Enhancing Good Stranger Skills: A Method and Study

Robert Hubal, Mike van Lent, Bob Marinier, Chris Kawatsu, Bob Bechtel
Soar Technology, Inc., 3600 Green Ct. Ste. 600, Ann Arbor, MI, 48105
{robert.hubal, vanlent, bob.marinier, chris.kawatsu, bob.bechtel}@soartech.com

ABSTRACT

A good stranger (GS) is a professional who can effectively integrate tact and tactics, in order to create positive outcomes in difficult social encounters. For military personnel, creating positive social outcomes enhances mission effectiveness and force security, and supports broader strategic and tactical objectives. Some evidence suggests that military personnel may come into situations with preconceived ideas, or frames, about how to behave, not all of which involve GS tactics. Deliberate training in a variety of such situations is required to gain more effective control of people and situations. As part of a large DARPA-funded program maximizing especially high-risk, high-consequence interactions occurring in unfamiliar social terrain, we investigated how to train military personnel on GS skills in order to adapt to and successfully manage these interactions. Training was based on a theoretical structure for GS skills-based interaction; generally this flow maps to the basic sequencing for most interactions that produce positive end states: An approach, a period of framing, orientation and sensemaking, followed by engagement in the evolving business of the encounter. This engagement often involves necessary rapport-building, trouble recovery, and appropriate departure. We conducted an experiment with students at the Infantry Basic Officers Leader Course at Ft. Benning using a browser-based tool developed under the DARPA funding. We presented 32 students with a series of storylines, some having multiple injects, and asked the students to demonstrate their perception of relevant cues in a scene as they observed the interaction depicted by the storyline. We found this training to have a positive effect in increasing behaviors associated with a GS frame. In this paper we detail the training approach, describe our study, and offer recommendations on improving GS skills training in military personnel.

ABOUT THE AUTHORS

Robert Hubal, PhD, is a lead scientist at SoarTech interested in using technology intelligently for training and assessment. He has conducted basic and applied research in decision-making, mental modeling, vigilance, intelligent tutoring, linguistic analyses, patterns of life, scientific visualization, and usability. He has applied research results to everyday and specialized domains such as clinical assessment of social and interpersonal skills, improved patient communications, law enforcement interactions, and warfighter/civilian sociocultural engagement.

Mike van Lent, PhD, is CEO of SoarTech and conducts R&D of artificial intelligence for serious games, dynamically individualized training environments, social-cultural behavior models, and computer generated forces. Prior to SoarTech, he worked at the University of Southern California's Institute for Creative Technologies as associate director for games research, where he led R&D of a range of game-based training technologies for Army applications including explainable artificial intelligence, social simulations, and intelligent tutors.

Bob Marinier, PhD, is an expert in modeling behavior using the Soar cognitive architecture for a range of domains including mine countermeasures planning, autonomous ground vehicle planning, and social skills training. He participated in RDECOM's MAGIC 2010 robotics competition as part of first-place Team Michigan, under which he was the primary implementer of the SAGE user interface.

Chris Kawatsu is an AI engineer interested in modeling and predicting student performance. Mr. Kawatsu holds a B.S. in physics from the California Institute of Technology and an M.S. in computer science from Lawrence Technological University.

Bob Bechtel, PhD, has over 30 years' experience applying advanced computing technologies to problems in the defense and commercial sectors. At SoarTech, he is a principal investigator and program manager on projects in a variety of areas ranging from robotics to political-social modeling.

Enhancing Good Stranger Skills: A Method and Study

Robert Hubal, Mike van Lent, Bob Marinier, Chris Kawatsu, Bob Bechtel
Soar Technology, Inc.

INTRODUCTION

The mission of the Defense Advanced Research Projects Agency's (DARPA) Strategic Social Interaction Modules (SSIM) program was to maximize Warfighters' ability to adapt to and successfully manage all interactions, especially high-risk, high-consequence interactions, on unfamiliar social terrain. DARPA and others identified a number of positive outcomes from the application of "good stranger" (GS) skills, including de-escalation of conflict, reduction of unnecessary use of force, security and boundaries, reduced mutual perceptions of humiliation or disrespect, respect of culture, increased flow of actionable information, ethical decision making, legal decision making, correction of errors and misperceptions, increased mutual perceptions of trust and respect, effective communication, enhanced cooperation with host nation allied forces and civilian populations, and improved cross-cultural intelligibility. Overall, creating positive social outcomes enhances mission effectiveness and force security, and supports broader tactical and strategic objectives (Hubal et al., 2015).

We investigated a training system that presents an environment for teaching and assessing social skills in dynamic social and tactical situations. Other SSIM performers had documented gaps in the current training of GS skills across military schoolhouses (Logan-Terry & Damari, 2015). Our goal was to enable a focus on sensemaking, social affordances (i.e., aspects of or cues within the situation that represent social opportunities), and the intentions that lead to actions in high consequence social situations. The approach involved providing decision points where the student's selection of an option causes a particular branch of the situation to happen, and annotation frames (described further below) on which the student must note relevant cues or label important features of the scene. As such the system simulates decision-making and portrays consequences.

MOTIVATING THE TRAINING OF GOOD STRANGER SKILLS

The development of a GS may require some deviation in the way some military personnel think about situations. Even once the GS perspective is perceived as a valuable capability, the change in behaviors must be followed by practice and refinement to become natural.

