Developing and Evaluating an Intelligent Tutoring System for Advanced Shiphandling

Jason H. Wong, Lauren Ogren
Naval Undersea Warfare Center
Newport, RI
jason.h.wong.civ@mail.mil, lauren.ogren@navy.mil

Stanley Peters, Elizabeth O. Bratt
Stanford University
Stanford, CA
peters@stanford.edu, ebratt@stanford.edu

ABSTRACT

The goal of an Intelligent Tutoring System (ITS) is to improve training efficiency by monitoring student performance and providing automated tutoring advice with the goal of increasing student learning and throughput. Traditional ITS development has focused on static problems, such as math and physics (Koedinger, Anderson, Hadley, & Mark, 1997; VanLehn, et al., 2005). Recent systems have targeted dynamic environments, such as Navy shipboard damage control and basic shiphandling maneuvers (Iseli, Koenig, Lee & Wainess, 2010; Peters, Bratt & Kirschenbaum, 2011). The research described here examines the advanced shiphandling task of mooring to a pier, which is the graduation exercise at the Surface Warfare Officers School (SWOS). To develop an ITS for mooring, many variables were considered, including ownship parameters (e.g., engine and rudder status), predicted future paths, and student behavior (e.g., number of orders, gaze direction). This development process involved creating and vetting a task analysis with SWOS subject matter experts (SMEs) and several iterations of system prototype testing. An effectiveness evaluation of the prototype was conducted with twenty novice shiphandling students at SWOS, split into groups that received either human or ITS tutoring only for a mooring to a pier scenario. Afterward, all students completed another mooring scenario without any tutoring. Across both runs, performance was evaluated using ship parameters, student behavior, and instructor scoring metrics. Analyzing a wide variety of performance measures showed no differences between the two groups, suggesting that the ITS was able to tutor as effectively as human instructors. Future work will involve developing additional advanced shiphandling scenarios and examining how the student-to-teacher ratio can be increased using a combination of ITS tutoring and instructor supervision.

ABOUT THE AUTHORS

Dr. Jason H. Wong is a Human Factors Scientist with the Naval Undersea Warfare Center. He designs and conducts experiments to examine warfighter performance in complex systems and the efficacy of innovative training methodologies. His current projects include examining the cognitive dynamics of distributed teams.

Ms. Lauren Ogren is a Human Systems Engineer at the Naval Undersea Warfare Center. She performs user-centered redesigns of interfaces through user interviews, task analyses, and applying usability principles. Her current projects concentrate on Navy user requirements, usability concerns, and the analysis of physiological sensors for unobtrusive performance measurement.

Prof. Stanley Peters is professor emeritus of linguistics at Stanford University and a past director of its Center for the Study of Language and Information. He has a research focus on interactive spoken language technology, including artificially intelligent tutoring systems. Prof. Peters is a Fellow of the Linguistic Society of America.

Dr. Elizabeth O. Bratt is a Senior Research Engineer in the Computational Semantics Laboratory of the Center for Study of Language and Information (CSLI) at Stanford University. Her research at CSLI has focused on how natural language interfaces can support learning in intelligent tutoring systems for simulation-based training.

Developing and Evaluating an Intelligent Tutoring System for Advanced Shiphandling

Jason H. Wong, Lauren Ogren
Naval Undersea Warfare Center
Newport, RI
jason.h.wong@navy.mil, lauren.ogren@navy.mil

Stanley Peters, Elizabeth O. Bratt
Stanford University
Stanford, CA
peters@stanford.edu, ebratt@stanford.edu

INTRODUCTION

In this paper, we review the complex and dynamic nature of the shiphandling domain, the strengths of ITSs for assisting instruction in this area, the extension of our ITS for the mooring exercise, and present an evaluation of its effectiveness compared with human instruction.

