# Can Role-Play with Virtual Humans Teach Interpersonal Skills?

**Matthew Jensen Hays,**
**Julia C. Campbell,**
**Matthew A. Trimmer**

**University of Southern California**
**Institute for Creative Technologies**

**Playa Vista, CA**

**hays,campbell,trimmer@ict.usc.edu**

**Joshua C. Poore,**
**Andrea K. Webb**

**Charles Stark**
**Draper Laboratory**

**Cambridge, MA**

**jpoore,awebb@draper.com**

**Teresa K. King**

**Naval Service Training**
**Command**

**Great Lakes, IL**

**teresa.king@navy.mil**

## ABSTRACT

Interpersonal and counseling skills are essential to Officers' ability to lead (Headquarters, Department of the Army, 2006, 2008, 2011). We developed a cognitive framework and an immersive training experience—the Immersive Naval Officer Training System (INOTS)—to help Officers learn and practice these skills (Campbell et al., 2011). INOTS includes up-front instruction about the framework, vignette-based demonstrations of its application, a role-play session with a virtual human to practice the skills, and a guided after-action review (AAR). A critical component of any training effort is the assessment process; we conducted both formative and summative assessments of INOTS. Our formative assessments comprised surveys as well as physiological sensor equipment. Data from these instruments were used to evaluate how engaging the virtual-human based practice session was. We compared these data to a gold standard: a practice session with a live human role-player. We found that the trainees took the virtual-human practice session seriously—and that interacting with the virtual human was just as engaging as was interacting with the live human role-player. Our summative assessments comprised surveys as well as behavioral measures. We used these data to evaluate learning produced by the INOTS experience. In a pretest-posttest design, we found reliable gains in the participants' understanding of and ability to apply interpersonal skills, although the limited practice with the virtual human did not provide additional immediate benefits. This paper details the development of our assessment approaches, the experimental procedures that yielded the data, and our results. We also discuss the implications of our efforts for the future design of assessments and training systems.

## ABOUT THE AUTHORS

**Dr. Matthew Jensen Hays** is a cognitive scientist at the University of Southern California's (USC) Institute for Creative Technologies (ICT). He received his B.S. in psychology from Duke University, and his M.A. and Ph.D. in cognitive psychology from the University of California, Los Angeles. He works to integrate principles of learning and memory with simulations and game-based training systems. He also designs, develops, and conducts formative and summative assessments of these systems.

**Dr. Julia C. Campbell** is a research associate at the ICT. At USC, Julia earned an M.A. in communication and an Ed.D. in educational psychology. Her work focuses on cognitive task analyses and assessment of reaction and self-efficacy.

**Dr. Joshua C. Poore** received his M.A. and Ph.D. in psychology (with emphases in social psychology, statistics, and behavioral neuroscience) from UCLA. His graduate and post-doctoral work (NIH/NINDS) focused on the neural basis for social affiliation, trust, and attachments. At the Charles Stark Draper Laboratory, he works to develop applications for neuroscience and physiology in the interests of the Department of Defense and intelligence community.

**Dr. Andrea K. Webb** received her M.S. and Ph.D. in psychology from the University of Utah. Her areas of expertise are in physiology, quantitative methods, eye tracking, and deception detection. At the Charles Stark Draper Laboratory, she works to solve physiological and behavioral science problems in the national interest.

**Dr. Teresa K. King** is an education research analyst at the Naval Service Training Command. After serving in the Navy for 24 years, she earned an M.A. and Psy.D. in clinical psychology from the Illinois School of Professional Psychology. She previously received an M.S. in education from Troy State University. Her work focuses on research and evaluation of newly integrated interventions and systems.

**Mr. Matthew A. Trimmer** is a project director at the ICT. Mr. Trimmer received his B.S. and M.B.A. degrees from USC's Marshall School of Business in 2002 and 2007, with concentrations in cinema-television and technology commercialization. He has led several mixed-reality- and game-based training efforts at the ICT.