There are several reasons why training GS skills, along with instilling mechanisms of reinforcing GS behaviors, is important. These include:

- *Increased effectiveness*. In many situations, Warfighters with a GS perspective can accomplish their mission more quickly and effectively, and with better consequences and fewer adverse outcomes.
- *Extensibility*. GS skills translate to many aspects of Warfighters' lives. When under stress, otherwise routine interactions with family members, friends, inattentive clerks at a store, and other situations can turn to negative outcomes. GS skills have the broad potential to assist with challenges in everyday situations.
- *Norms*. GS skills are accepted behaviors of other, successful military units who are to be emulated.
- *Peer pressure*. Peers avoid those who cannot or do not effectively integrate tact and tactics. Non-GS behavior risks escalation, does not demonstrate benevolence or integrity, and may increase security concerns.
- *Performance expectations*. Leaders and experts expect subordinates to act using GS skills.
- *Direct feedback and consequences*. Seeing people's reactions to being treated well or poorly has a powerful effect. A GS approach has short- and long-term consequences. One important consequence is increasing safety if the citizenry is accepting rather than hostile.
- *Models*. Observing respectful interaction with a civilian that still gains compliance can illustrate the advantages of practicing GS skills. The opposite may also hold; witnessing forced or coerced compliance through intimidation or command may lead the Warfighter, particularly a novice, to tend toward a non-GS perspective. In such cases, the adverse consequences of this tendency should be made clear, and more appropriate skills (e.g., when intimidation is modeled, GS skills such as reducing provocation) should be stressed.

LEARNING OBJECTIVES AND STORYLINES

Learning Objectives

During the SSIM program our team defined a set of learning objectives (LOs) to delineate the skills involved in being a GS. (We note that language *per se* was *not* an emphasis of the SSIM program.) Table 1 lists ten high-level LOs, each of which was broken down into a hierarchy (not shown) of sub-objectives.

Table 1. Learning Objectives

LO	Description
1. Initiating/ Reinitiating the Encounter	The GS uses initiating skills to project authority, calmness, and, as necessary, positivity and respect, and to adapt to discomfort or confusion on the part of the other persons that may require changing tactics and de-escalating or recovering from trouble.
2. Attending to Nonverbal Cues	Nonverbal communication includes hand gestures, eye contact and eye gaze, and body positioning. These can be used to aid or enhance comprehension, indicate the topic of conversation, signal an intention, and convey meaning. They can also add emphasis, determine who is being addressed, and indicate power relationships. Nonverbal cues thus represent an important source of situational information.
3. Making Sense of the Encounter	Sensemaking is a deliberate effort to understand events. It involves models that people develop for explanations of others' behaviors. These models are based on individual experience and social and cultural factors, and rely on paying attention to details of the encounter and questioning of assumptions to better understand how and why certain activity is occurring.
4. Taking the Perspective of Someone Else	The GS needs to accurately take the perspective of others. This skill requires insight into others' thoughts, motivations, concerns, and decision-making to identify goals and priorities, anticipate how others are likely to act or react, and explain behavior. Perspective-taking is only possible if the GS has some understanding of his own belief and value systems, and how others' differ.
5. Building Rapport with Someone Else	Building rapport is very useful for achieving desired positive outcomes. It often has a long-term component, such as contributing to trust that might encourage others to take actions in support of the warfighter's tactical and strategic objectives. This state can be achieved by identifying shared goals, especially when the other's goals are acknowledged, respecting cultural boundaries, and repairing conflicts by using trouble recovery techniques.
6. Recognizing "Affordances" of the Encounter	Sensemaking is about figuring out what is going on in the situation. Recognizing social affordances is about figuring out what actions are available. The GS is interested in creating a social situation favorable to his objectives, combining tactfulness with tactical impact, and not just passively accepting the details of a situation.
7. Balancing Tact and Tactics	It is important for the GS to learn how to combine (1) self-protection and safety, (2) the ability to control emotions and get an objective sense of the situation, and (3) taking decisive actions that become tactical advantages. Adopting a security-focused approach does not rule out simultaneously trying for a positive social outcome. Such an integration of tact and tactics can actually lower the immediate threat level, and thus improve everyone's safety.
8. Repairing the Encounter	Warfighters must always remain aware of the inherent instability of social interactions. A GS mindset encourages the warfighter to take steps to lower tensions before they rise out of hand. Recovery should be sought whenever it is safe and appropriate to do so, but the warfighter must always have the agility to move rapidly into and out of force.
9. Appraising Outcomes of the Encounter	The goal of the GS is to cause positive outcomes. A priority for any warfighter is to establish authority, as this relationship enhances the security under which he performs his missions. However, true authority requires a mutual understanding and respect between the two individuals, and the GS must understand how the integration of tact and tactics supports the intent in the current situation.
10. Attending to and Acting on Tactical Cues	Sometimes appearing too friendly or uncertain creates an appearance of weakness that could result in the need for more force than would have been required by taking a firmer stance. One characteristic of a GS is the ability to take decisive action, sometimes even at great cost to another's interests, yet without causing humiliation, which may foster resentment. Compliance is best gained when requests issued to individuals contain concrete, manageable actions that lead to acceptable resolutions.

The training system presents storylines and injects. A large curated list of situations (storylines) generated during the SSIM program is available for use for training or assessment of social interactions skills. Variations on each storyline (injects) involve different initial and branching conditions, different annotation screens, different response options, and different expected outcomes. Different storylines and injects make sense for different LOs. The tool provides an authoring interface to define LOs and metrics to facilitate social interaction skills assessment.

