SANDIA REPORT

SAND2011-0166
Unlimited Release
January 2011

Real-time Individualized Training Vectors For Experiential Learning

Elaine M. Raybourn, Nathan Fabian, Matthew R. Glickman, Eilish Tucker, Matt Willis

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2011-0166
Unlimited Release
January 2011

Real-time Individualized Training Vectors For Experiential Learning

Elaine M. Raybourn, Nathan Fabian, Matthew R. Glickman, Eilish Tucker, Matt Willis
Cognitive Systems, Scalable Analysis and Visualization
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS1188

ABSTRACT

Military training utilizing serious games or virtual worlds potentially generate data that can be mined to better understand how trainees' learn in experiential exercises. Few data mining approaches for deployed military training games exist. Opportunities exist to collect and analyze these data, as well as to construct a full-history learner model. Outcomes discussed in the present document include results from a quasi-experimental research study on military game-based experiential learning, the deployment of an online game for training evidence collection, and results from a proof-of-concept pilot study on the development of individualized training vectors. This Lab Directed Research & Development (LDRD) project leveraged products within projects, such as Titan (Network Grand Challenge), Real-Time Feedback and Evaluation System, (America's Army Adaptive Thinking & Leadership, DARWARS Ambush! NK), and Dynamic Bayesian Networks to investigate whether machine learning capabilities could perform real-time, in-game similarity vectors of learner performance, toward adaptation of content delivery, and quantitative measurement of experiential learning.

ACKNOWLEDGMENTS

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. The authors would like to thank Karn Gustafson for assisting with the quasi-experimental study and Warren Davis for participating in discussions on clustering and temporal segmentation.

CONTENTS

FIGURES

TABLES

NOMENCLATURE

AAR	After Action Review
DBN	Dynamic Bayesian Networks
LDRD	Lab Directed Research & Development
LSA	Latent Semantic Analysis
DOE	Department of Energy
Sandia	Sandia National Laboratories
STTR	Stability, Security, Transition, and Reconstruction

1. INTRODUCTION

Over the course of the past several years the training and education community has begun to see more studies identifying the characteristics that constitute game effectiveness (Belanich, et. al., 2004; Orvis, et. al., 2006; Beal, 2006; Rowan & Brown, 2008; Raybourn, 2009; Raybourn et. al., 2010; Brown, 2010). These contributions address the question, "What can be learned about the use of games in training?" For example, Rowan & Brown (2008) indicate that serious games, when executed properly, can provide an effective and efficient means of blended training. The cognitive and affective learning possible in a game-based experiential environment is valuable for both individual and collective training. Beal (2006) learned that games are most effective in training when focused on specific training objectives and when facilitated by experienced instructors as opposed to used as a stand-alone tool. Rowan & Brown (2008) also found that serious games are an effective means to address tactical training requirements. Belanich and others (2004) learned that certain design and interface features of games enhance trainee motivation. In follow-on work examining the use of a first-person shooter, Orvis and others (2006) learned that characteristics such as prior videogame experience, goal orientation, and self-efficacy can also impact motivation. The findings from these studies address primarily kinetic game-based training missions.

The past several years have also seen the growth of non-lethal (a.k.a. non-kinetic) cross-cultural engagement training. While the United States military is adept at performing kinetic operations, gaps have been identified in home station training in the area of cross-cultural, non-kinetic engagements (Wong, 2004). Stability, Security, Transition, and Reconstruction operations (STTR) require non-lethal, or non-kinetic[1] competencies to succeed such as languages, regional and technical expertise, intercultural communication, interpersonal skills, and adaptive thinking.

Notable efforts are made throughout the training and education development community to prepare troops with the non-kinetic skills needed upon deployment. Command training centers and schoolhouses may provide live-action, constructive or virtual simulation, and/or game-based training exercises for rehearsing kinetic/non-kinetic missions in a more blended approach. There are also a number of game-based training applications aimed at learning languages, leadership, decision-making, negotiation, team-building, communication, and cultural awareness ranging from web-based advanced distributed learning to interactive video vignettes to single-player and multi-player commercial or government game-based training solutions. These serious games and related applications, although not discussed here, contribute to the resources available for home station STTR training.

Military training utilizing serious games or virtual worlds potentially generate data that can be mined to better understand how trainees' learn in experiential exercises. Few data mining approaches for deployed military training games exist. We sought to investigate whether we could develop individualized training vectors for trainees who participate in game-based training or training administered in a virtual world. Individualized training vectors are defined in the present report as individualized, real-time paths or trajectories that measure and predict future

[1] The US Army changed the term "non-kinetic" to "non-lethal" in 2007. In the present paper, non-kinetic is used to refer to civilian engagement techniques that do not involve the use of force.

performance by comparing/contrasting trainees' dynamic performance to those of experts or baseline standards. Opportunities exist to collect and analyze these data, as well as to construct a full-history learner model. Outcomes discussed in the present document include results from a quasi-experimental research study on military game-based experiential learning, the deployment of an online game for training evidence collection, and results from a proof-of-concept pilot study on the development of individualized training vectors.

This Lab Directed Research & Development (LDRD) project leveraged products within projects, such as the Titan Toolkit (Network Grand Challenge), Real-Time Feedback and Evaluation System, (Special Forces JFKSWCS Adaptive Thinking & Leadership, DARPA DARWARS Ambush), and Dynamic Bayesian Networks (STANLEY) to investigate whether machine learning capabilities could perform real-time, in-game similarity vectors of learner performance, adapt content delivery, and quantitatively measure experiential learning.

The following sections describe research conducted during the first year of the LDRD to investigate the efficacy of the use of multiple roles in ill-defined domains such as game-based non-kinetic military training and to determine whether it was feasible to use informatics and Latent Semantic Analysis (LSA) technologies to interpret game play data toward the development of an individualized training vector.

2. PROBLEM STATEMENT AND RATIONALE

While the body of research regarding the study of tactical training game effectiveness is growing as described in the above section, empirical study of non-lethal, or non-kinetic, game effectiveness is lacking. Eventually pushing past tactical, kinetic, game effectiveness studies our community will better understand best practices for how we can make learning itself more effective with games.

This research seeks to go *beyond game effectiveness* per se to understand how we can make *game-based learning more effective,* especially learning focused on training non-lethal, or non-kinetic engagement skills. Making learning more effective is an opportunity for out-of-the-box thinking (Raybourn, 2007a). For example, in most cases the training community designs and develops game-based training solutions that leverage the dominant paradigm of how users conventionally engage in entertainment games. In particular practice environments or trainers are often make the assumption that in order to engage trainees cognitively, experientially, and affectively we need to keep them busy in the game by "doing." However, what would happen if we also kept them "busy" by observing and thinking in a game?

In two cases, such as the DARWARS Ambush! NK and Adaptive Thinking & Leadership, multi-player game-based training systems have been deployed which provide trainees with opportunities to play multiple training roles designed to exercise intercultural communication, adaptive thinking, and metacognitive skills (Raybourn 2006, 2007a,c; 2009). Real-time feedback supported the use multiple roles in negotiation and cultural awareness non-kinetic training for Special Forces (Raybourn et. al., 2005) however more research is needed.

2.1. Research Objectives

The purpose of this research is to investigate the utility of the inclusion of multiple roles focused on both "doing" and "metacognitive thinking" in multiplayer game-based training. In particular, multi-player, first person perspective games are fast paced, and often task-oriented. While it might seem counter-intuitive at first to create multi-player training roles for real-time metacognitive skill development in which players observe and evaluate the performance of others, we believe this approach may offer out-of-the-box solutions for training metacognitive agility and adaptive thinking. Therefore in the spirit of desiring to make game-based learning more effective, we studied the inclusion of multiple roles in a non-kinetic game-based training mission. The present paper presents our first analyses of an empirical study investigating multi-role experiential learning in the multi-player game-based training modules transitioned in 2007 to PEO-STRI, DARWARS Ambush! NK.

3. WHY MULTIPLE ROLES?

3.1. Experiential Learning Theory

The incorporation of multiple roles in our game-based training approach was inspired by Experiential Learning Theory, metacognition, and our previous work with the US Army John F. Kennedy Special Warfare Center and School (USAJFKSWCS). During an ethnographic investigation at the Special Warfare Center and School, The first author learned that role-playing, observing others model behavior, reflecting to analyze best practices, and providing constructive peer feedback were key elements to the way Special Forces trained across their education curriculum (Raybourn, 2009). In addition she noted that each of these elements could be part of the same game-based training system, but should be trained differently—which required *thinking differently* about the design of conventional first-person, game-based training (Raybourn, 2007a).

Experiential Learning Theory defines learning as "the process whereby knowledge is created through the transformation of experience" (Kolb, 1984; p. 41). According to Kolb knowledge results from the combination of both grasping and transforming experience. The constructs for creating knowledge include concrete experience and reflective observation for grasping experience, and abstract conceptualization, and active experimentation for transforming experience (Kolb et. al., 2000). Learning is characterized as a cycle of creative tension among these four learning modes. The cycle is expressed as such: concrete experiences form the basis for reflective observations. These observations form abstract concepts that provide a framework for new implications of actions that can be taken. These implications are then tested in active experimentation to guide the formation of new actions. Multiple roles were introduced into two multi-player game-based systems in order to provide trainees with different cognitive experiences at the same time and regarding the same training content so that they could better learn from each other during debriefings and after action reviews [AAR] (Raybourn, 2006, 2007a,c).

3.2. Metacognition

Metacognition has been defined a number of ways over the years. A good working definition of metacognition is higher order thinking that involves active control of one's learning process to include knowledge of persons, task, and strategy (Flavell, 1979; White et. al., 1999). Thus *metacognitive agility* is defined in the present paper and has been defined by the author previously (Raybourn, 2007c) as possessing the ability to analyze the way one or others think, discern different tasks or problems requiring different types of cognitive strategies, and employ those strategies to enhance learning and performance. Knowledge is considered to be metacognitive in nature if it results in *strategic* use toward the accomplishment of a goal. For example, knowing one's strengths and weaknesses with respect to a given task and using this information strategically (through task analysis, planning, monitoring, evaluating, and reflecting) to meet a goal or improve performance is exercising executive or metacognitive skills (Veenman et. al., 2005).