SIMULATION-BASED TRAINING FOR SHIPHANDLING

Many activities that require application of theoretical knowledge in operating expensive, potentially dangerous equipment are taught with the aid of high fidelity simulators. Research confirms that simulators effectively accelerate the acquisition of practical knowledge in such domains (Gredler 2004). Conning a ship, that is, driving the ship by ordering the amount of engine power to use, what position to put the rudder in, and so on, is a prime case in point. Masters of the art of shiphandling have a keen intuitive sense of ships' position and momentum, forces acting on them, and their future path, based on observing and interpreting a wide range of informative cues, along with a subtle understanding of the use of engines, rudder, etc. to direct their movement. To guide learners in developing correct mental models based on their practice at conning simulated ships, it is common for an experienced shiphandler to supervise each student constantly throughout the exercise in order to accelerate development of the many essential shiphandling skill components, including observation, interpretation, mental modeling of complex interactions, planning, and timing of actions in a dynamic environment. We discuss these aspects of shiphandling in more detail below.

Visual observation is critical for awareness of angles and both absolute and relative motion, especially angular relative motion and visual flow. When a student is approaching a pier or passing by other ships, noticing these details is critical, because the student will need this information as factors in conning decisions. Simulation-based training aims to provide practice in this area.

The conning officer must understand how the ship is affected by the multiple forces determining the ship's trajectory, and how to change the forces he controls to make the ship move in the desired direction. He must observe and interpret multiple sources of information, in order to determine whether to order changes in the ship's rudder and engines, or changes in a tug's actions, that will counteract or supplement the forces of the current, wind and the ship's momentum appropriately. The conning officer must interpret the effects of these changes, and order further changes of the appropriate amount at the appropriate time. Conning a ship skillfully requires quickly integrating a broad array of information into an intuitive mental conception of physical forces on the ship (whether or not the conning officer consciously thinks of this cognitive construct in terms of physics), which permits the officer to anticipate how his ship will respond to changes of forces he controls, and let him gauge whether the ship actually responds as expected.

One example of the complexity of the conceptual model for shiphandling is that it must include how the ship's pivot point moves over ground under various conditions of current and wind, and various speeds and engine power settings, because the conning officer must understand the details of how the bow and stern will move separately, not simply the ship as a whole. Another example of the complex skills in shiphandling is control of the ship's heading, which both influences the ship's movement over ground and determines what part of that motion is lateral and what part is parallel with the ship's keel.

Typically an instructor supervises each learner throughout the simulation experience, in order to provide the most effective instruction. One-on-one tutoring by expert instructors is known to be the most effective aid to learning (Bloom 1984, VanLehn 2011); however, the labor costs to provide this intensive, expert instruction can be prohibitive. Digital artificially intelligent tutoring systems (ITSs) can serve to supplement expert tutoring at low ongoing cost.

ITSs for Simulation-Based Training in Dynamic Domains

ITSs have been shown to be highly effective in teaching students to solve word problems in domains such as mathematics and physics (VanLehn 2011, Graesser et al. 2005, Koedinger et al. 1997). Such systems implement a combination of detailed software models: for the subject domain, for individual students, for skillful tutoring, and for interaction between a learner and the ITS. While learning to conn a ship shares many similarities with learning to solve academic physics problems, it also contains many stark differences. Recent advances have yielded ITSs that are effective tutors of students learning to solve dynamic problems – problems that change while learners are thinking about how to solve them (Iseli et al. 2010, Murray 2006, Martens and Himmelspach 2005, Peters et al. 2011). Thus, ITSs are a promising approach to increasing the consistency and effectiveness of simulation-based learning and reducing ongoing instructor costs.

Beyond aiming for simple parity with human instruction, digital ITSs may have advantages over human instructors in the breadth of relevant detail they can incorporate into their model of a student's current performance, because ITSs can fully utilize the instrumentation of learners and their simulated world as used by virtual-reality simulators for presenting realistic visual and auditory stimuli. A shiphandling ITS connected to a ship simulator can continuously measure where a student is looking, what controllable forces he orders, what other forces are acting on the simulated ship, and how it is responding. From these measurements plus information about the learning objectives, the shiphandling knowledge required, and standards of acceptable performance, the ITS can assess a learner's degree of proficiency at the observational, interpretive, control, and anticipatory skills involved in the lesson. Based on this estimate of the student's proficiency, it can decide when to offer assistance and what tutoring to provide.

Another area in which an ITS may have especially strong capabilities is in the consistency of its assessment. An ITS is never distracted and the measurements it relies on are objective. Accordingly, its assessments can be consistently comprehensive, and unaffected by human variation in attention or leniency.