An important consideration during training is to take account of dependencies among LOs, to help determine in what order they should be introduced. For instance, demonstration of an ability to initiate an encounter typically precedes that for repairing it. Similarly, it is critical to ensure that, across all storylines and injects that a student experiences, all LOs are assessed, so that a complete picture is drawn of the student's competencies and areas for improvement. We developed modules that augment the training system by helping an author with these considerations. One such module, for instance, tracks LO assignment across all storylines to ensure full coverage.

Storylines & Injects

The storylines used in the current study are listed in Table 2. Sources included social tactical decision games, expert interview vignettes, observations of training during the Marines' Infantry Officer Course at 29 Palms, stories used during the gap analysis to evaluate student "frames" (described further below), discussions with specialists, and the scenario working group of performers and cadre from the Army's Infantry Basic Officer Leaders Course (IBOLC).

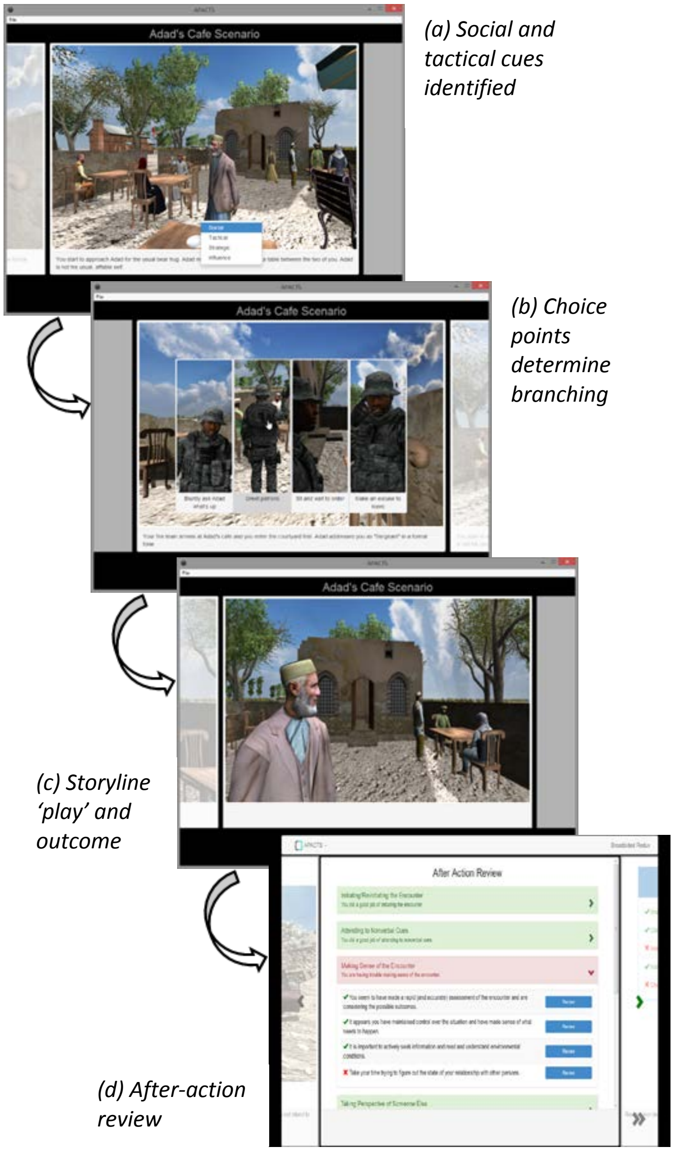
Each storyline, particularly the more complex ones, can have some number of injects. If the storyline is the 'backstory', then the inject is a variation on that theme. There are different types of injects, such as those changing initial conditions, those having different branching conditions or options, and those modifying the expected outcomes. Some variations are analogous to others, just with superficial differences; these alternatives can be used for pre- and post-testing.

Note there is a many-to-many relationship between storylines and LOs. That is, each storyline/inject may bring up one or more decision points or annotations that require the demonstration of a given LO, and across all injects all or any subset of LOs may be addressed. Consequently, the response(s) associated with a given LO may occur within only one storyline or across many.

BROWSER-BASED PERCEPTUAL AND COGNITIVE TOOL

The system used for this study targets a middle ground between immersive constructive or live training and more didactic classroom-based or virtual (non-simulation) training. By introducing graphical elements (pictures, short video clips) and involving interactivity through branch points and annotations, it enables the student to practice not only the cognitive portions of the target good stranger skills, but also some of the perceptual skills. It is integrated with tailoring (Wray, 2013) and coaching components, to include bookmarking and after-action review, to gain insight into the student's proficiency and to guide the student's training experience. We packaged the system in a form that can run inside any web browser (including Chrome, Firefox, and Internet Explorer) for deployment remotely and to tablets and other mobile devices. The implementation also allows for ease of updating and capture of student statistics.

Figure 1 depicts several frames of the system in use.



You start to approach Adad for the usual bear hug. Adad … table between the two of you. Adad is not his usual, affable self.

Your fire team arrives at Adad's cafe and you enter the courtyard first. Adad addresses you as "Sergeant" in a formal tone.

