Oaks, CA, US: Sage Publications, Inc.

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., Massey, C., O'Neil, H., Pellegrino, J., Rothman, R., Soulé, H., & von Davier, A. (2017). *Collaborative Problem Solving: Considerations for the National Assessment of Educational Progress*. Alexandria, VA: National Center for Educational Statistics.

Flor, M., Yoon, S.-Y., Hao, J., Liu, L., & von Davier, A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 31-41). San Diego: Association for Computational Linguistics.

Gabelica, C., Van den Bossche, P., Segers, M., & Gijselaers, W. (2012). Feedback, a powerful lever in teams: A review. *Educational Research Review, 7*(2), 123-144.

Gelman, A. (2013). Commentary: P Values and Statistical Practice. *Epidemiology, 24*(1), 69-72. doi: 10.1097/EDE.0b013e31827886f7

Graesser, A. C., Cai, Z., Hu, X., Foltz, P. W., Greiff, S., Kuo, B.-C., & Shaffer, D. W. (2017). Assessment of collaborative problem solving. In R. Sottilare, A. Graesser, X. Hu & G. Goodwin (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Assessment Methods* (Vol. 5, pp. 275-286). Orlando, FL: US Army Research Laboratory.

Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., & Germany, M.-L. (2018). Challenges of Assessing Collaborative Problem Solving. In E. Care, P. Griffin & M. Wilson (Eds.), *Assessment and Teaching of 21st Century Skills: Research and Applications* (pp. 75-91). Cham: Springer International Publishing.

Gregersen, N. P. (1996). Young drivers' overestimation of their own skill: An experiment on the relation between training strategy and skill. *Accident Analysis & Prevention, 28*(2), 243-250.

Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. C. (2017). Initial Steps Towards a Standardized Assessment for Collaborative Problem Solving (CPS): Practical Challenges and Strategies. In A. A. von Davier, M. Zhu & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 135-156). Cham: Springer.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139-183): North-Holland.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology, 77*(6), 1121.

Kulik, J. A., & Fletcher, J. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research, 86*(1), 42-78.

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of educational research, 58*(1), 79-97.

Kumar, R., Ai, H., Beuth, J. L., & Rosé, C. P. (2010). Socially capable conversational tutors can be effective in collaborative learning situations. In *International Conference on Intelligent Tutoring Systems* (pp. 156-164). Berlin: Springer.

McGrath, J. E. (1984). *Groups: Interaction and performance* (Vol. 14): Prentice-Hall Englewood Cliffs, NJ.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegemann, D. Leutner & R. Brünken (Eds.), *Instructional design for multimedia learning* (pp. 181-195). Münster: Waxmann.

Organisation for Economic Cooperation and Development. (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris: OECD Publishing.

Organisation for Economic Cooperation and Development. (2017). *PISA 2015 Results (Volume V): Collaborative Problem Solving*. Paris: OECD Publishing.

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. doi: 10.1002/hrm.21628

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "big five" in teamwork? *Small group research, 36*(5), 555-599.

Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *Journal of Educational Measurement, 54*(1), 12-35.

Simsarian Webber, S. (2002). Leadership and trust facilitating cross-functional team success. *Journal of Management Development, 21*(3), 201-214. doi: doi:10.1108/02621710210420273

Soller, A. (2001). Supporting social interaction in an intelligent collaborative learning system. *International Journal of Artificial Intelligence in Education (IJAIED), 12*, 40-62.

Sottilare, R., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2012). The Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: US Army Research Laboratory–Human Research & Engineering Directorate (ARL-HRED).

Sottilare, R. A., Baker, R. S., Graesser, A. C., & Lester, J. C. (2018). Special Issue on the Generalized Intelligent Framework for Tutoring (GIFT): Creating a Stable and Flexible Platform for Innovations in AIED Research. [journal article]. *International Journal of Artificial Intelligence in Education, 28*(2), 139-151. doi: 10.1007/s40593-017-0149-9

Sottilare, R. A., Burke, C. S., Salas, E., Sinatra, A. M., Johnston, J., & Gilbert, S. B. (2018). Designing Adaptive Instruction for Teams: A Meta-Analysis. *International Journal of Artificial Intelligence in Education, 28*(2), 225-264.

Sottilare, R. A., Holden, H., Brawner, K., & Goldberg, B. (2011). Challenges and Emerging Concepts in the Development of Adaptive, Computer-based Tutoring Systems for Team Training. *U.S. Army Research Laboratory, Human Research and Engineering Directorate, Orlando, Florida*.

Sottilare, R. A., Shawn Burke, C., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2017). Designing Adaptive Instruction for Teams: a Meta-Analysis. *International Journal of Artificial Intelligence in Education*. doi: 10.1007/s40593-017-0146-z

Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, Shared Mental Models, and Coordinated Performance: An Empirical Link Is Established. *Human Factors, 41*(1), 61-71. doi: 10.1518/001872099779577273

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227-265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197-221.

Walton, J., Bonner, D., Walker, K., Mater, S., Dorneich, M., Gilbert, S., & West, R. (2015). The Team Multiple Errands Test: A Platform to Evaluate Distributed Teams. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (pp. 247-250). Vancouver: ACM.

Walton, J., Gilbert, S., Winer, E., Dorneich, M., & Bonner, D. (2015). Evaluating Distributed Teams with the Team Multiple Errands Test. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)* (pp. 1-12, Paper 15264). Orlando, FL.

Zachary, W., Cannon-Bowers, J., Bilazarian, P., Krecker, D., Lardieri, P., & Burns, J. (1998). The Advanced Embedded Training System (AETS): An Intelligent Embedded Tutoring System for Tactical Team Training. *International Journal for Artificial Intelligence in Education, 10*, 257-277.

Acknowledgments

The research described herein was supported by a cooperative agreement with the U.S. Army Research Laboratory. 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.

Highlights for "Evaluation of an Intelligent Team Tutoring System for a Collaborative Two-Person Problem: The Surveillance Team Tutor"

- An Intelligent Team Tutoring System was built for a two-person collaborative task.
- A unique method of filtering feedback was used to avoid overwhelming learners.
- A pilot study compared no feedback, individual feedback, and team feedback.
- Feedback affected performance, teammate ratings, and accuracy of self-assessment.
- Recommended requirements for the ideal team task for tutoring are provided.