Effective Pedagogy for Dynamic Tasks

Science of learning research shows that dynamic tasks are most effectively taught in order of increasing difficulty, with earlier learning objectives laying foundations for later ones (Gonzalez 2012, VanLehn 2011). Lessons should be structured to reinforce shared building blocks while teaching new ones and to pose challenges that deepen learners' knowledge of previously acquired skills. Dynamic problems are also best approached in order of increasing stressfulness.

Human instructors generally follow this model in simulation-based training of shiphandling, by starting with easier tasks involving simpler skills, and increasing to more complex ones. For example, getting a ship underway from a pier is both easier and less stressful than mooring one. For getting underway, the ship begins securely attached to a pier, and generally moves an increasing distance from hazards of collision or running aground. Peters et al. (2011) showed that a digital ITS could tutor this task as effectively as human instructors. Mooring, however, begins with the ship in motion and takes it successively closer to collision hazards, including the pier that it must eventually contact. Some shiphandling skills are common to both tasks, such as maneuvering with steerageway (i.e., when the ship is going at sufficient speed for the rudder to have an effect) and maneuvering without steerageway (using tugs and engines along with the rudder). Other skills are not common to the two tasks, and certain challenges create different difficulties for them. For instance, it is easier to maneuver the ship in a direction opposite the environmental forces of current and wind. Thus, environmental forces pushing the ship toward the pier are easier to handle when getting underway and harder when mooring, while environmentals pushing away from the pier are the reverse.

SHIPHANDLING TASK ANALYSIS AND ITS DEVELOPMENT

Developing an ITS assisting a conning officer in the tutoring of shiphandling in simulation requires explicitly specifying the information a human instructor would use, namely:

- what learning objectives a given lesson involves
- what constitutes an acceptable 'solution of the problem'
- what knowledge the student must deploy to achieve sub-objectives in the lesson
- what knowledge the student has previously demonstrated possession of
- what things are generally helpful to say to students struggling with particular sub-objectives or specific aspects of acceptable solutions
- when a student is likely to learn more from continuing to try on his own, than from being offered help

We elicited information about these matters from subject matter experts: master mariners and uniformed instructors at the U.S. Navy's Surface Warfare Officers School (SWOS) in Newport, Rhode Island. We have used a similar process for studying other shiphandling exercises in our previous development (Peters, Bratt & Kirschenbaum, 2011), but this paper reports specifically on our work on the Mooring to a Pier evolution. The initial task analysis collected data from shiphandlers at a variety of skill levels. We recorded moment-by-moment ship information, including lateral and longitudinal speed, heading, rudder direction, and engine power, along with audio recordings of the students and their instructors. Eighteen runs were collected from novice ship handlers (students in the Advanced Shiphandling and Tactics [ASAT] class), six runs from intermediate students (students in the Prospective Commanding/Executive Officer class), and eight runs from experts (Lieutenant instructors and Master Mariners). This provided a picture of how Mooring was completed by ship handlers at a variety of skill levels, which helped to set parameters of acceptability for the ITS to employ. Our knowledge acquisition and analysis also contributed to our collaboration with a research team developing an automated assessment engine for shiphandling, to interface with our ITS (Koenig 2014). Figure 1 contains a small snippet of the task analysis that was completed for Mooring to a Pier, as diagrammed by the UCLA CRESST team.

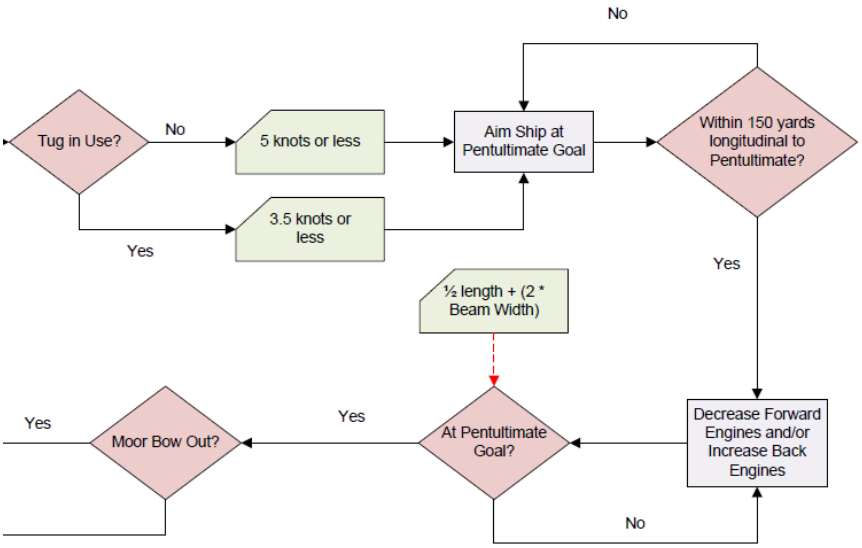

Figure 1. A partial task analysis for Mooring to a Pier (from UCLA CRESST).