# Can Role-Play with Virtual Humans Teach Interpersonal Skills?

**Matthew Jensen Hays, Julia C. Campbell, Matthew A. Trimmer**
**University of Southern California Institute for Creative Technologies**
**Playa Vista, CA**
**hays,campbell,trimmer@ict.usc.edu**

**Joshua C. Poore, Andrea K. Webb**
**Charles Stark Draper Laboratory**
**Cambridge, MA**
**jpoore,awebb@draper.com**

**Teresa K. King**
**Naval Service Training Command**
**Great Lakes, IL**
**teresa.king@navy.mil**

## TRAINING INTERPERSONAL SKILLS

### Is it Necessary?

The Navy is a complex organization, comprising over 360,000 Officers and Enlisted Sailors. Leadership is an integral part of daily life for most of these Sailors; all but Seaman Recruits (the lowest rank of Enlisted Sailor) have subordinates. As a result, the Navy places a strong emphasis on leadership and devotes a great deal of training to developing it (Wagner, 2010).

A critical component of leadership is being able to interact effectively with subordinates, particularly when they are having difficulties. These difficulties span a wide range of topics. Many military members and their families encounter significant financial distress (DOD, 2006; Simmons, 2008). Another common problem is alcohol abuse and the related health and legal issues that stem from it (e.g., DUI arrests; Stahre, Brewer, Fonseca, & Naimi, 2009). We conducted interviews with Officers to identify additional topics. We found that strained interpersonal relationships—with peers, superiors, and significant others—topped the list.

Each of the above issues can distract a Sailor and interfere with his/her duties. It is critical for Navy leaders to be able to address them in a productive way. Despite the Navy's focus on leadership, Sailors have access to very little formal interpersonal skills training. Some Officers reported that they had been directed to participate in role-play sessions with untrained partners, but this training is unlikely to be effective (Holsbrink-Engels, 2001).

This lack of effective training can be especially problematic for Sailors who earned their commission by attending Officer Candidate School (OCS), Officer Development School (ODS), or the United States Naval Academy. When they become Ensigns at perhaps 22 years of age, they are immediately placed in command of subordinates who may be at least a decade older than they are. These subordinates likely have many more years of Navy and general life experience. They may also have spouses, children, property, and associated concerns about which the young Officer may know very little. With limited prior experience in a leadership role and even less experience counseling subordinates, a lack of interpersonal-skills training can have a significant negative impact.

### What Needs to be Taught?

In response to this training gap, our Institute[1] began to design and develop a way to help young Officers learn and practice interpersonal and counseling skills. The resulting training experience was simultaneously developed for the Navy and Army[2] (Campbell et al., 2011), but the principles of leadership and effective interpersonal communication are not branch-specific. As a result, Army Field Manual (FM) 6-22 was a primary source for the doctrine we integrated into our design.

We supplemented that doctrine with information from a *cognitive task analysis* (CTA) we conducted with several experienced and inexperienced Officers. CTA is an interview technique that helps experts fully explain the concepts, processes, and principles that underlie their decisions and actions in their area of expertise (Clark & Estes, 1996). From the CTA, there emerged two categories of situations in which Sailors need to rely on interpersonal skills with their subordinates. The

[1] Disclosure note: The ICT is a University-affiliated Army Research Center at USC. The majority of the ICT's work consists of basic research as well as the design, development, and refinement of research prototypes. The ICT explicitly does not serve as a production contractor and all deliverables of government-sponsored projects (like INOTS) have full government use rights.

[2] The Army version of INOTS was previously known as VOLT. It is now known as ELITE.

first category is when a subordinate has a job-performance problem. The second is when a subordinate informs the Officer that s/he is having difficulty dealing with a personal issue. We worked with the Subject Matter Experts (SMEs) and drew on the content of FM 6-22 to generate strategies that should be used in these two situations.