Incorporating a role that exercises more strategic thinking, reflection, and self-regulation can provide trainees with a unique view of different sides of the same coin. According to Livingston (1996) "simply providing knowledge without experience or vice versa does not seem to be sufficient for the development of metacognitive control." Menaker et. al. (2006) take this notion further by arguing that experience alone is not enough to make an activity "cognitive." Therefore in developing training to exercise cognitive or metacognitive agility, designers should make the diverse cognitive processes explicit to trainees so they can utilize these skills again or in diverse settings. Multiple roles can help trainees better experience and identify their executive skills (metacognition).

Experiential Learning Theory and metacognition contributed to the design of multiple training roles which provide opportunities for 1) addressing the four experiential learning theory modes in multiplayer first-person game platforms 2) honing one's meta-level thinking about the strategies employed by performers in game-based training while providing constructive performance evaluation feedback to others. Honing metacognitive agility is also integral to becoming a competent intercultural communicator (Bennett, 1984; Raybourn 2009) which is a key capability of successful STTR operations and the basis for the Simulation Experience Design Method (Raybourn, 2006, 2007a) used by the first author to design multiple training roles.

4. EXPERIMENTAL ENVIRONMENT

DARWARS Ambush! NK [non-kinetic] (Raybourn et. al., 2008) was developed to provide DARWARS Ambush! (Roberts et. al., 2006) with non-kinetic mission modules. The modules were transitioned to PEO-STRI in 2007. DARWARS Ambush! NK consists of immersive multiplayer scenarios in a fictitious environment and builds on commercial computer game technology (Operation Flashpoint, developed by Bohemia Interactive Studios). The DARWARS Ambush! NK platform includes roles for an instructor, soldiers, local nationals, observer/evaluators, and non-player-characters (NPC). Role-play is centered on exercising non-kinetic stability operations competencies within two different scenarios. Much like the Adaptive Thinking & Leadership training game (Raybourn et. al., 2005) role-players use headsets with

microphones to communicate and interact with others during the game-based training. Reflective Observer/Evaluators provide real-time performance evaluations.

Real-time injects that influence the actions taken by role players in the scenario help the instructor create opportunities for adaptive thinking and demonstration of leadership skills as the situation dynamically changes (Raybourn et. al., 2005, 2008). The training design also includes a proprietary method of collecting real-time in-game assessment and feedback from observer controllers, subject matter experts, or peer learners in the role of observer/evaluators (Raybourn, 2006, 2007c). Adapting this approach to a multiplayer environment, one with multiple observers and multiple trainees, necessitated some changes to the DARWARS Ambush! environment.

DARWARS Ambush NK! missions include a socio-cultural human terrain overlay for the DARWARS Ambush! geographical map. A workshop was conducted at Ft. Lewis, WA with training developers in February 2007. The lessons learned and invaluable contributions of Ft. Lewis subject matter experts helped generate the DARWARS Ambush NK! training materials (Raybourn et. al., 2007b) designed to aid home station training developers in creating non-kinetic engagement missions for convoy and dismounted STTR operations training.

5. MULTIPLE ROLES USED IN STUDY

The Cordon and Knock mission module (Raybourn et. al., 2007b; Raybourn et. al., 2008) was used as the context for our empirical study. Portions of the multiple roles discussed below are excerpts from DARWARS Ambush! NK mission documentation and an unpublished paper that accompanied a poster presentation at the December 2008 Army Science Conference.

5.1. Player Role

Figure 1. Player Interface

A Cordon and Knock mission was designed to hone player's listening, communication, and problem solving in an intercultural setting (see Raybourn, 2006, 2007a for mission design model). A concluding task was comprised of one kinetic, novel situation dilemma that allowed the player to demonstrate leadership and creativity (Raybourn et. al., 2005). The player receives

an Operations Order to bring a local national (LN) from the village back to the FOB for questioning. As a trainee the goal of this mission is to successfully conduct tactical questioning in the intercultural setting. If successful, the trainee learns that the local national's cousin is an Imam who is cooperating with US Government. The questioning allows the trainee to practice aspects of intercultural communication such as cultural awareness, language, listening, cultural norms, and some nonverbal communication. The trainees' objective is to negotiate with the local national to return to the Forward Operating Base (FOB) for questioning willfully and voluntarily. Other tasks executed by the trainee include investigating a nearby marketplace where questionable equipment is for sale (e.g., weapons, night vision goggles) and communicating with merchants (non-player character text dialogue) who can provide additional information as to the whereabouts of the local national to be interviewed. A player view (3rd person perspective) is shown in Figure 1. Players interact with the interface similar to many commercial first-person perspective games including Operation Flashpoint and more currently VBS-1 and 2.

5.2. Reflective Observer/Evaluator Role

An approach to training metacogntive agility and adaptive thinking is to give trainees concrete practice, reflective observation, abstract conceptualization and active experimentation with evaluating their own actions and those of others. Non-kinetic engagement training such as rapport building, negotiation, questioning, interviewing, etc. is aimed at improving communication and cultural awareness skills. A goal of the Reflective Observer/Evaluator Role is to provide trainees the opportunity to reflect on communication events, speech acts, and verbal strategies that are enacted in player roles (Raybourn 2009).

The creation of a new trainee role, the role of the peer Reflective Observer/Evaluator, allows both trainees and experts such as observer controllers to provide real-time, in-game performance & feedback evaluation (Raybourn, 2007c). Up to 20 trainees may perform the Reflective Observer/Evaluator role in a given DARWARS Ambush! NK training session. This role enables more trainees to actively reflect on and evaluate the effects of one's actions and the ways they might have responded or acted if they were in the player's shoes. This approach places evaluators in the training event, and gives them the ability to assess the player's performance and comment on events as they unfold.

The Reflective Observer/Evaluator interface shown in the following screenshot allows users to track the activities of any character in the mission from that character's point of view with the expressed purpose of evaluating performance in real-time. Evaluations are initiated by the instructor who selects the performance criteria from a competency drop-down list. The instructor sends requests for evaluation to all the Observer/Evaluators. Presently there are 10 general non-kinetic engagement competencies from which to choose including several items related to cultural awareness, leadership, communication, and adaptability.

Figure 2. Reflective Observer/Evaluator interface

The interface above shows the character centered in the scene from a third-person perspective. As that character navigates through the virtual world, the display will update automatically. Two controls at the top of the screen adjust the relative viewpoint: the **Pan** slider slews the view to the left or right of the character's own body's orientation; the **Zoom** slider moves the viewpoint closer or farther from the character (Raybourn et. al., 2007b, 2008).

Observer/Evaluators attach themselves to any character or team in the mission using the drop-down list at the upper left of the screen. They can also switch between third- or first-person perspectives using the adjacent drop-down list button.

6. METHOD

Recall that the purpose of this research is to investigate the utility of the inclusion of multiple roles focused on both "doing" (enacting) and "metacognitive thinking" (observing and evaluating) in multiplayer game-based training. The previous sections describe roles that comprise an approach to training designed to exercise the skills needed for adaptive thinking, communication (Raybourn et. al., 2005), intercultural competence, self-awareness, reflection (Raybourn, 2007a), and related non-kinetic skills. A quasi-experimental design was used to measure the effects of participation in a multi-player, multi-role game-based training mission. All attempts were made to replicate training as it might occur at a schoolhouse. The following research questions were addressed with a quantitative, qualitative, and pilot validation study:

RQ1: Do participants, regardless of role (whether in player or observation/evaluation roles), report change with respect to their learning?

RQ2: Are there significant differences among groups participating in different roles in non-kinetic engagement training, especially when one role requires more active participation that the other?

RQ3: What does the visualization of trainee performance data coming from unscripted game play transcription offer toward further understanding of the empirical results?

6.1. Procedure

Participants arrived in groups of two and completed demographic questionnaires and a pre-test questionnaire designed to baseline their learning expectations. Following the completion of questionnaires both participants received training on the interface commands and maneuvering characters in the game. They were allowed time to familiarize themselves with the game controls and interface. Next, participants were trained on how to operate the Reflective Observer/Evaluator interface and taught how to evaluate the player's performance (e.g. what to look for, definitions of terms and corresponding behaviors, etc). Participants then watched a video of a 10 minute power point briefing administered by a member of the U.S. Army (in uniform) on the mission and the player's mission objectives. For example, the briefing consisted of the local geographical area, culture and society, the key items to be on the lookout for, and the operations order to bring a local host national from a fictitious city back to the FOB for questioning. Participants were also told that the US Army unit had been working on relationship building with a key individual of the area. After the video briefing, participants self-selected the role of player or Reflective Observer/Evaluator. The training mission lasted approximately 25 minutes. During the mission players were in the role of a commander or the Observer/Evaluator. Commanders had one squad member (played by a confederate of the experiment) assigned to them. Observer/Evaluators listened to the communications and observed/evaluated the gameplay. They provided real-time evaluations on the player's performance in key moments. This feedback was logged and not made visible to the player so to not alter events. After the training mission had concluded both participants completed posttests and switched roles. All attempts were made to replicate real-world training events as the authors have witnessed/conducted at military battle command simulation centers and schoolhouses.

The present study investigates the first half of game participation, that is, participation in one role during the first 25 minutes of the training exercise. We also augmented the quantitative study with further investigation such as stochastic player performance analysis using latent semantic analyses and graph visualizations. Our long-term approach consists of constructing an algorithmic mapping from data to performance evaluation that models performance evaluations provided by a human rater/coder. We provide more detail on this procedure in the remaining sections.

6.2. Research Participants

Eighty-five members of Sandia National Laboratories volunteered for the present study. Most of the 85 participants were novices with little to no military experience, only 13% reported ever being or currently a member of the US Armed Forces. They ranged from ages 18 – 64. Only 12 females participated in the study. Sixty-two percent reported being European American. Thirty-three percent had master's degrees and 11% had doctoral degrees. Ninety-nine percent reported having no computer game-based training, although 64% had played single-player games, and 26% reported playing multi-player games.

In addition to the quantitative study of the 85 participants, 78 transcripts from the same volunteer pool of 85 members of Sandia National Laboratories were analyzed for the present study. Seven transcripts were omitted from the study because they were incomplete, inaudible, etc. In the following section we provide the results and discussion of the quantitative data analysis.