Shiphandling ITS Architecture

The ITS software developed and assessed in this research monitors multiple variables once per second via the ship simulator's Application Program Interface (API). These include variables for the conning officer's observations (azimuth and elevation of gaze; viewpoint from the bridge or a bridge wing), the ship (position, heading, speed, rudder and engine settings, etc.), the environment (current, wind, water under keel, etc.), and the conning officer's decisions (orders given to helm, lee helm, tug, etc.). SWOS uses the Conning Officer Virtual Environment (COVE) ship simulator that Computer Sciences Corporation (CSC) built in its VShip simulation system. From the variables monitored, the ITS calculates other values such as the ship's distance from other objects and its projected path, position, and heading at a range of future times (assuming current forces on the ship do not change).

From information about a lesson's objectives and environmental conditions, the ITS calculates an "ideal" voyage plan (a trajectory with speeds and headings at various positions), as well as a range of satisfactory ones based on

information about acceptable "solutions." It also estimates positions at which action is required, such as points in time where the rudder needs to be moved, etc.

As a lesson progresses, the ITS monitors the conning officer's information gathering (particularly gaze, measured by head tracking) and decision-making performance (especially orders to the helm, lee helm, and tug), along with the resulting ship performance (track, speed, heading, proximity to hazards, etc.). It compares these with the range of acceptable behaviors in the lesson to assess the quality of the conning officer's performance. Additionally the ITS is able to estimate the student's current proficiency in the shiphandling skills at each stage of the lesson.

When a conning officer's performance falls outside the acceptable range, the ITS makes a decision whether to intervene with tutoring based on factors including how far from acceptability the performance is, the student's estimated proficiency at pertinent skills, and imminence of serious hazards. If it decides to tutor, the ITS provides spoken comments or recommendations corresponding to the unacceptable performance, the learner's hypothesized deficiency, and any imminent hazards. For example, heading control is less intuitive without steerageway than with, so the ITS devotes particular resources to assessing the student's proficiency in this condition, providing pointers at opportune times when deficiencies are detected.

Typically, compromises must be reached regarding how often the ITS will voice tutoring advice. Master Mariners value the opportunity for students to learn from their mistakes, and in simulation there is no safety risk or financial cost to crashing. Therefore, SMEs generally want the ITS to provide only minimal coaching, allowing students to experience results of their actions. However, the advantages of real-time tutoring are well documented, and calling the student's attention to important factors in a complex situation can be valuable. Striking the right balance of how much and what kind of tutoring advice to provide under which circumstances requires continual consultation with the Master Mariners in the development process. In a related concern, the content of these tutoring utterances are typically short and to the point, as students sometimes get frustrated when the tutor speaks for too long (Di Eugenio et al. 2008).

Overall, the process of developing an ITS for a particular shiphandling evolution requires multiple rounds of data collection and prototype testing. Initial data collection occurs from ship handlers at a variety of skill levels, and that feeds a task analysis that is vetted by SMEs. Once a prototype ITS is created, it is tested with Master Mariners to refine the content and frequency of ITS utterances. Once a prototype has been refined to the point where SMEs are satisfied, an assessment is conducted to examine the efficacy of an ITS on students.

EVALUATING ITS EFFECTIVENESS

The goal of an ITS is to augment instructors with technology that monitors student performance and provides automated feedback. Therefore, the primary test of ITS effectiveness is whether there is any difference in student learning when taught by a instructor or by an ITS. For COVE ITS, an evaluation of effectiveness was completed with actual students in the Advanced Shiphandling and Tactics (ASAT) course at the Surface Warfare Officers School (SWOS). Students used COVE to learn to carry out the Mooring to a Pier (or "Mooring") shiphandling evolution, in which they conn a ship through a channel, across a basin, and then land aside a pier.