After Action Review

Initiating/Revitalizing the Encounter
You did a good job of initiating the encounter

Attending to Nonverbal Cues
You did a good job of attending to nonverbal cues

Making Sense of the Encounter
You are having trouble making sense of the encounter

✓ You seem to have made a rapid (and accurate) assessment of the encounter and are considering the possible outcomes.

✓ It appears you have maintained control over the situation and have made sense of what needs to happen.

✓ It is important to actively seek information and read and understand environmental conditions.

✗ Take your time trying to figure out the state of your relationship with other persons.

Taking Perspective of Someone Else

(a) Social and tactical cues identified

(b) Choice points determine branching

(c) Storyline 'play' and outcome

(d) After-action review

Figure 1. Screenshots of Training System

Table 2. Full List of Storylines Used in Study

Storyline	Description
An Invasion of Privacy (source: vignette from parallel project)	Student is on his way to a scheduled meeting with the local chief of police to build a relationship with him. While moving down a street along the side of his compound a soldier yells "IED!". Student ducks through the nearest door and finds himself in a courtyard inside the compound. The police chief is standing on the other side of the courtyard smoking a cigarette. He has a gun on his hip and starts yelling at the student.
Breaking Bread (source: S. Flanagan)	Student is the leader of a unit on patrol tasked with searching a house based on a tip that the house may contain an illegal weapon stash. Student is in the kitchen with a local national soldier leading the search and with an agitated local national civilian. The local national soldier, hungry after a long day, picks up a piece of bread off the kitchen counter and starts eating it. The civilian becomes outraged and picks up a kitchen knife. How does the student react?
Broadsided (source: IBOLC students)	Rumors are that locals are getting into accidents with U.S. vehicles to collect compensation from the U.S. Government. You respond to an accident with a civilian vehicle and an out-of-control soldier. A group of locals converges and tensions rise.
By Any Means Necessary (source: IBOLC cadre)	Three local national soldiers were just lost in an IED. Student is ordered to search a suspect's house, with a local national soldier. The search turns up clean. But this is a high-value individual's house and he must be tactically questioned immediately. The local national uses uncalled-for force on the suspect and his family to get information "by any means necessary".
Café Conundrum (source: J. Schmitt)	Student is going to a meeting with the indigenous owner of a local café with whom he has been developing a friendly relationship. Student believes the café owner has some intelligence he is willing to share about growing insurgent activity. When student arrives, however, he finds café owner acting nervous and very differently than usual. How does student deal with the situation?
Disaster Relief (source: Macrocognition)	Student's squad is providing security assistance during a disaster relief operation. As student finishes a regularly scheduled neighborhood walk-through, a team member spots a stranger running towards a nearby house. Student is familiar with the family that lives in the house. Student approaches the house and asks the homeowner if he has seen strangers in the area or heard anyone trying to enter his house. The man says "No" and starts to close his door. What does student do?
Illegal Weapon (source: Macrocognition)	Student is at a checkpoint searching for weapons. An Iraqi man enters the checkpoint. The man's background check comes out clean but student finds a pistol in his car. The driver protests that he needs the pistol to protect his family. Student knows that Iraqi men are only allowed one AK-47 per male per household and that pistols are not allowed in vehicles. How would student handle this?
Neighborhood Patrol (source: Macrocognition)	Student's squad is patrolling a neighborhood to provide security. These patrols are usually uneventful as the neighborhood has been peaceful for quite a while. A team member spots a young child dropping what looks to be a grenade from a second-story window of a home. He yells, "Run! Run!" Fortunately, there was no explosion and no panic. Student sees that it was actually a stone. What would student do?
Not Enough Room (source: S. Flanagan)	Student is the leader of a unit conducting a nighttime mission to detain a low level member of the local insurgent group. A local informant has come along to guide student to the right house. Unit and informant pile into two HMMWVs; space is tight. At the house student finds not only the low level member of the group but also a higher level target. There isn't room for everyone in the HMMWVs because student hadn't expected to detain two people. Who does student leave behind?
Ordnance Detection (source: Macrocognition)	As part of a U.N. peacekeeping mission, student has orders to support an ordnance detection team as they search a wooded area. Because this can cause dangerous explosions, families living nearby were ordered to leave their homes and land for several days. They were promised monetary compensation for their trouble. When student arrives, he finds that one family has not started to prepare. How would student get the family out so the detection team can start its work?
Security vs. Convenience (source: Macrocognition)	Student is stationed at a U.S. base that also houses a contingent of Afghani soldiers. A gate near the Afghani quarter is very close to their latrine but it must be kept locked at night for security reasons. The Afghanis have been told that the gate must always be locked. The gate, however, is often left unlocked. What would student do?
Stolen Knives (source: J. Schmitt)	Student is deployed to a base in the Marshall Islands as the dining facility director. Student is a bit picky and brought his own knives. Sometimes, when management tasks take student away, he returns to find one or two knives missing, though they turn up sometime later. Talks with locals about the importance of honor and respect for private property do not change their behavior; they feel property left in plain sight becomes communal property. But student isn't happy with this situation, feeling locals need to follow his rules.
Survivor's Guilt – A	Based on intelligence, student goes into a village, and with an interpreter forcibly enters a house,

Death of Innocents (source: E. Amdur)	trying to find a wanted subject. Things happen very fast and very badly. The house contained an extended family who were under death threats from the enemy (the intelligence was false). The patriarch of the family, with the door suddenly kicked in, grabbed his weapon, and student's team, perceiving a threat, began firing. Not only was the patriarch killed, but also two women and four children. The interpreter is devastated.
Suspicious Behavior (source: Macrocognition)	Student is assisting at a checkpoint near a neighborhood known to harbor insurgents. Protocol requires that student thoroughly check every car that enters and exits. A local resident student recognizes drives up to the checkpoint. The man is usually pleasant and co-operative. Today, though, he says he needs to rush and asks that the check be hurried. How would student handle this?