7. QUANTITATIVE STUDY

The authors tested the hypotheses described below with the quantitative and qualitative analysis. The first hypothesis tested self-reported change in learning for trainees in one of the two roles (either player or observer/evaluator) and was measured as the difference between pre- and posttest scores as calculated by a paired sample t-test. The decision rule for paired sample t-test was five percent significance. The second hypothesis was tested by mean differences of independent sample t-tests calculated on the difference scores of pre- and posttest questionnaires completed by participants in each of the two conditions: Group 1 Player and Group 2 Observer/Evaluator. The decision rule for independent sample t-test was five percent significance.

The first null hypothesis stated that *participants, regardless of role (whether player or Observer/Evaluator), would report no change regarding their learning after participation in the non-kinetic training mission*. Paired sample t-tests on the pre- and posttest means indicate that the posttest mean statistically significantly increased after participation in the non-kinetic training mission as compared to the time of the pretest. The null hypothesis was rejected. Results suggested that players reported learning about their communication by interacting in the training mission ($t= 2.8$, [df = 36], $p<.009$), that the training was a good use of their time ($t= 3.4$, [df = 36], $p<.002$), that they learned something about cultural awareness by interacting with the mission ($t= 4.0$, [df = 36], $p<.000$), that the training mission was an engaging way to practice communication skills ($t= 6.2$, [df = 35], $p<.000$), that that the skills learned during the training mission are helpful in solving problems and making decisions($t= 3.5$, [df = 36], $p<.001$), and that the training mission was difficult ($t= 2.7$, [df = 34], $p<.012$).

Reflective Observer/Evaluators reported believing the training mission was an engaging way to practice communication skills ($t= 2.6$, [df = 41], $p<.014$), that they learned something about cultural awareness by interacting with the mission ($t= 2.4$, [df = 41], $p<.023$), and that that the skills learned during the training mission are helpful in solving problems and making decisions ($t= 2.4$, [df = 41], $p<.02$).

The second null hypothesis stated that *there would be no significant differences among the two groups participating in different roles of non-kinetic engagement training*. The second hypothesis was tested by mean differences of independent sample t-tests calculated on the difference scores of pre- and posttest questionnaires completed by participants in each of the two conditions: Group 1 Player and Group 2 Observer/Evaluator. The null hypothesis was rejected. Participants in Group 1 Player reported learning about their communication by interacting in the training mission ($t= 2.9$, [df = 83], $p<.005$), that the training was a good use of their time ($t= 2.3$, [df = 83], $p<.022$), that the training mission was difficult ($t= 2.2$, [df = 83], $p<.03$), and that they learned more about their strengths and weaknesses by participating than they would have if they did not participate ($t= 1.2$, [df = 83], $p<.05$, equal variances not assumed). In other words, players reported learning tasks that were specifically designed to address communication and self-awareness.

However, on other tasks that were specifically designed with Observer/Evaluators in mind there were no significant differences among the two groups even though both groups reported

learning. That is, when taken independently both groups reported statistically significant learning, but when the means of the two groups were compared, they were not statistically significantly different. For example, there was no statistical difference on previously salient items such as believing the training mission was an engaging way to practice communication skills (t= .82, [df = 83], p>.4), learning something about cultural awareness from interacting with the training mission (t= .81, [df = 83], p.>.4), and that the skills they learned from interacting with the mission would be helpful in solving problems and making decisions (t= .12, [df = 83], p>.9).

7.1. Discussion

The results indicate that contrary to popular expectations participants in both roles (Player and Reflective Observer/Evaluator) reported statistically significant learning. That is to say, trainees don't have to "play" as a protagonist in order to learn in game-based non-kinetic training. One may also observe, analyze, and evaluate another's performance in game-based non-kinetic training and still report engagement and learning. The Observer/Evaluator role was designed to provide an opportunity for real-time reflection and meta-cognitive learning.

Both Group 1 Player and Group 2 Observer/Evaluator exhibited significant change in learning after participation in the training mission from the time the pre-test was taken. Those participating in Group 2 Observer/Evaluator reported believing that the mission was an engaging way to practice communication, that they had learned about cultural awareness, and that the skills they learned from the mission were useful for problem solving and decision making. . While Group 1 Players reported significant difference on certain items closely associated with performing a communication task (which was the purpose of the mission for the player), it is important to note that there were a number of commonalities between the reported learning experiences for Group 1 Player and Group 2 Observer/Evaluators. Both groups learned, and in the end, there were only a few items that players reported learning more than Observer/Evaluators.

The Reflective Observer/Evaluator role focuses trainees on providing performance evaluations in real-time for behaviors such as cultural awareness, communication, leadership, and adaptability. It is possible that observation/evaluation is more complex activity than originally thought—therefore requiring concrete experience with the training topic. Recall a principle from Experiential Learning Theory: concrete experiences form the basis for reflective observations (Kolb et. al., 2000). It may be possible that trainees need to have a direct concrete experience with the training topic before they can identify salient behaviors in themselves and others for evaluating performance. Evaluation can be a rather abstract task that may require "graduate level" understanding of the training objectives. After all, in order to evaluate others fairly one must understand the phenomenon very well and be able to articulate the rationale for evaluations. Players' tasks may have been more straightforward, and the reward perhaps more immediate. Follow-on research is needed to better address these issues.

What do the lessons learned from this empirical study mean for the future use and design of game technology for training? First, the role for Reflective Observer/Evaluators was much more engaging than most would imagine. In a non-kinetic engagement mission where small

groups or key individuals (commanders) practice the act of intercultural communication or negotiation, introducing a role for Reflective Observer/Evaluators can be a force multiplier in developing a shared understanding of collective skills practiced in game-based training. By drawing every trainee into the same mission they learn from each other (Raybourn, 2007a). Ultimately it is learning from each other that we hope to engender with this training.

Second, the results of this study help us see that we have not yet fully explored what it means to learn and train with multiple roles in games. Does one always have to "do," to learn in games? Should the tasks always be concrete and procedural? The results would suggest that one can also learn in roles that may be more abstract and conceptual honing different ways of "thinking" and metacognition (Kolb et. al., 2000) Games can potentially provide different roles which can also be played more than once or in a different order to potentially enhance experiential learning in new ways. Further research is warranted.

8. QUALITATIVE STUDY

8.1 Data Analysis Toolkit: Titan

The Titan Toolkit (Wylie & Baumes, 2009) addresses a growing trend that involves merging scientific visualization, which is physics-based and has a real-world representation, with information visualization, which is notionally abstract and has no well-defined representation. In the present paper, we used the algorithms developed in Titan to visualize and interpret military training in the form of game-play activity data combined with abstract qualitative meta-information (e.g., trainee performance or individual learning processes) in the game space. In doing so, we established a foundation on which to build a generalized utility for analyzing games and bridging from concrete game activities to qualitative meta-level performance and learner modeling.

8.2 Data Analysis Procedure

The initial step involved transcription of all the communication utterances in the game-based training mission. Two human rater/coders performed a content analysis rating the trainees on a scale of 1 to 5 as to how well they performed the task (see Table 1 for categories corresponding with 1 – 5 ratings). The videos of game play time-stamped events were also transcribed. The videos used in this study were from the Reflective Observer/Evaluator point of view so that comments from trainees in this role could be coded. The videos of game-play data were analyzed for type of speech (statement, question, answer, or comment.), category of speech (unscripted spoken or scripted text dialogue), and other information (i.e. vocal tone if any). In the category of "type of speech" a comment was help provided out of character by a research confederate or a trainee asking a question that did not pertain to the mission, such as "How do I use my gun? The videos were coded a second time by two different coders who were familiar with the training mission and had been trained to recognize core competencies of non-kinetic engagements in the game-based scenario. The inter-coder reliability score (Pearson's correlation coefficient, r =.79) was strong. The coders focused on salient events and noted the Reflective Observer/Evaluator ratings and comments. Examples of these ratings on a Likert-type scale (1= strongly disagree to 5= strongly agree) include items such as "Able to speak targeted phrases in foreign language" and "Effectively communicates with team." In some cases Reflective Observer/Evaluators also augmented rankings with open-ended statements. The human coders used transcripts, Reflective Observer/Evaluator ratings, comments, and trainee performance categories to rate the skill level of the trainees using a Likert-type scale (1 = unskilled player to 5 = skilled player). The final rating depended on trainees' ratings in the categories of Communication, Observation, Cultural Awareness, and Mission as well as trainees' overall effectiveness in gaining information, their use of communication strategies, and lastly the Reflective Observer/Evaluator's ratings. Table 1 illustrates the coding system used.

Table 1. Human Coding Process

	Ratings				
	1	2	3	4	5
Communication	*Poor to little communication with Confederate *Poor to little communication with Host National *Doesn't Attempt to use Arabic greetings	*Poor communication with Confederate *Poor communication with Host National	*Okay communication with Host National *Needs prompting from Confederate *Doesn't utilize Confederate *Attempts Arabic phrases	*Good communication with Host National *Utilizes Confederate *Engages Confederate in mission *Attempts Arabic Phrases	*Uses Arabic Phrases *Good communication/utilizes confederate *Great communication with Host National
Cultural Awareness	*Poor to no Cultural Awareness	*Little Cultural awareness	*Attempts cultural awareness	*Demonstration of Cultural awareness	*Cultural Awareness more prominent across behaviors and communication
Observation	*Not aware of surroundings	*A little aware of surroundings	*Observant, but might not spot crowd or BOLO	*Might remember Fayez/BOLO *Aware of surroundings	*Concern for safety *Spots BOLO *Aware of crowd, binoculars or BOLO
Mission	*Narrow View of mission *Tries to arrest Host National *Forceful in getting Host National back to base *Forgets mission	*Tries to arrest Host National *Forceful		*Asks pointed questions *Concern for finishing mission non-kinetically	Asks pointed questions but is not forceful Concern for finishing mission non-kinetically

8.3 Human Coding Results

There were four main categories evaluated that make up the aggregate 1-5 rankings that players received: Communication, Cultural Awareness, Observation, and Mission. Communication had subcategories such as communication with the Confederate and Communication with the Host National, attempting to use the Arabic Language. Observation included noticing the BOLO, and watching/noticing the crowd. Cultural awareness included remembering conversations with vendors and utilizing information learned in the market, expressing concern for civilians, etc. Mission included how the trainees viewed their mission and if they remembered mission objectives. Of the 78 trainee transcriptions coded, 7 trainees were rated a 1 (as characterized by the categories in Table 1 pertaining to 1 – 5), 16 were rated as 2, 17 were rated 3, 29 were rated 4, and 9 were rated 5. Transcripts were omitted if they were unable to be coded due to the video not picking up on player conversations or a video not completely recording a session.