The evaluation of effectiveness of COVE ITS for Getting Underway from a Pier was carried out in 2011 with 58 students from SWOS (Peters et al. 2011). Results showed no statistical difference in student performance based on whether they were tutored by an instructor or by the ITS. Getting Underway is a simpler task for students, whereas mooring is used as a graduation exercise for ASAT students. So examining ITS effectiveness in this more difficult scenario is a more stringent evaluation of ITS tutoring ability.

The general design of this effectiveness study gave students two runs in the Mooring scenario. The first run was tutored (either by an instructor or the ITS), where students were given advice and feedback as they performed the scenario ("tutored run"). Then, students got a second run where they were asked to perform the same scenario, but without any tutoring ("test run"). The primary hypothesis is that, during the unassisted test run, students tutored by COVE ITS will not perform significantly differently than students who were tutored by an instructor. If this turns out to be the case, then the data will suggest that the instructor and the ITS were equally able to coach students and provide a similar level of training that carried over into the unassisted test run.

Methodology

Participants

Volunteers were recruited from among the DDG (Destroyer class ship) student population of several ASAT classes. Content in the ASAT class is at a fairly basic level for students, who have had at most one tour at sea. A total of 20 participants were recruited (13 males and 7 females). Students ranged in age from 23 to 31 with an average of 34.5 months in the Navy. The students were randomly assigned to the experimental group (ITS tutor) or control group (Human tutor) so that each group contained 10 participants.

Lab Setup

The Conning Officer Virtual Environment (COVE) laboratory at SWOS consists of 12 individual computer stations. Each computer station includes an instructor station and a student simulation area. The instructor station allows an instructor to load a shiphandling scenario and monitor student progress. This station also contains the ITS software, which is separate from COVE but can communicate via the COVE application programming interface (API). The ITS can gather data from COVE and provide tutoring advice over headphones based on this data.

Figure 2. COVE station setup.

The student's simulation area contains four screens of instrument indications plus a head-mounted visual display (Figure 2) that affords the student a 360-degree immersive view of the ship's surroundings. The student also has a communications radio available, which allows for voice commands to the computer. The computer responds to the voice commands as the helmsman and lee helmsman on a surface ship would in real life.

Procedure

At the beginning of the COVE session, students received a pre-briefing on how to perform the Mooring scenario from their instructor. After this briefing, students all performed the Mooring scenario for the first time (their tutored run). Students in the Human Tutor group received instruction and tutoring advice from their assigned instructor. Instructors have varying styles – some like to provide advice during the run, while others withheld comments until the end (or if the student crashed). Audio of the entire session was recorded for later analysis. Students in the ITS Tutor group only received guidance from the ITS (though an instructor would observe the run). This system provided recommendations, warnings, and prompts if the student was giving orders that were outside predetermined parameters as determined in the task analysis and prototype testing.

This first run (the tutored run) provided students with the ability to learn how to perform the mooring scenario and receive feedback on their performance either by a human instructor or the ITS. Additionally, instructors graded the first run using the Conning Officer Shiphandling Assessment (COSA, Lee [2011]), which is a subjective performance assessment where instructors rate students along a Likert-like scale based on shiphandling factors (e.g., situational awareness, decision making, and maneuver). COSA is the standard shiphandling grading form used by SWOS, so instructors were already familiar with how to grade students in this way.

After completing the tutored run, all students were debriefed by their instructor. Then, students were asked to perform another Mooring run (the test run). No tutoring – by instructor or ITS – was provided to the student. An instructor was asked to observe the run, provide a debrief after the scenario was complete, and grade the student with the COSA. Two surveys were provided to the students. A basic biographical questionnaire asked questions pertaining to student demographics, time in the military, time at sea, and video game experience. The second questionnaire was an ITS Usage questionnaire that asked students their opinions on the feedback provided by the ITS. This test run examines how well students learned to complete the exercise during the tutored run. In general, human instructors were only asked to provide tutoring to students during the first run for students in the Human Tutor condition. They were asked to remain quiet during the first run for students in the ITS Tutor condition and

during every student's second (test) run. Only one instructed run was performed because instructor availability at SWOS currently limits students to this much instruction.