One strategy (I-CARE) should be used when a subordinate is exhibiting a performance problem. The Officer should perform the following steps:

Initiate communication; state the performance issue
Check for underlying causes
Ask questions and verify information
Respond with a course of action
Evaluate by following-up

The other strategy (LiSA-CARE) should be used when a subordinate comes to the Officer with a difficult personal issue. The Officer should perform the following steps:

Listen without interruption
Summarize in a neutral style
Ask for confirmation of your understanding

These steps should be followed by the CARE procedure described just above. (For more on I-CARE and LiSA-CARE, including descriptions of the sub-steps and associated skills, please see Campbell et al., 2011.)

## INOTS OVERVIEW

Having identified the strategies that lead to effective interpersonal interactions, we needed to establish how to use them. Existing training and FM 6-22 explain what Officers should do in the above situations, but usually at a very abstract level. As a result, Officers may know that they need to "initiate communication," but not know *how*. We therefore relied heavily on the CTA to create a training experience that made these skills concrete and provided a structured opportunity to practice applying them.

### The INOTS Experience

The INOTS experience totals approximately three hours. Ideally, it is spread over at least two days, with a homework assignment on the first night. It is instructor-led and can accommodate up to 50 students at a time.

Training begins when the instructor introduces the students to basic counseling and interpersonal skills. Students receive a handout that provides the I-CARE/LiSA-CARE strategy set as an organizational framework for those skills. The instructor also connects each step and sub-step to what a leader might actually say and do.

**Figure 1. The Virtual Human Role-Player**

Students then receive a homework assignment in which they review real-world case studies. For each, the students must generate examples of how to apply each skill in the framework, specifying what they might say and do to address a subordinate's performance or personal problems.

The next day, the instructor facilitates a student-led review of the homework. The instructor then shows the class video vignettes that demonstrate the interpersonal skills being used correctly and incorrectly. An additional video reviews the relationship between the I-CARE/LiSA-CARE framework and each correct and incorrect action taken by the characters in the vignettes. Instructors may tailor, to some extent, the training content and delivery, and they are encouraged to provide examples from their own background to demonstrate the skills.

Next, the class participates in a semi-structured role-play exercise between a single student and a life-sized virtual subordinate (Figure 1). The interaction is driven by a turn-based branching narrative; after the subordinate responds, the student chooses to say one of three pre-scripted responses. At each of these decision

points, the rest of the class uses hand-held voting devices ("clickers") to indicate which response they prefer (independent of the decision made by the role-playing student). The system tracks each vote for each student. Each vote is scored (correct, incorrect, or mixed) based on links to the I-CARE/LiSA-CARE framework. From these data, INOTS builds cognitive models of each student in the class.

When the scenario ends, the system uses the student-model data to assist the instructor in conducting an after-action review (AAR). The system provides the instructor with per-student, per-decision, and per-skill data. It also generates talking points for decisions on which many students voted incorrectly, or skills associated with mostly incorrect votes (i.e., topics that most students appeared to misunderstand).

The instructor can then select a new scenario and conduct another role-play session. At the time the present paper was submitted, INOTS featured two scenarios. *Pushing the Line* focuses on a performance problem: Gunner's Mate Second Class (GM2) Cabrillo shoves one of his subordinates during an argument. *Gunner's Troubles* focuses on a personal problem: GM2 Cabrillo learns that his wife may have been unfaithful.

**Assessing INOTS**

The INOTS training approach is based on empirical research in the areas of instructional design and cognitive psychology (Campbell et al., 2011). Each training topic establishes context for the content that follows, which substantially improves later recall (Bransford & Johnson, 1972). The first topic of instruction is the importance of interpersonal skills, followed by the definition of the skills themselves, followed by the I-CARE/LiSA-CARE framework. INOTS also combines direct instruction (Schwartz & Bransford, 1998) and interactivity (Evans & Gibbons, 2006) in order to harness the benefits of both approaches. The vignettes serve as enriched demonstrations, which improve memory for content (Cooper & Sweller, 1987). The vignettes and the role-play scenarios both feature strong narrative, which also boosts learning (Fernald, 1989). Finally, the AAR serves to delay feedback until the trainees have had a chance to further process the information—a powerful educational technique (Gaynor, 1981).