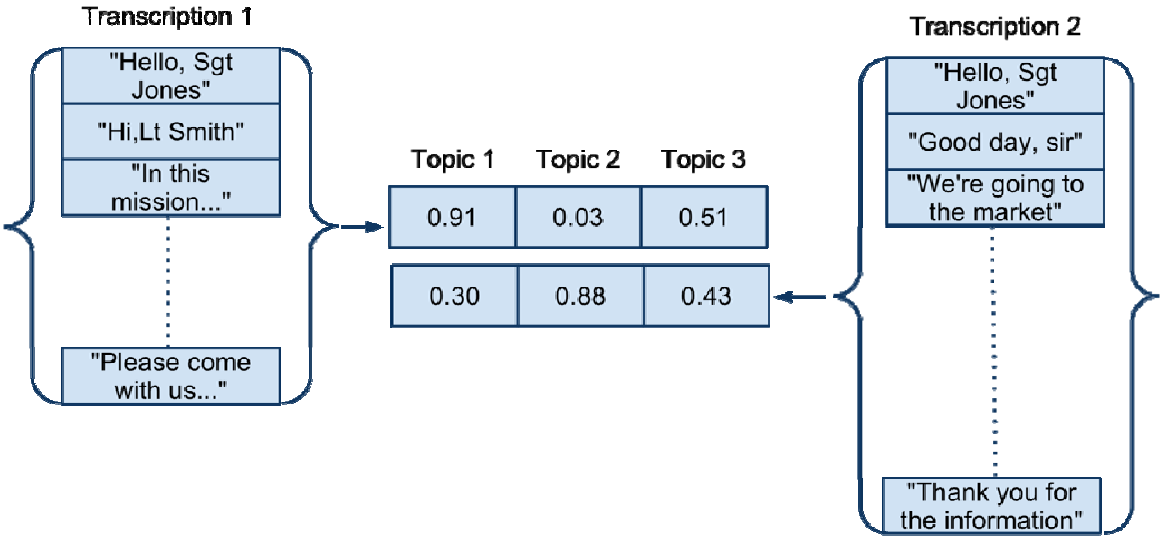


Figure 3. Transcription to Topic Space Diagram

8.4 Automated Coding Results & Discussion

Following Landauer et al. (Landauer et al, 2003) to automatically score written text, we then used Titan to cluster the trainee's individual terms from the transcriptions collectively into topics via Latent Semantic Analysis (LSA) to produce vectors of weights over transcriptions in a fixed dimension topic space, (see Figure 3). For example, each trainee's transcription becomes a position in this topic space which enables an ordinary distance comparison between trainees' texts in the space. Using these topic-space vectors, we then applied standard machine learning techniques (Duda et al, 2001) to establish a model that maps trainees to scores. With the constructed model, we can predict scores for new trainees.

LSA (Landauer et al., 1998) works by taking a per-document term-frequency count and turning this into a set of topic weights. Each term is weighted for relevance to a topic, and similarly, each document is weighted for relevance to the same set of topics. This approach produces two sets of weights—one for mapping terms to topics and one for mapping documents to topics. We use the document weights to interpret similarity between players. Topics are chosen as a parameter to the system, in this case we settled on 9 topics by experimentation.

In order to account for the temporal aspects of the game data (similar to the approach found in Brants et al, 2002) each transcription is broken into three distinct sub-documents (see Figure 4). We chose these in such a way as to break up the final human-to-human negotiation that occurs within the game into three equally sized segments. For instance, if one player spent 15 minutes negotiating the segments would each by 5 minutes long. If another player negotiated for 20 minutes, that player's segments would be 6.6 minutes long. We decided to break up this final dialogue in this way, based on the human rating process where the final dialog was the most important factor in the overall score. Finally after running each sub-document through LSA we concatenated the vectors back into one longer vector of dimension 27.

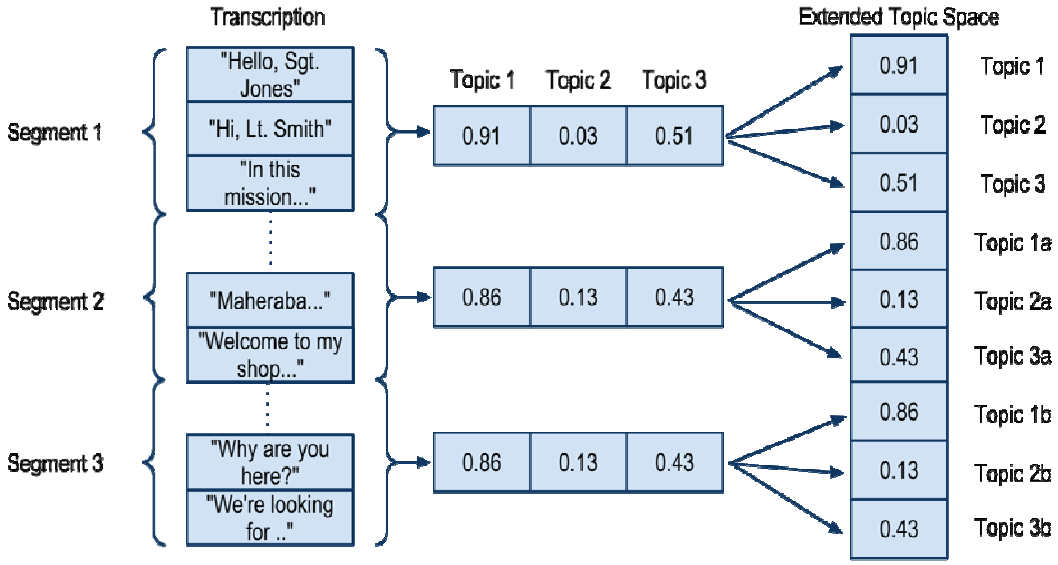


Figure 4. Splitting Transcriptions Temporally

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
ah	shot	tent	shop	target	uh	uh	uh	uh
uh	hello	shot	inch	medical	bodies	confirmed	bodies	americans
nasir	safe	green	tv	neutralized	lieutenant	injured	village	leave
um	shop	crates	32	permission	return	rocket	tribe	tribe
sergeant	shooting	injured	greetings	lock	medical	care	leave	strange
okay	rpg	medical	models	clear	touch	bodies	terrorists	weapon
saiful	buy	night	flat	confirmed	happened	appears	members	cousin
hello	greetings	bodies	panel	force	throwing	permission	ambulance	night
talk	islam	tribe	ahead	uh	rocks	commander	crowd	village
help	injured	strange	electronics	commander	team	medical	green	understand

Figure 5. Top terms in each topic

Figure 5 shows the top ten weighted terms assigned by LSA to each of the nine topics. Note that, these topics correspond well with some of the categories the human judge used in the rating scheme, for instance Topic 5 matches well with "communication with confederate." In Figure 6 the association of document segments to topics, thus there are three vertices that belong to a document.

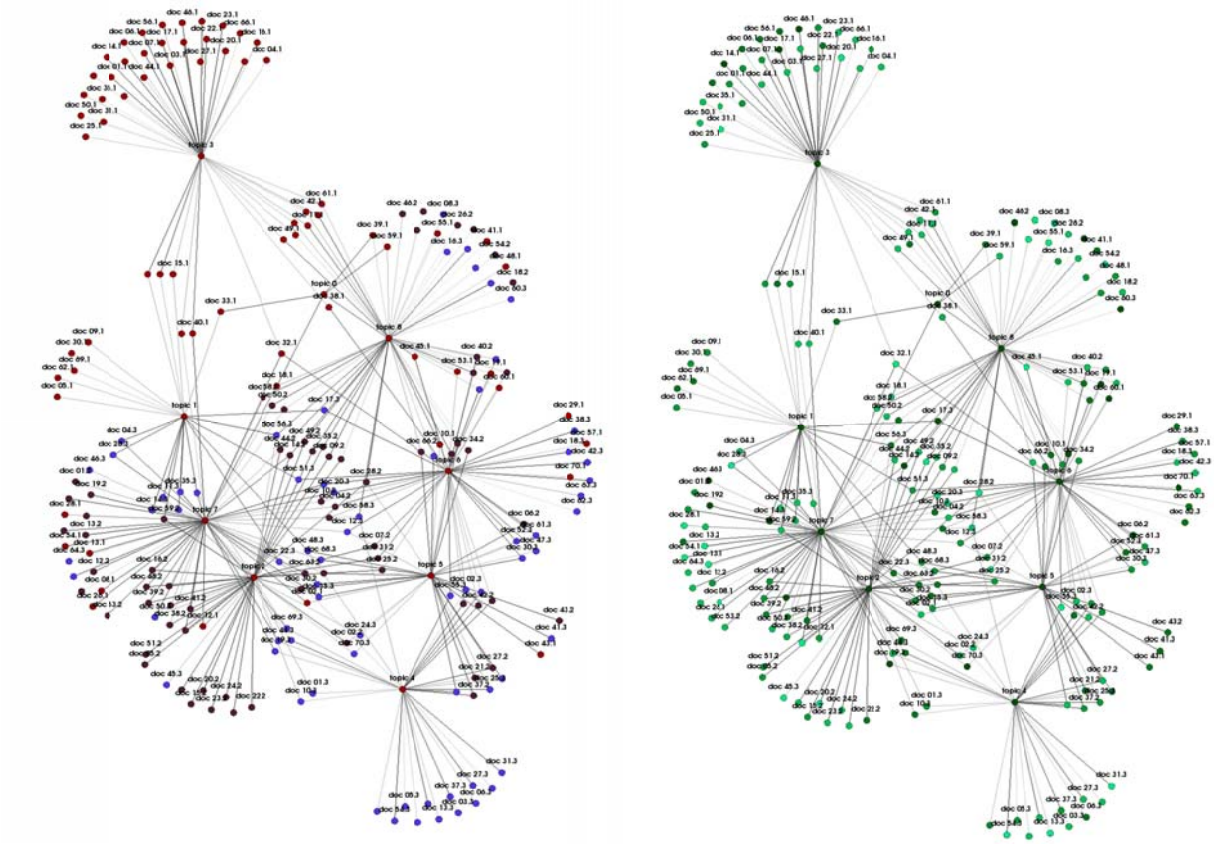


Figure 6. Topc topics in document segments

Left image vertices colored by segment. Right image vertices colored by human score.