Results

ITS Questionnaire

An ITS Usage questionnaire was administered to students after they completed their scenarios. The experimental group was asked six questions on a standard 1-7 Likert scale. Overall, students said they listened to the ITS' advice instead of tuning it out (n=10, M=4.2, SE=0.45), but they were not more likely to change their behavior in response to ITS comments (n=10, M=3.3, SE=0.54). Students felt that the ITS gave slightly too much feedback (n=10, M=3.8, SE=0.44). Finally, they did find the ITS more useful than not (n=10, M=3.8, SE=0.44) for the Mooring scenario.

Instructor/COSA Scores

COSA forms were filled out by observing instructors in both Runs 1 and 2 across Human and ITS Tutored groups. For the first run, subjective assessment from the instructors across both groups was not significantly different (M=78.0%, SE=0.06% for Human Tutor, M=77.8%, SE=0.07% for ITS Tutor, $t(18)$=.38, p=.71). A similar result was found for the students' second unassisted run (M=69.4%, SE=0.05% for Human Tutor, M=71.1%, SE=0.09% for ITS Tutor, $t(18)$=.21, p=.84). This COSA form is part of the assessment process at SWOS, so this instrument is a trusted measure of student performance. No statistical difference between scores in both runs indicates that students were able to perform the Mooring scenario similarly well, regardless of tutoring method. A total of 348 students (174 students in each group) are needed to achieve 80% power at two sided 5% significance level. With current student throughput rates at SWOS, this number would be difficult to achieve. Further, Cohen's effect size (d =.10) for this analysis, falls below Cohen's convention for small effect (d=.20), and indicates similar results between groups.

Outcome Metrics

There are several dimensions by which success can be considered for Mooring. When the exercise is completed, there are three benchmarks that should be met to indicate a successful mooring: the bow must touch down no greater than five degrees off of the pier heading (i.e., the bow is lined up with the pier), longitudinal speed must be no greater than 0.5 knots, and the ship must be no more than 30 feet laterally away from a target on the pier. Binary versions of these metrics (successful/failure) were averaged to create an overall score, and there was no difference in the second (test) run (M=62.50% for Human Tutor, M=60.71% for ITS Tutor, t(18)=0.16, p=0.88). Cohen's effect size fell below convention for small effect (d=.08 for overall success metrics).

Process Metrics

Delving into mooring performance data allows analysis beyond the outcome metrics. These process metrics reveal how students went about the mooring task, and this is where the ITS is primarily designed to tutor. Factors such as speed, heading, and tug use all contribute to various performance process metrics.

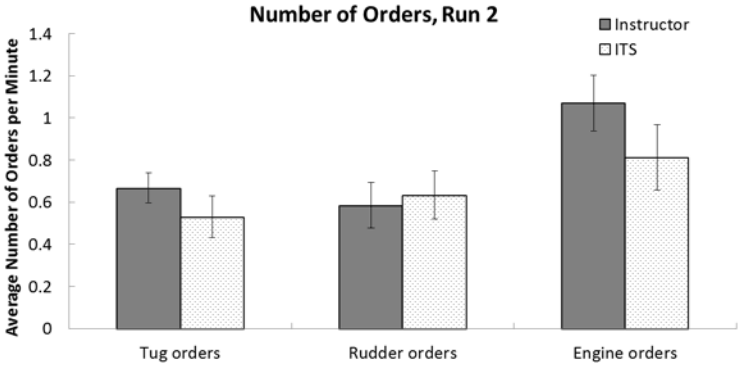

Figure 3. Average number of orders issued per minute.

One such area of these metrics involves the number of orders given by the student relating to the tug, rudder, or engines. Students with greater control over the ship, tug, and environmentals, should issue fewer orders. Those with less of a grasp of how these factors impact shiphandling will likely issue more orders. Figure 3 reveals the number of orders per minute given to the tug, rudder, or engine across conditions for Run 2. Run 1 was not analyzed for number of orders because this was the initial learning run where the student was provided tutoring advice during the run. There were no significant differences related to number of orders in Run 2 across conditions (t=1.13, p=0.27 for tug, t=0.32, p=0.76 for rudder, t=1.27, p=0.22 for engine).