In frame (a) of Figure 1 the student will have stepped through a number of background frames and at this point is required to annotate cues in the scene that have social or mission-critical importance. (In the mode run during the experiment described in this paper, no immediate feedback on annotations was provided.) In frame (b) the student is presented with four choices to indicate how the storyline should flow. In frame (c) (and across subsequent frames) the student observes the effects of that branching decision. Finally, in frame (d) the student is presented with an after-action review that links the previous annotations and decisions to learning objectives.

IBOLC STUDY

To test how effective a system focused on perceptual and cognitive skills could be in training GS skills, we conducted a study with recent graduates of IBOLC. The study took place in late September and early October 2014 at Ft. Benning, GA. Each student used the training system over the course of approximately 2½ hours, running through all of the training storylines. The intent was to see if even in these rather restricted conditions, with a short timeframe and students who had already completed training, we could influence GS behaviors.

Methods

A total of 14 storylines were developed for the study (Table 2), and for every run-through students' actions within the storylines were captured. Each storyline contained some number of annotation frames and decision points, and every annotation and decision point was associated with one or more LOs. One storyline and a variant, Invasion of Privacy and Unexpected Intruder, were used as pre- and post-tests, meaning that they were run through just once each, with no feedback or guidance given to students. For the remaining storylines, we integrated coaching and review and encouraged re-running of storylines to address non-GS responses and explore response options. Prior to IBOLC, we conducted two testing trials with convenience samples to determine approximately how long each storyline took to run through and to refine the study instructions. Each storyline was vetted by at least one expert and one outsider, all of whom were SSIM performers who had not participated in the development of the storyline.

A total of 32 students took part in the study, all Second Lieutenants and all male. These students had recently completed IBOLC and were awaiting Ranger School or waiting to be assigned to their next units. We assigned students randomly to one of six presentations of storylines, depending on session (morning or afternoon) and blocking. Because the storylines differed in the length of time required to complete and in their number of injects, blocks of storylines were assigned *a priori* to 10, 15, or 20 minute time segments (see Table 3).

Procedures

The study was conducted using the following procedures. The first ten minutes were given to a welcome and overview. We described the study as a test of one of the software products coming out of the SSIM program. We told students that over the next three hours (though neither session turned out to require that much time) they would be seeing a series of storylines with different injects, and their job would be to run through the stories and answer questions, mark up the frame, or make choices as indicated on the screen. We gave each student a grouping to use for determining the order of storylines. One of the experimenters then briefed the students on the SSIM program.

The next ten minutes were assigned to a tutorial for students to learn to use the training system. They learned to sign in, load a storyline and variant, and start the storyline. A number of features of the software were then discussed, including scrolling back and forth through a storyline; fast-forwarding through introductory material (e.g., on a second run); different kinds of frames, images, videos, and choice frames; coaching; and adding and removing annotations. An after-action review was presented, with discussion of color coding, LOs and their descriptions, specific feedback for an LO, and how to jump back and forth between the review screen and student action frames.

Table 3. Storyline Groups

Time Allocated	Storyline Grouping		Storylines Involved (blocks shown for Session 1)		
	Session 1	Session 2	Block 1	Block 2	Block 3
15	Romeo	Indigo	Illegal Weapon	Ordnance Detection, Neighborhood Patrol, Not Enough Room	Suspicious Behavior, Disaster Relief
20	Echo	Delta	Stolen Knives, Suspicious Behavior	Café Conundrum	By Any Means Necessary, Survivor’s Guilt
10	Juliet	Lima	By Any Means Necessary	Suspicious Behavior	Breaking Bread
10	Yankee	Juliet	Broadsided	Survivor’s Guilt	Not Enough Room, Ordnance Detection
15	Indigo	Romeo	Disaster Relief, Not Enough Room, Security vs. Convenience	Illegal Weapon	Security vs. Convenience, Broadsided
20	Delta	Mike	Breaking Bread, Survivor’s Guilt	By Any Means Necessary, Stolen Knives	Café Conundrum
20	Mike	Echo	Café Conundrum	Broadsided, Disaster Relief, Security vs. Convenience	Neighborhood Patrol, Illegal Weapon
10	Lima	Yankee	Ordnance Detection, Neighborhood Patrol	Breaking Bread	Stolen Knives

During this time we emphasized four points. First, we informed students that the same images could be used with different text, so they should be sure to read the text as it might indicate a variant or inject. Second, we told them different decisions could lead to different paths through a storyline, so to explore to learn how the different paths play out and what coaching was associated with the different choices. Third, we asked them to try to make the best choices they could, and to keep playing until they got a good ending (i.e., all indicators on the review screen suggesting GS behavior, though the term “good stranger” was not used in the explanation). We told them, though, that it may not be possible, for a given storyline and inject, to get all good indicators, even though they felt they acted as they felt most appropriate, so that some choices, as in the real world, do not have clear-cut right and wrong answers. We assured them that the study was more interested in seeing their responses as the decisions came up, and for them to do the best they could for the presented situation, as they saw it. Fourth, we told the students that we took some artistic license with characters, settings, clothing, etc. They could let us know when they saw something in the scenes that did not quite fit, but what we most cared about was how they thought about and perceived the different situations.