Vertices are labeled with "doc x,y" where x is the document number and y is the segment number. In the left image we colored the vertices by the three segment numbers, from red to black to blue. In the right image, vertices are colored by the score (1 to 5, from dark green to bright green) given to the entire document. In Figure 6, there is no direct association between topics and rating. Although it is apparent from the graph on the left that certain topics are important early, such as topic 3, and some late, such as topic 4, there is no clearly defined trend for all the topics strictly in terms of segments.

Figure 7 depicts the average topic weight across all documents at each segment in time. This shows again that topic 3 is strong in the early but not later segments. There are some patterns recognizable from Figures 5 and6, but to complete the picture we elicited a more complicated mapping between topics and segments. To construct this mapping, we used a support vector machine (SVM) provided in the WEKA machine learning toolkit (Hall et al, 2009). This builds a classifier of our 27-dimensional topic-space vectors. We used the human rater's scores as train and test class labels.

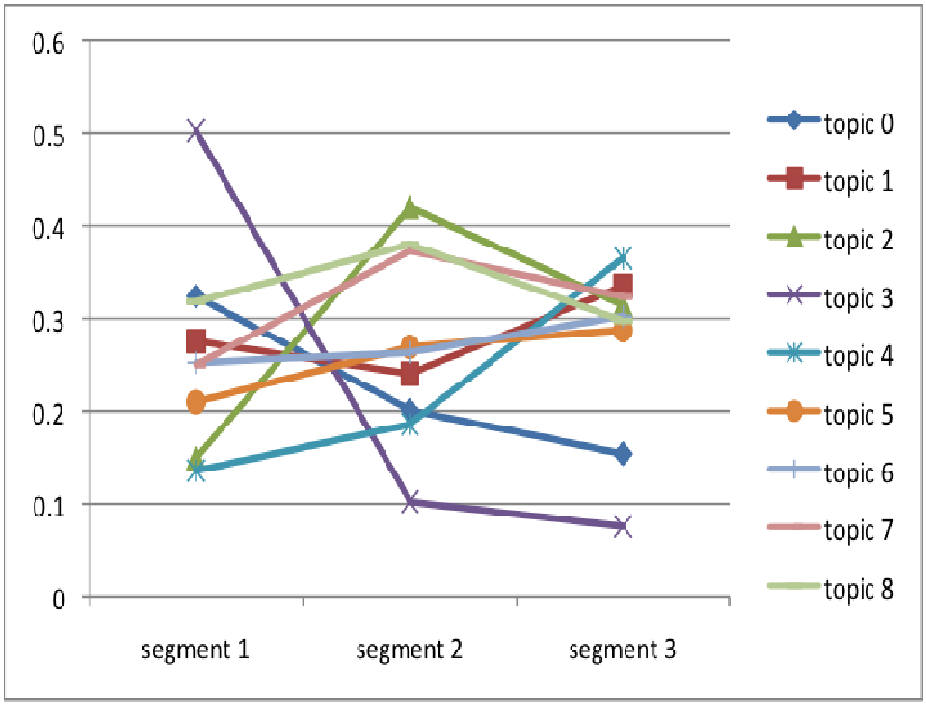


Figure 7. Average topic weight vs. temporal segment

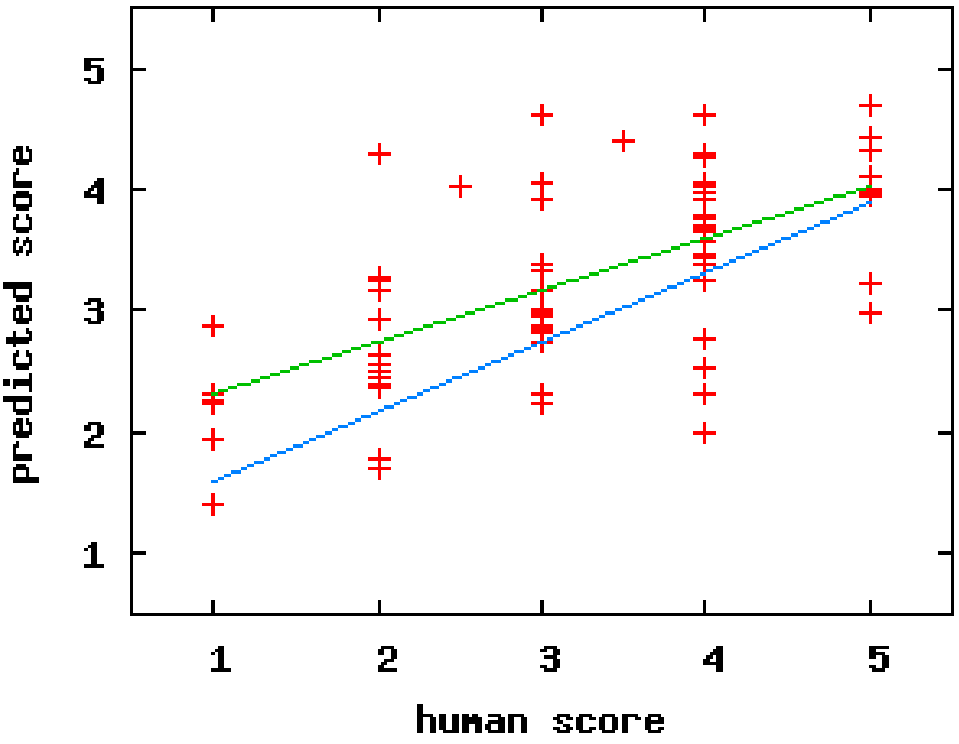


Figure 8. Correlation between LSA-SVM predicted scores and human scores

The model constructed by the SVM predicts the human rater's scores with an average Pearson's correlation coefficient of 0.63 over a 10-fold cross validation. The mean absolute error was 0.70. In other words, there was a disagreement of usually less than one point within the 1-5 Likert-type rating scheme used by the coders. The scatter view, in Figure 8, shows a comparison between the human rater's scores on the horizontal axis and the automatic prediction on the vertical. The upper green regression line through the data shows the regression between the predicted and actual scores. Inter-coder reliability with a second rater's scoring was 0.79 Pearson's correlation coefficient and its regression (without the corresponding points) is shown as the lower blue line in the scatter plot.

Despite not yet achieving the same level of correlation as that of the two humans, the automatic approach using LSA provides a strong baseline from which to improve. However, indications from our experiments with other methods of direct manipulation on the frequency data, e.g., reweighing and preclustering, or different algorithms (e.g., Latent Dirichlet Allocation, Neural Networks) increase the performance only minimally (or not at all). Thus, it is likely these results represent the highest performance available using a term-frequency approach in isolation.

Returning to the research questions of the present study: RQ3—What does the visualization of trainee performance data coming from unscripted game play transcription offer toward further understanding of the empirical results? Now that we have evidence which supports using this approach to reliably corroborate human rater/coder ratings for unscripted communication in a non-kinetic engagement mission we believe visualizations can be used to further identify trainee strengths and weaknesses especially when against pre- and post-test self-report on

learning. For instance, we could use the visualization to identify discrepancies between performance (as evidenced through game play behaviors, events, transcriptions, and Reflective Observer/Evaluator ratings) and self-report of perceptions of learning to better understand if any trainees perceive they learned even though their performance may have been rated poorly. .

With the goal of improving the accuracy of prediction in the present study, two avenues that could be explored are sentiment analysis (Lin and He, 2009) and automatic segmentation (Brants et al., 2002).

Sentiment analysis assigns a positive or negative score to terms using prior knowledge. This may lend itself to effectively weigh frequency counts and indicate whether the dialogue has a positive tone or a negative one. While LSA may do an effective job of clustering synonyms into topics, individual words within a topic may have dramatically different sentiment or tone. We expect such information would be crucial in a human rater's determination, conscious or not, of whether or not the dialog was successful.

By segmenting the transcription at three known points to incorporate a small amount of temporal information, we increased performance dramatically over applying LSA on the document as a whole. The approach used by Brants et al, (2002) is a fully automated method for segmenting a document into discrete regions. Using this in combination with sentiment analysis may give us a more complete picture of the conversation including when information is introduced to the discussion and the tone of the actors at various stages. The following sections describe an online game that was developed to test our approach for creating individualized training vectors.

9. ONLINE GAME DEVELOPMENT FOR LARGE-SCALE INDIVIDUALIZED TRAINING VECTOR ANALYSIS

The Inspect – Collect game (Figure 9) was developed internally at Sandia National Labs using the Unity Game Development Tool to follow the mission narrative of the DARWARS Ambush NK! and provided an opportunity to collect additional log information necessary to perform the analysis automatically and in real-time within the game environment. The evidence collection mission involves a player entering a suspect's house to quickly inspect and collect any available evidence in support of the suspected criminal activity. While this mission is less culturally motivated than the first, there are still protocols to follow and learn as part of the training. When taken in conjunction with the mission in DARWARS Ambush! NK the combination of the two missions from the two games form a complete training experience that includes cross-cultural training, tactical questioning, situational awareness, adaptability, and tactical site exploitation (TSE) procedures.