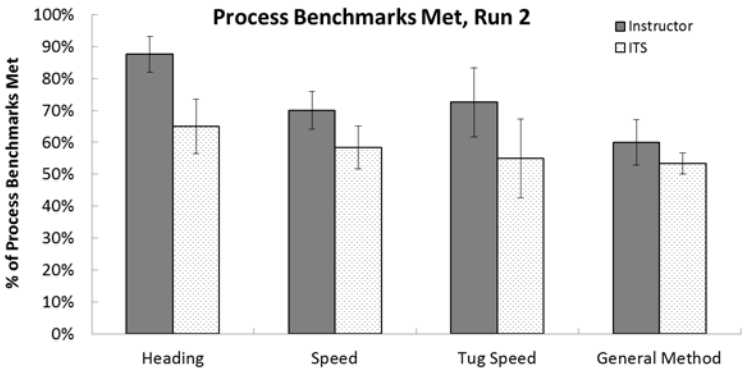

Figure 4. Four process benchmarks compared across conditions.

Along with the number of orders given, there are a variety of process benchmarks related to heading (e.g, leading and/or touching down stern first), speed (e.g., sternway while approaching the pier), tug speed (e.g., activating the tug with greater than 5 knots of speed), and general methods (e.g., ordering tug at full speed, ordering engines at a speed greater than two-thirds). Each of the benchmarks contains several criteria that were either met or not, and these were scored so that meeting these benchmarks led to a 100% score. Figure 4 contains the data for each of these four process categories in Run 2. Analysis shows a significant difference for heading (t=2.21, p<0.05) and no difference between the other process metrics across Human and ITS Tutored conditions in Run 2 (t=1.23, p=0.21for speed, t=0.23, p=0.82 for tug speed, t=1.13, p=0.27 for general method). Finally, Cohen's effect size for most process metrics (all except heading) fell below convention for small effect size indicating similar results between groups.

Tutoring Metrics

For every run for each student (regardless of whether the student was in the Human Tutor or ITS Tutor condition), the ITS passively recorded scenario data and which tutoring utterances were (or would have been) given to the student based on their real-time shiphandling performance. Therefore, the number of utterances that would have been given (because a performance tripwire was triggered) can be compared by students in both conditions for the second (test) run. Examining this data (Figure 5) shows no difference in the number of ITS utterances that would have been made to students across both conditions (M=14.8 for Human Tutor, M=14.8 for ITS Tutor, t(18)=0.00, p=1.00). This suggests that student performance in the second (test) did not differ based on the source of their tutoring in their first (tutored) run.

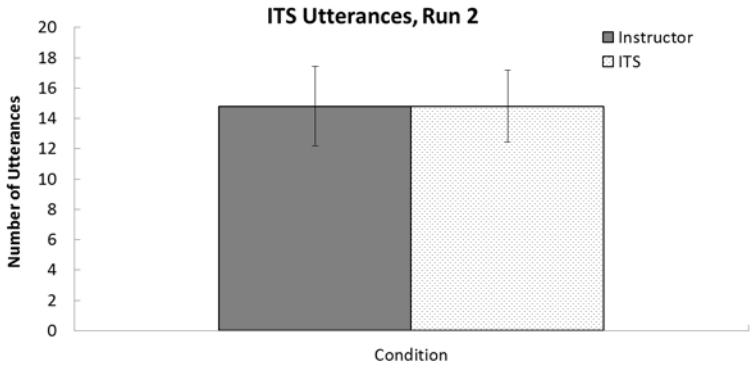

Figure 5. Number of ITS utterances.

DISCUSSION

Developing an ITS requires data collection for the task to be tutored from a variety of skills levels. One of the keys of an ITS is to understand acceptable performance parameters, so understanding merely how a novice or expert accomplishes the task does not provide a complete picture of acceptable performance. Additionally, as dynamic problems such as shiphandling can unfold in nearly infinite ways, more data is useful to understand what kinds of solutions are acceptable.

ITS development also requires constant discussion and compromise with subject matter experts about how much feedback to give and how to provide feedback that provides maximal input in the shortest amount of time. Some students felt that the refined tutoring advice in the Mooring ITS was still too long, so a constant refinement process is needed to try and provide the ideal amount of feedback to be helpful without being cumbersome. Additionally, certain utterances that satisfy the subject matter experts might be viewed differently by course instructors and/or students, and meeting the needs of all groups can be difficult.