The next five minutes were used for pre-testing. All students were given the same storyline as a pre-test, though they did not know it was a pre-test. The instructional material took up the bulk of the session. Each student ran through the storylines in the order given to him; the student had a choice which order of storylines to run when there were multiple to a block, but could not revisit storylines in previous blocks. All students followed these instructions. The next five minutes were used for post-testing. All students were given the same storyline as a post-test, though again they did not know that it was a post-test. The final fifteen minutes were given to debrief and feedback. We thanked students, took notes on comments and suggestions they made, and fully explained the study and its goals.

Table 4. Storyline Timing

Storyline	Predicted	Actual
Pre-test	5:00	3:44
Ordnance Detection	5:00	2:45
Security vs. Convenience	5:00	2:56
Disaster Relief	5:00	3:03
Survivor’s Guilt	10:00	4:36
Neighborhood Patrol	5:00	4:51
Breaking Bread	10:00	4:52
Not Enough Room	10:00	5:04
Suspicious Behavior	10:00	5:40
By Any Means Necessary	10:00	5:49
Stolen Knives	10:00	7:06
Illegal Weapon	15:00	8:05
Broadsided	10:00	8:40
Café Conundrum	20:00	10:08
Post-test	5:00	1:26
Average:		5:34

Analyses

As a check to see if the perceived complexity of storylines (their length, variations, and/or number of decision points) matched actual interactions with the storylines, we determined how much time on average students spent during each storyline. Table 4 shows these results. Though the actual times were significantly lower than the predicted times ($p<0.001$ by a one-tailed paired t-test), the times were highly correlated (r=0.86), suggesting that students did require more time to navigate through more complex storylines.

The student data were analyzed in three different ways. First, the raw scores from the pre- and post-test storylines were compared to see if students demonstrated a significant change from pre- to post-test. Second, students' proficiency levels were estimated using Microsoft TrueSkill (Herbrich, Minka, & Graepel, 2006). Third, students' decisions were rated based on cognitive "frames".

Change in Responses

A variety of data were collected related to the students' performance during the trial. At each point where a student made a decision in the storyline, the student's choice was recorded along with the time and a correct, incorrect, or neutral assessment (as determined by our experts) for one or more LOs. A choice could be assessed as correct in one LO and incorrect in another LO. In addition to decision points, some storylines presented students with image annotations. In these storylines students were asked to annotate images with LOs. Storyline authors had previously noted areas on the image that associated with LOs. Students' responses were considered correct if they marked an area with the same LO as the storyline author. Responses were considered incorrect if they did not mark the area with the same LO as the storyline author. Extra LOs marked by students (either inside or outside of annotation areas) did not impact students' scores.

Raw scores were computed for the pre- and post-test storylines by dividing the number of correct evaluations by the total number of evaluations. Students started the training with a high level of competence; the average pre-test score was 84%, while the average post-test score was 96%. Possible raw score values depended on the path that students took through each storyline. This path dependence caused a non-normal distribution of raw scores among the students; therefore, a Wilcoxon signed rank test was performed to confirm that the increase in score was statistically significant. The test confirmed (p<0.02 for a two sided test) that the score increase was significant.

TrueSkill Analysis

TrueSkill is a Bayesian rating system which predicts player performance (e.g., for online games) and estimates proficiency level. It represents each player's proficiency level as a normal distribution parameterized by a mean and a variance, which correspond to the player's proficiency level and the uncertainty in the proficiency level respectively. After a game is played between two players, the players' means and variances are updated based on a win or a loss.

For the IBOLC data, the "games" are played between players and decisions (which branch to take in a storyline or which LO to annotate in an image). The concept is that prior to every evaluation, the mean and variance of the student and the decision point can be used to predict the outcome of the evaluation (i.e., whether or not the student will make the correct decision). Therefore both players and decisions that occur within a storyline are assigned proficiency levels. Since to our knowledge this is a new application of TrueSkill, we assessed the quality of the game outcomes predicted by TrueSkill.

Proficiency levels for every student and decision point were generated using the following procedure. Each student's and decision point's initial proficiency level was set to a default value. All of the evaluations that occurred during the trial were sorted based on the time they occurred. After each evaluation a TrueSkill update was performed on the player proficiency level and the decision point proficiency level.

Systems using TrueSkill typically use a matchmaking system which prefers matches where the predicted outcome is close to 50%. The IBOLC trial did not use a matchmaking system and thus has a wide range of predicted outcomes. This allows for a unique analysis of the accuracy of TrueSkill's predictions. We took each of the approximately 8,000 evaluations from the trial data and binned it according to the predicted chance that the student would make the correct choice at that point. For example, in the trial data there are some 3,500 evaluations which TrueSkill predicted the student would provide the correct answer about 95% of the time. In actuality, for those evaluations, students chose the correct answer just over 96% of the time. Indeed, for all of the TrueSkill predictions, they were in very close agreement with the actual performance of students observed during the trial. Therefore, for this group of students, TrueSkill provides a good model of both student and evaluation proficiency level. It is also important to note that the values predicted by TrueSkill use only information available up until the time of that evaluation. For example, if an evaluation occurred one minute into the trial, the TrueSkill prediction uses only the outcomes of evaluations that occurred before one minute.