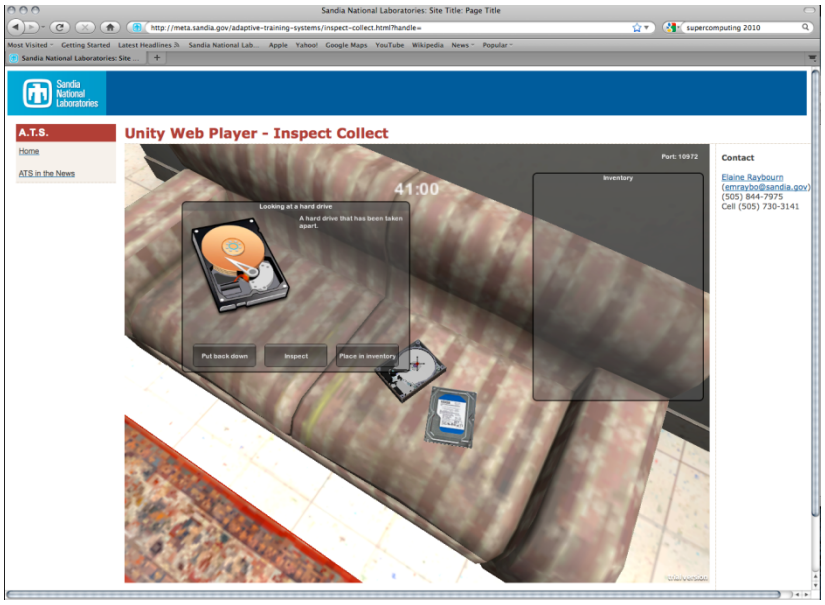


Figure 9. Web Deployed inspect-collect mission

The Player User Interface focuses on minimal screen clutter. By decreasing the boarder (amount of content on the screen) between the user and the game content the user can mentally juxtapose themselves into the simulation, a central tenant of Csikszentmihalyi's flow theory (Csikszentmihalyi, 1990). The player interface is developed from a first person perspective with an avatar window so the player can view their nonverbal communication from the front, unlike a 3rd person perspective that traditionally focuses on the back of the avatar. The player interface has a timer at the top center. The timer begins at 45 minutes and counts down. Every time a player interacts with their environment, such as inspecting or picking up an object, deducts time from their timer. The timer can go into negative time if the player clicks on too many objects, inspects objects that are not relevant or is not efficient in their decision making.

The Observer/Evaluator User Interface is designed to provide the user with the most information about the current inspect-collect mission. Observer/Evaluators were implemented in the TSE mission with the additional capability to selectively pose either standard questions or free-form questions to the trainee during the course of the mission. The Observer/Evaluator can see in real time what the player views and can prompt questions to the player in real time, to remind the player about something they missed or ask a question about their search process. This would pause the mission and not count against trainees in the overall scoring. Also, the extended capability and internet deployment allows multiple observers to watch a single trainee and pose questions or score performance.

The discrete nature of the mission in the inspect-collect allows us to automatically generate a numerical score based on a priori performance specification, which helps to automate the analysis without needing a human judge to intervene. Capability has also been implemented to change the nature of rooms on an individual basis, by virtue of either adding complexity to the baseline through extraneous, distracting materials (e.g., excess trash), or removing complexity from the baseline by revealing or obviating certain evidence (e.g., exposing a hard drive which

was previously hidden in a couch). This moves toward the goal of adapting the game in real-time to make the learning more affective.

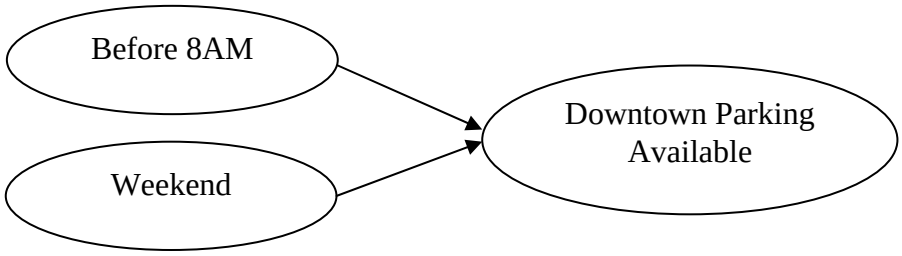
A research goal this effort was to adapt a game automatically to a trainee, or create an individualized training vector. Our approach to this research was to capture total in-game task performance and attempt to derive or infer the individual's vector for the latter portion of the task performance from analyses of prior performance in the same session. This approach could allow us to adapt the game in real time— to change the latter segment of the game itself or to steer the trainee back to a pedagogically correct path. Thus we asked the final research question:

RQ4: Can this approach be used for future games development of player/learner models and game adaptation algorithms? The following section describes a validation pilot study that was conducted to demonstrate the development of individualized training vectors.

10. DYNAMIC BAYESIAN NETWORK STUDY

10.1 Dynamic Bayesian Networks

Bayesian Networks (Pearl 1986) are graphical models that represent conditional dependencies between random variables. For example, the simple Bayesian network in the example below indicates that the availability of downtown parking is conditionally dependent upon both (a) whether or not the time is prior to 8AM and (b) whether or not the day is a weekend. The same network further implies that whether or not it is before 8 AM is independent of whether or not it is a weekend.



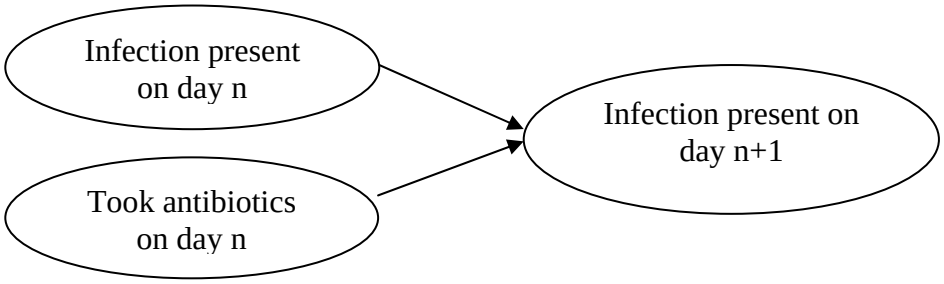
Simple example of a Bayesian Network

Associated with each node is a Conditional Probability Table (CPT). An example CPT for downtown parking availability is shown in Table 2. Note that knowing whether it is before or after 8AM and whether or not it is a weekend isn't sufficient information to completely determine whether or not it's possible to find a parking space downtown. Rather, as shown in the CPT, the probability of finding a parking space is affected by these two conditions.

Before 8AM	Weekend	T	F
F	F	0.05	0.95
F	T	0.5	0.5
T	F	0.4	0.6
T	T	0.9	0.1

Table 2. A Conditional Probability Table for Downtown-Parking-Available

Dynamic Bayesian Networks (DBNs, Murphy, 2002) are a particular variety of Bayesian networks which represent conditional dependencies over time. For example, the DBN shown in Figure 10 indicates that the presence of an infection on a given day is conditionally dependent upon the presence of an infection on the previous day and whether or not the potentially infected individual took antibiotics on the previous day (a full specification of this DBN would require CPT's for each node).



An example of a DBN

Bayesian Networks are models, i.e. abstract, simplified descriptions of aspects of the world. The task of devising a DBN may be broken into two subtasks: (1) devising the structure of the model, and (2) populating the associated CPT's. One valid way to devise the model structure is for a human to employ knowledge of the target domain to create a suitable structure by hand. Alternatively, the structure can be created in an automated fashion, typically via a search process that evaluates possible structures to find the best one (as defined with respect to some specified criteria). Once a structure is identified, CPT's can be populated straightforwardly, simply by using the observed frequencies of joint events. For instance, in the CPT shown in Table 2 the probabilities of 0.9 and 0.1 indicated in the row corresponding to before 8AM on weekends may be derived from data in which parking spaces were found to be available downtown 9 out of 10 times among observations that were made before 8AM on weekends.

To understand how DBNs may be used for classification, consider the case of the DBN in the example above. Imagine we built two versions of the model: One with data from cases with

bacterial infections and the other with data from cases with viral infections. We would then expect that probabilities in the two models should be different: For viral infections, the probability of an infection being present on day *n+1* would be essentially unaffected by whether or not the infected individual took antibiotics on day *n*. These two DBN models would thus reflect two different underlying situations. Then, when presented with data for a new case, we could calculate the *likelihood* of each DBN generating the data in question. If the new data reflects a case where the presence of the infection appears conditionally dependent upon taking antibiotics the day before, the likelihood scores should indicate that DBN built from bacterial infection is a better model of the data. Similarly, if the course of infection appears independent of antibiotic use, the viral infection model should appear more likely.

To support the goal of building automated adaptive training systems, we wanted to assess the potential for using an Individualized Training Vector (ITV) to make inferences about trainees in real-time as they interact with a training simulation. We chose DBNs as a mechanism for experimentation for multiple reasons: (1) Once trained, they can be used to progressively assess an ongoing stream of events with minimal computational overhead, and (2) they are relatively straightforward to train, interpret, and reason about.

10.2 Deriving Individual Training Vectors in the Game with DBNs

Two kinds of information are logged from the TSE game: (1) trainee movements, and (2) discrete events occurring in the game. A short sample segment of logged discrete events is shown in Table 3.

Time	Event
3058.34	Threshold to room non-room
3060.78	Timer adjusted by 1:00
3060.78	Object Jeans picked
3061.795	Object Jeans viewed
3062.659	Timer adjusted by 1:00
3062.659	Object Bed picked
3064.153	Object Bed viewed
3067.673	Timer adjusted by 1:00
3067.673	Object shoes_f0 picked
3069.864	Object shoes_f0 PlaceInInventory
3072.92	Threshold to room non_room

Table 3. Data sample from TSE game

For DBN analysis to be computationally tractable, a small set of time-varying discrete random variables must be identified. It's not hard to imagine a great variety of ways to transform the logged TSE data into this form. We decided to try one that was extremely straightforward. We chose to subsample the event data to discard all but five of the most common types of events: (1) crossing the threshold of a room, (2) selecting an object for manipulation, (3) viewing a selected object, (4) searching a selected object, and (5) placing a selected object into the trainee's inventory. Figure 12 shows the same sequence of data depicted in Figure 11 after the application of our subsampling process.

Event Token
Threshold
Picked
Viewed
Picked
Viewed
Picked
PlaceInInventory
Threshold

Table 4. The data subsampled to only five selected events

Note that with explicit clock time now discarded, the implicit notion of time in our model is based simply upon events; time advances by one step whenever one of the five designated events occurs. Given a simplified data stream as shown in Table 4, we defined 5 discrete random variables: *searched, viewed, placein, picked,* and *threshold*. At any given time-step, the variable corresponding to the current event is set to 1 and all 4 other variables are set to 0.