However, it is essential to verify whether the design is effective and to explore how it might be improved. To that end, we conducted two experiments designed to evaluate the INOTS experience. Experiment 1 was a formative evaluation of the role-play session with the virtual human. Experiment 2 was a summative evaluation of the complete INOTS experience: the I-CARE/LiSA-CARE framework and instruction, the vignettes, and the role-play session with the virtual human.

## EXPERIMENT 1

Experiment 1 was designed to explore how well the virtual human functioned as a role-player. We chose to focus on the interaction with the virtual human because, to our knowledge, INOTS is the first implementation of a full-size virtual human in a classroom environment[1]. Although full-size virtual humans have been used for one-on-one role-play-based instruction (Saleh, 2010), it is important to examine how they can contribute on a larger scale with greater throughput.

We compared role-play with the virtual human to a gold standard: role-play with a live human in the same scenario. We used surveys to evaluate the participants' responses to the role-players, their perception of the realism of the virtual human and the social characteristics of the interaction, and the experience as a whole. Because self-report measures are not always reliable, we also collected physiological data during the role-play sessions.

**Method**

**Participants**

The participants were 21 members of the Naval Reserve Officer Training Corps (NROTC) at the University of California, Los Angeles. They were mostly male (81%) and their average age was 19.8 (*SD* = 1.57).

**Design**

The participants were randomly assigned to one of two between-subjects conditions that determined with which *role-player* they would meet: virtual or live. The *virtual* role-player operated as described above. The *live* role-player sat in the room with the participant at approximately the same "distance" as the rear-projected virtual human. The live role-player followed the same branching narrative as did the virtual role-player. That is, the choices made by the participant elicited the same responses from either role-player.

[1] To be clear, whereas the typical INOTS experience includes up-front instruction and demonstrations, the participants' experience in Experiment 1 was limited to the interaction with the role-player.

**Measures**

We used self-report surveys and physiological sensor equipment to examine how the role-players affected the participants. The participants responded to the survey items on a Likert scale (1 = low, 7 = high). The survey asked how engaging the experience was, how natural the participants found different aspects of the role-player to be, and how evocative the participants found different aspects of the role-player's acting to be.

We used two physiological sensors to supplement the self-report data. One sensor monitored the participants' heart rate (measured as the *inter-beat interval*). Another sensor measured the electrical conductance of the participants' skin—the *galvanic skin response* (GSR)—which varied with the participants' perspiration. The physiological data associated with each sensor corresponds to the participants' emotional state.

**Procedure**

The participants arrived at our Institute and provided consent to participate. Next, the physiological sensors were applied to their bodies. The participants then met with their assigned role-player (virtual or live) in the *Pushing the Line* scenario described above. After the meeting, the sensors were removed and the participants completed the survey described above.

**Results and Discussion**

Tables 1a and 1b present the results gathered from the survey items.

| | Engaging | | Natural: Body Language | | Natural: Spoken Reponses | | * Natural: Option Wording | |
|---|---|---|---|---|---|---|---|---|
| Role-Player | M | SE | M | SE | M | SE | M | SE |
| Live | 5.10 | .50 | 6.00 | .36 | 5.80 | .41 | 5.00 | .38 |
| Virtual | 5.36 | .48 | 6.27 | .34 | 6.00 | .40 | 6.09 | .36 |

**Table 1a. Self-Report Data from Experiment 1.**
Note: *M* indicates mean value. *SE* indicates standard error of the mean.

| | Evocative: Body Language | | Evocative: Speech | | Evocative: Vocal Intonation | | Evocative: Facial Expression | | * Evocative: Gaze | |
|---|---|---|---|---|---|---|---|---|---|---|
| Role-Player | M | SE | M | SE | M | SE | M | SE | M | SE |
| Live | 4.90 | .42 | 5.10 | .32 | 5.20 | .35 | 5.50 | .35 | 5.60 | .38 |
| Virtual | 5.36 | .40 | 5.27 | .30 | 5.73 | .34 | 5.45 | .33 | 4.36 | .36 |

**Table 1b. Self-Report Data from Experiment 1 (Continued).**
Note: *M* indicates mean value. *SE* indicates standard error of the mean.