Given that the performance of TrueSkill in this domain has now been assessed, the change in student proficiency level over the course of the IBOLC trial may be analyzed. For each LO, the students' proficiency levels and variances at the end of each storyline were averaged together to measure the performance of the class. Figure 2 shows the change in average proficiency level from pre-test to post-test.

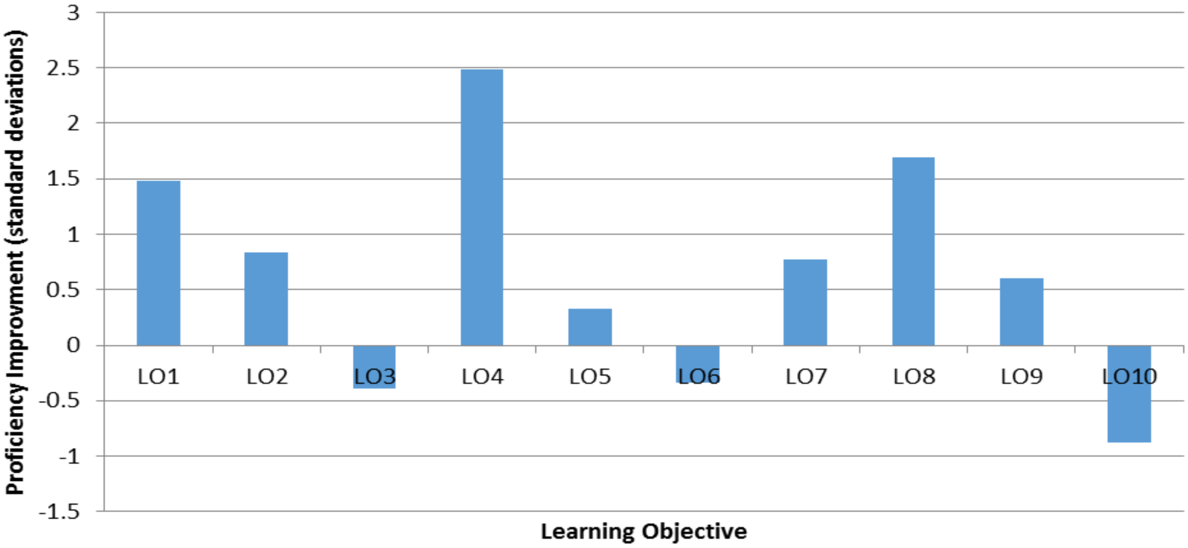


Figure 2. Confidence in Change in Proficiency Level from Pre-test to Post-test for Each LO

Change in proficiency level is shown in terms of the number of standard deviations for that particular LO. A change of two standard deviations indicates a confidence level of about 95% (that there was a pre/post change) while a change of one standard deviation indicates a confidence level of about 68%. Therefore students' proficiency levels increased with a high confidence level in LO1, LO4, and LO8, while changes in proficiency level in other LOs were not statistically significant. These three LOs (1, 4, and 8) in particular, and (others such as LO2, LO7, and LO9 that showed improvement, if not significantly so) make sense to have been influenced by the training, since, as described in Table 1, they involve the reasoning and perceiving that are involved in initiating an encounter, attending to nonverbal cues, taking another person's perspective, appraising the outcome of an encounter, and determining when and how to repair the encounter gone awry. In contrast, LO10, as an example, focuses on tactical cues and may not be best addressed by this training tool.

Frame-based Analysis

We also undertook a reanalysis of these data using the cognitive frames put forth by other SSIM performers (Klein et al., 2014) in their gap analysis. In that work, four identities or "frames" are posited to describe how different individuals approach decisions within storylines. According to that theory, a GS tries to gain trust, understanding the long-term consequences of his or her actions. Meanwhile, a Mission-focused individual makes decisions that optimize mission completion, a Rules and Procedures follower uses established protocol when possible, while an Authoritarian employs his or her authority to gain or coerce compliance.

For this reanalysis, we first needed to re-map all of the decision points as indicating one or more of these four frames. Two coders independently rated each decision point, resulting in initial agreement on 70% of assignments. The two coders then discussed each of the remaining 30% of decisions to resolve their discrepancies. The entire set was then used in the reanalysis.

Figure 3 shows results. Along the *x*-axis is time, indicated by the different storylines that different students experienced (e.g., for the first storyline, some students ran through Illegal Weapon, others Suspicious Behavior, etc.). At each time, the storylines afforded some number of possibilities for students to demonstrate GS, Mission-focused, Rules-based, or Authoritarian frames; these are plotted along the *y*-axis. To calculate the percentages, we summed all instances of a frame (e.g., all times students made GS decisions) then divided by the number of all frames for those decisions. (Note the percentages should not necessarily sum to 100% since the denominator was restricted to choices where that frame was available.) As can be seen from the figure, the majority of decisions taken were GS decisions, with a possible trend upward across storylines. More importantly, there were Mission-focused, Rules-based, and Authoritarian decisions at each point. Recall, as discussed above, that the storylines were developed so that not all decisions would involve GS choices. At each time point there were a mix of types of decisions, however over the course of training the system appeared to lead to more GS decisions in relation to the other frames, as was intended.