For our initial experiment, we chose to attempt learning to discriminate between novice and expert trainee behavior in real-time. Without true expert and novice behavior traces, we came up with the next best thing: One research team member played the TSE game 20 times, 10 performing as an expert and 10 performing as a novice. We then used freely available SBNet[2] software to develop DBNs based upon these data (transformed as described above). An example DBN is shown in Figure 10.

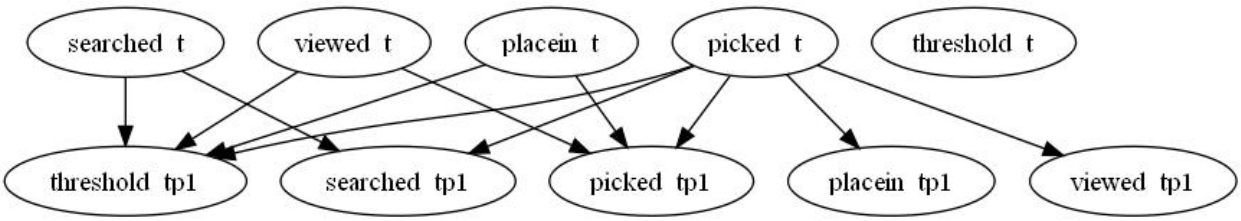


Figure 10. A DBN trained on data from the TSE game, transformed as described

In building our DBNs, we selected an option provided by SBNet specifically for the purpose of classification. Rather than building individual DBNs focused upon the strongest dependencies in a given dataset, the Approximate Class-conditional Likelihood (ACL, Burge & Lane 2005) measure was used to favor dependencies in the data that emphasize the differences between two sets of data, and thus increase discrimination power.

To make maximal use of our data, we employed *leave-one-out validation*. That is, for each of the 20 novice and expert runs, we applied SBNet to build two DBNs—one model of expert behavior and one model of novice behavior—based upon the *other* 19 runs. We then derived Individual Training Vector (ITV) of length 2 for each run by calculating the likelihood of these

[2] SBNet was created by Dr. John Burge. At the time of this writing, the SBNet software was available for download at `http://www.cs.unm.edu/~lawnguy/sbnet/index.html`

two corresponding DBNs with respect to the run in question. The results for all 20 runs are plotted in Figure 14.

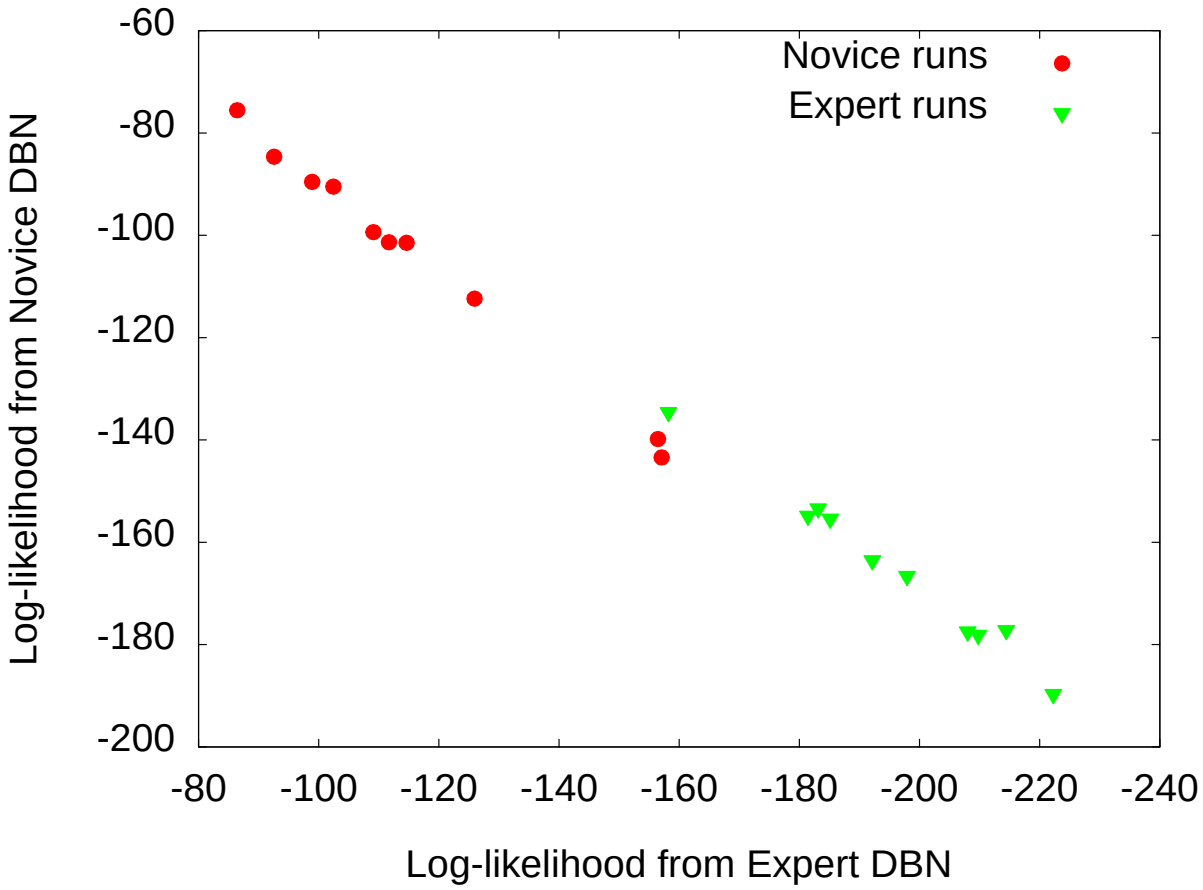


Figure 11. Independent Training Vectors for each of the 10 expert runs and 10 novice runs, as calculated using DBNs trained using the remaining 19 runs.

Note in Figure 11 that it is possible to draw a linear decision boundary that clearly divides the ITVs from 9 out of 10 expert runs from the rest of the runs, permitting correct classification of 95% of the runs. Furthermore, this level of performance was achieved using a maximally simple set of features drawn from the data. It appears highly likely that significantly better performance could be achieved from using a set of features that were optimized to discriminate between novices and experts.

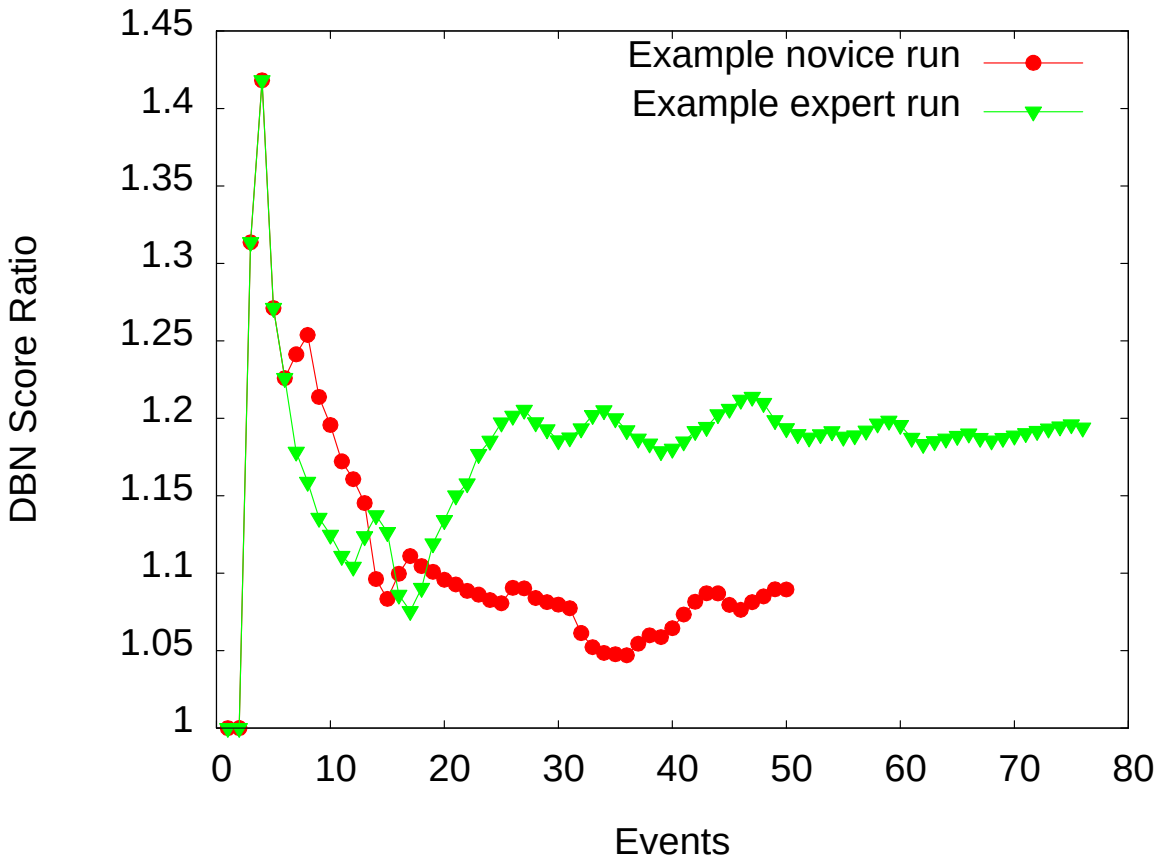


Figure 12. Divergence over time of the ratio between the two DBN scores in the ITV for one sample novice run and one sample expert run.

Figure 12 illustrates how ITVs may be used to automatically adapt training content to the user in real-time. We randomly selected 1 expert run and 1 novice run. For each run, starting with the first event in the run we ran each successive event by the two corresponding novice and expert DBNs to compute a progressively evolving ITV from the likelihood scores. The ratio of these two components of the ITV is plotted in Figure 12. As we can see, for the first 20 or so events it is not possible to discriminate between the runs using the DBN score ratio. However, after about 20 or so events, the two ratios begin to diverge and remain separate for the rest of the runs. After collecting the proper statistics over a range of runs, the system should have a profile that should enable it to become progressively more certain (in real-time) whether the trainee in question is a novice or an expert, and use this information to adapt training content.