Data from the evaluation experiment indicate that there are no significant differences between students who were tutored by a human instructor and students who were tutored by the ITS. This is especially clear in the second, unassisted, test run. This pattern holds across instructor observations via COSA scores, outcome metrics of overall mooring evolution success, process metrics of the methods students use to accomplish the task, and the utterances that the ITS recorded/spoke to give students advice. While one cannot prove the null hypothesis (that student performance is similar between human tutored and ITS tutored conditions), data across a variety of metrics suggests no difference. These results replicate the previous COVE ITS effectiveness study (Peters et al. 2011) in a more stringent test environment. That study found no performance difference between instructor and ITS tutoring amongst 58 students learning the Getting Underway from a Pier shiphandling evolution, which is a simpler task than the Mooring evolution examined here. From subjective to objective measures and outcome to process metrics, data collected from twenty students suggests that an ITS is capable of tutoring students as well as a human instructor.

From these results, we conclude that an ITS can be used to effectively tutor students. We do not suggest that an ITS can completely replace instructors, but it can valuably augment human instruction. By allowing the ITS to focus on students' detailed process metrics (e.g., speed, course, etc.), instructors can focus on bigger-picture issues such as shiphandling strategies and intentions.

Future development of the ITS includes understanding and implementing ITS for other shiphandling scenarios, such as the difficult evolution of Underway Replenishment, where a ship connects to an oil tanker for refueling in the open ocean. Additionally, future work will examine how many students a human instructor can tutor at a single time with the aid of an ITS monitoring each student - ideally, increasing the student-to-teacher ratio will save on costs and increase student throughput. Overall, an ITS for dynamic problem solving requires real-time student monitoring but has great potential at improving learning efficacy and efficiency. Another area of future work involves the interfacing of the ITS with additional technologies to support learning, such as the automated assessment engine (Koenig 2014). An ITS can mitigate limitations of instructor availability by enabling students to cost-effectively receive more instruction in and experience of shiphandling, thereby increasing their proficiency.

ACKNOWLEDGEMENTS

The authors thank the students, instructors, and staff of the Surface Warfare Officers School (SWOS) in Newport, RI for their assistance, time, and support in developing and testing the ITS. We would also like to thank Dr. Ray Perez, Program Officer at the Office of Naval Research (ONR) for support, advice, and oversight of the COVE ITS program. This project was undertaken with ONR funding to Stanford University and the Naval Undersea Warfare Center, Division Newport. We thank CSC for providing access to its COVE/VShip simulator for this research.

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*, Vol. 13(6), 4-16.

Di Eugenio, B., Fossati, D., Haller, S., Yu, D., & Glass, M. (2008). Be brief, and they shall learn: Generating concise language feedback for a computer tutor. *International Journal of Artificial Intelligence in Education,* Vol. 18(4), 317-345.

Gonzalez, C. (2012). Training decisions from experience with decision-making games. In Durlach, P.J., & Lesgold, A.M. (eds.) *Adaptive Technologies for Training and Education*, 167-178.

Graesser, A.C., Chipman, P., Haynes, B.C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education*, 48, 612-618.

Gredler, M. E. (2004). Games and simulations and their relationships to learning. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (2nd ed., 571-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, J.. (2011) *COSA [Conning Officer Shiphandling Assessment] at SWOS [Surface Warfare Officers School]*. Presentation given at the 2011 Human Systems Integration Symposium, Vienna, VA.

Koedinger, K., Anderson, J., Hadley, W., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30-43.

Koenig, A. (2014). Automated assessment methodology for games & simulations. To be presented at the Second Annual Mobility and Modern Web Conference, September 2014, at the University of California, Los Angeles.

Martens, A., & Himmelspach, J. (2005). Combining intelligent tutoring and simulation systems. *In Proceedings of the 2005 Westen Simulation Multiconference. SIMCHI 2005, International Conference on Human-Computer Interface Advances for Modeling and Simulation.*

Murray, W. R. (2006). Intelligent tutoring systems for commercial games: The virtual combat training center tutor and simulation. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 66-71

Peters, S., Bratt, E., & Kirschenbaum, S. (2011). Automated support for learning in simulation: Intelligent tutoring of shiphandling. *Proceedings of the 2011 Interservice/Industry Training, Simulation, and Education Conference*, Orlando, FL.

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