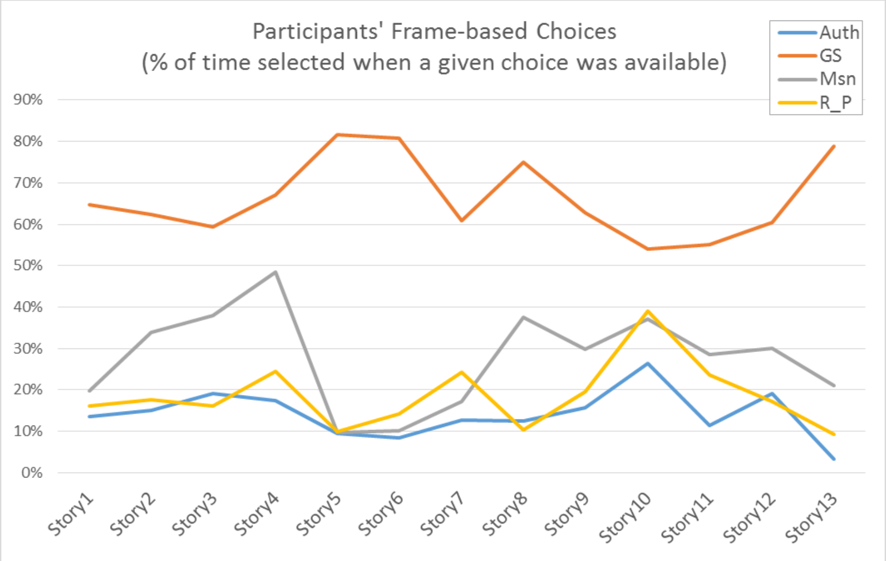


Figure 3. Frame-based Choices across the Experimental Session

CONCLUSIONS

In this study we studied whether or not a training tool that focuses on perceptual and cognitive skills can influence the GS skills that IBOLC students demonstrate. We showed that there were changes in students' behavior and that they often use a GS frame—though, appropriately, not always since not all decisions demand GS skills. A recent independent study (Halverson et al., 2015) confirmed these findings, showing that the training tool decreased students' use of dominance frames and increased GS frame usage relative to a control group.

We are investigating the tool's use in military and non-military domains for similar skills. For instance, we are interested in health-related assessments of social competency (Hubal, 2012); this tool might encourage an individual to exhibit behaviors aligned with underlying interpersonal constructs like inquisitiveness or impulsiveness. Similarly, we see the tool as useful for gauging vigilance and emotional stress control in border control and mobile health environments (Basner & Rubinstein, 2011; Kizakevich et al., 2014). We have continued to refine the tool and plan a number of improvements to address the needs of these applications, including greater control over the sequence of storylines, flexible frame layout, more complex and prioritized annotations, allowance for students to indicate confidence in choices, and optimized display for different delivery platforms.

ACKNOWLEDGEMENTS

The authors thank Ellis Amdur, Scott Flanagan, Kent Halverson, Leon Pryor, Sarah Quinn, John Schmitt, and the SSIM Task Area 1B team for their valuable input to this document. This work was supported by contract W911NF-11-C-0273 from DARPA. The opinions expressed in this document are those of the authors and do not necessarily reflect the view of DARPA.

REFERENCES

Basner, M., & Rubinstein, J. (2011). Fitness for duty: A 3 minute version of the Psychomotor Vigilance Test predicts fatigue related declines in luggage screening performance. *Journal of Occupational and Environmental Medicine*, *53*(10), 1146-1154.

Halverson, K.C., Lucia, L., Horn, Z., Lande, B., Keeney, M., Weil, S., & Diedrich, F. (2015). Evaluation of social skills training approaches: APACTS and MAST. Technical report submitted by Aptima, Inc. to the U.S. Army Research Office, Research Triangle Park, NC under DARPA Contract #W911NF-11-C-0266, March 6, 2015.

Herbrich, R., Minka, T., & Graepel, T. (2006). A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, *19*, 569-576.

Hubal, R. (2012). The imperative for social competency prediction. In S.J. Yang, A.M. Greenberg, & M. Endsley (Eds.), *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction* (pp. 188-195). Springer-Verlag.

Hubal, R., van Lent, M., Wender, J., Lande, B., Flanagan, S., & Quinn, S. (2015). What does it take to train a good stranger. In *Proceedings of Cross-Cultural Decision Making* (pp. 5076-5083). AHFE International.

Kizakevich, P.N., Eckhoff, R., Weger, S., Weeks, A., Brown, J., Bryant, S., Bakalov, V., Zhang, Y., Lyden, J., & Spira, J. (2014). A personal health information toolkit for health intervention research. In B.K. Wiederhold & G. Riva (Eds.), *Annual Review of Cybertherapy and Telemedicine* (pp. 35-39). IOS Press.

Klein, G., Klein, H.A., Lande, B., Borders, J., & Whitaker, J.C. (2014). The good stranger frame for police and military activities. *Proceedings of the Human Factors and Ergonomics Society*, *58*(1), 275-279.

Logan-Terry, A., & Damari, R.R. (2015). Key culture-general interactional skills for military personnel. In *Proceedings of Cross-Cultural Decision Making* (pp. 5092-5099). AHFE International.

Wray, R. (2013). Tailoring culturally-situated simulation for perceptual training. In D.D. Schmorrow & D.M. Nicholson (Eds.), *Advances in Design for Cross-Cultural Activities (Part I)* (pp. 45-54). CRC Press.