11. FUTURE RESEARCH

The present LDRD sought to investigate the use of multiple roles in non-kinetic game-based training, and develop an approach for the design of individualized training vectors for experiential learning. A quantitative and qualitative analysis was conducted on participation in one role during the first 25 minutes of a longer training exercise in which participants self-selected roles and then after the first session was complete, switched roles. As the participants were not randomized into group assignment there is some chance that the results may be biased however there is no indication from examining the results to suggest this concern. Participants were not randomized into groups in order to closely follow the quasi-experimental protocol that attempts to replicate a true training event. A future repeated measures analysis will further investigate multi-role experiential learning. Future research will address the order in which one plays both roles and whether there are any notable differences among multi-role learning experiences. Research is also currently underway by the author to identify salient factors in communication performance, recall of information, social behavior modeling, and the development of new strategies in multi-role game-based training.

The present paper sought to go *beyond game effectiveness* to understand how *game-based learning can more effectively* focus on training adaptable non-kinetic engagement skills by introducing multiple learning roles that exercise different cognitive skills. To do so we introduced multiple learning roles that exercise different skills (reflection, perspective taking, adaptability, and intercultural communication). This approach is different in spirit and design from more conventional means of interaction and engagement used in training and entertainment games. The training and game design approaches used by the author were empirically tested and the results presented in the present paper. Contrary to current trends in military game development, the present study illustrates (albeit not entirely) that experiential learning can be supported by approaches designed to facilitate trainee mastery of reflective observation and abstract conceptualization as much as performance-based skills. By providing trainees with the opportunity to switch roles, and play from different perspectives we can engender the adaptive behaviors needed to excel in non-kinetic engagements and STTR operations.

Finally regarding RQ4—can this approach be used for future games development of player/learner models and game adaptation algorithms? As mentioned, the larger strategy guiding our incorporation of automatic scoring is to adapt a game's content automatically to a trainee's individual strengths and weaknesses. We examined the effectiveness of temporal state space learning including the use of Dynamic Bayesian Networks (DBN) to create learner models. The automatic segmentation of transcriptions for LSA lent itself to construction of state space transitions, which were necessary as inputs to a DBN learning algorithm. We believe that this state-transition model is a useful general representation of activity in any game, exemplified by the more esoteric conversation space used in this LDRD, and thus as an approach, generally applicable.

We applied advanced analysis and visualization techniques to discover and interpret the underlying trends and patterns associated with communication behaviors that can someday potentially lead us to better predict how well trainees will perform in role-playing scenarios and whether game systems can respond to individual strengths and weaknesses as trainees learn

communication strategies. The results are applicable to the assessment and design of First-Person game-based learning systems designed to enhance trainee intercultural communication, interpersonal skills, and adaptive thinking.

The ratings garnered from the Reflective Observer/Evaluator roles were used along with trainee transcripts, events, and coded behavior toward the assessment of trainee performance for the long-term goal of adapting training game content to account for individual strengths and weaknesses. One limitation of codifying behavior into a single number rating is that it artificially limits expressiveness of that rating. Further work will include looking at the original dimensions of rating and processing a multidimensional value. This would potentially allow a much more individualized value in the result.

While the preliminary results of the DBN experiments are encouraging, they are not yet sufficient to support strong conclusions. First, statistical analysis is required to establish confidence in the ability to do real-time discrimination. Next, it is necessary to validate the ability to make inferences from data sets that purport to support real-world training objectives, and to employ this approach to dynamically adapt training to enhance learning in a verifiable manner.

If successful, this technology could support a strongly data-driven adaptive training system. Such a system would be most obviously useful when both (a) human instructors are in short supply and training objectives are unknown, and (b) relevant learner state is difficult to infer by other means. In many real-world situations, however, a purely data-driven approach to automated training may be neither needed nor desirable. Nonetheless, there may be situations in which data-driven, real-time learner assessment and training adaptation may still have a role to play in ill-defined training and learning domains.

Due to factors such as individual differences, changing instructional context goals, objectives, and changing learning platforms, teaching and training present continual challenges. While it's unlikely computer systems alone will outperform live humans in meeting these challenges, humans and computers together may outperform each on their own. Because computer-based systems both perceive and process data in a distinctly different manner than humans, the two may prove highly complementary. It is thus in hybrid systems which play to the strengths of both humans and computers that data-driven approaches to learning may be most useful.

12. REFERENCES

Beal, S. A. (2009). Exploring the Use of a Massive Multiplayer Game (MMPG) to Train Infantry Company Commanders. I/ITSEC 2009 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, November 30- December 04, Orlando, Florida, USA.

Beal, S. A. (2006). Lessons learned from evaluating training games for Infantry leaders. I/ITSEC 2006 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, December 4-7, Orlando, Florida, USA.

Belanich, J., Orvis, K. L. Mullin, L. N. (2004). Training game design characteristics that promote instruction and motivation. I/ITSEC 2004 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, December 6-9, Orlando, Florida, USA.

Bennett, M. J. (1986). A developmental approach to training for intercultural sensitivity. *International Journal of Intercultural Relations, 10,* 179-96.

Brants, T., Chen, F., & Tsochantaridis, I. (2002). Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis. In International Conference on Information and Knowledge Management (CIKM), McLean, VA, USA.

Brown, B. (2010). A Training Transfer Study of Simulation Games. Unpublished Master's Thesis, Naval Postgraduate School, Monterey, CA.

Burge, J. & Lane, T. (2005). Learning class-discriminative dynamic Bayesian networks, Proceedings of the 22nd international conference on Machine learning, p.97-104, August 07-11, 2005, Bonn, Germany.

Duda, R.O., Hart, P.E., and Stork, D.G. Pattern Classification. John Wiley & Sons, 2001.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34, 906-911.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

Kolb, D. A. (1984). Experiential learning: Experience as the source of learning and development. New Jersey: Prentice-Hall.

Landauer, T. K., Foltz, P. W., Laham D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic Essay Assessment. Assessment in Education, Vol. 10, No. 3.

Lin, C. and He, Y. (2009). Joint Sentiment/Topic Model for Sentiment Analysis. CIKM '09: Proceedings of the 18t ACM conference on Information and Knowledge Management. 275-384.

Livingston, J. A. (1996). Effects of metacognitive instruction on strategy use of college students. Unpublished manuscript, State University of New York at Buffalo.

Menaker, E., Coleman, S., Collins, J., & Murawski, M. (2006). Harnessing experiential learning theory to achieve warfighting excellence. I/ITSEC 2006 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, December 4-7, Orlando, Florida, USA.

Murphy, K. P. (2002). Dynamic Bayesian Networks: Representation, Inference and Learning. University Of California, Berkeley.

Orvis, K., Horn, D. & Belanich, J. (2006). Videogame-Based training success: The impact of trainee characteristics - Year 2 (Technical Report 1188). U.S. Army Research Institute for the Behavioral and Social Sciences: Arlington, VA.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* (Elsevier) **29** (3): 241–288.

Raybourn, E. M., Deagle, E., Mendini, K., & Heneghan, J. (2005). Adaptive Thinking & Leadership simulation game training for Special Forces Officers. I/ITSEC 2005 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, November 28-December 1, Orlando, Florida, USA.

Raybourn, E.M. (2007a). Applying simulation experience design methods to creating serious game-based adaptive training systems. *Interacting with Computers* 19, Elsevier. 207-14.

Raybourn, E. M., Roberts, B., and Diller, D. (2007b). DARWARS Ambush! NK documentation available from PEO-STRI. http://www.peostri.army.mil/PRODUCTS/DARWARS/.

Raybourn, E. M., Roberts, B., Diller, D., & Dubow, L. (2008). Honing intercultural engagement skills for stability operations with DARWARS Ambush! NK Game-based training. Unpublished manuscript and poster presentation, Army Science Conference, 2008, Orlando, Florida, USA.

Raybourn, E.M. (2009a). Beyond Game Effectiveness Part I: An Empirical Study of Multi-role Experiential Learning. I/ITSEC 2009 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, November 30- December 04, Orlando, Florida, USA.

Raybourn, E.M. (2009b). Intercultural Competence Game that Fosters Metacognitive Agility and Reflection. In A.A. Okok and P. Zaphiris (Eds.). Online Communities, Lecture Notes in Computer Science (LNCS) 5621, 603-12. Springer-Verlag.

Raybourn, E. M. (2006). Simulation experience design methods for training the forces to think adaptively. I/ITSEC 2006 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, December 4-7, Orlando, Florida, USA.

Raybourn, E. M. (2007c). Training approaches for honing junior leader adaptive thinking, cultural awareness and metacognitive agility. I/ITSEC 2007 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, November 26-29, Orlando, Florida, USA.

Roberts, B., Diller, D. and Schmitt, D. (2006) Factors affecting the adoption of a training game. I/ITSEC 2006 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, December 4-7, Orlando, Florida, USA.

Rowan, P. A., & Brown, D. (2008). Games – Just how serious are they? I/ITSEC 2008 Proceedings, Interservice/ Industry Training, Simulation and Education Conference Proceedings, December 1-4, Orlando, Florida, USA.

Selmeski, B. (2007). Military cross-cultural competence: Core concepts and individual development. Armed Forces & Society Occasional Paper Series, (1) AFCLC Contract Report 2007-01. Canada: Royal Military College of Canada Centre for Security.

Veenman, M., Kok, R., & Blote, A. (2005). The relation between intellectual and metacognitive skills in early adolescence. Instructional Science (2005) 33: 193–211.

White, B., Shimoda, T. & Frederiksen, J. (1999) Enabling students to construct theories of collaborative learning and reflective inquiry: Computer support for metacognitive development. *International Journal of Artificial Intelligence in Education* 10, 151-182

Wylie B., & Baumes, J. (2009). A Unified Toolkit for Information and Scientific Visualization. In Proceedings of Visualization and Data Analysis (VDA). SPIE.

DISTRIBUTION

1	MS1188	John Wagner	1462
1	MS1188	Phil Bennett	1463
1	MS1188	Alan Nanco	6114
1	MS3408	John Mitchiner	1460
1	MS1323	David Rogers	1461
1	MS0899	Technical Library	9536 (electronic copy)
1	MS0359	D. Chavez, LDRD Office	1911

Sandia National Laboratories