**Self-Report Measures**

As can be seen in Table 1a, there was not a reliable main effect of role-player on the participants' ratings of engagement: $t(19) = .38$, $p = .71$. Although the live human was in the room with the participants, interacting with the virtual human appears to have been approximately as engaging.

There was not a reliable main effect of role-player on the participants' ratings of how natural the role-player's body language was: $t(19) = .55$, $p = .59$. There was not a reliable main effect of role-player on the participants' ratings of how natural the role-player's spoken responses were: $t(19) = .36$, $p = .73$. The various

components of the virtual human appear to be effectively conveying emotion. This finding is consistent with the engagement results above.

There was a marginally reliable main effect of role-player on the participants' ratings of how natural the wording of the conversation options was: $t(19) = 2.10$, $p = .07$. Because the wording of the options was identical for both role-players, this result may indicate that the participants felt it was more natural to have an interaction constrained by options with the virtual human than with the live human. If this interpretation is accurate, the limitation of the interaction to the branching narrative options may not come at the cost of immersion in the experience. This interpretation is consistent with the negligible difference in self-rated engagement with the virtual versus live role-player.

Table 1b presents the mean responses (and the standard errors of the means) to the remaining survey items. There was not a reliable main effect of role-player on the participants' ratings of how evocative they found the role-player's body language to be: $t(19) = .80$, $p = .43$. There was not a reliable main effect of role-player on the participants' ratings of how evocative they found the role-player's speech to be: $t(19) = .39$, $p = .70$. There was not a reliable main effect of role-player on the participants' ratings of how evocative they found the role-player's vocal intonation to be: $t(19) = 1.08$, $p = .29$. There was not a reliable main effect of role-player on the participants' ratings of how evocative they found the role-player's facial expressions to be: $t(19) = .09$, $p = .93$. These results suggest that the virtual human elicited emotion in its counterpart approximately as effectively as the live human.

There was a reliable main effect of role-player on the participants' ratings of how evocative they found the role-player's gaze to be: $t(19) = 2.38$, $p = .02$. This finding is unsurprising; the INOTS system does not yet support the capability for the virtual human to accurately make and sustain eye contact with the trainee. The live human role-player, of course, executes this task naturally. Nevertheless, the virtual human's inferior gaze appeared not to decrease the participants' immersion in the experience.

Taken together, the self-report data suggest that role-play with the virtual human was roughly comparable to role-play with the live human.

**Physiological Measures**

A software error prevented physiological data from being recorded for one participant. That participant's data are omitted from all reported analyses.

The physiological sensor data were first checked to ensure that they were reliably different from baseline while the participants were interacting with the role-players. The participants' heart rate reliably responded to the role-play session with the virtual human: $t(8) = 2.74$, $p < .05$. The participants' heart rate reliably responded to the role-play session with the live human: $t(10) = 3.79$, $p < .01$. The participants' GSR reliably responded to the role-play session with the virtual human: $t(8) = 2.86$, $p < .05$. The participants GSR reliably responded to the live human: $t(10) = 4.85$, $p < .01$. These verification analyses indicated that the sensors provided reliable data and that these data were suitable for comparing the virtual human to the live human.

We therefore compared the two role-players' effects on the physiological data. There was not a reliable main effect of role-player on heart rate (less baseline): $t(18) = 1.23$, $p = .23$. There was not a reliable main effect of role-player on GSR (less baseline): $t(18) = .17$, $p = .87$. Thus, the physiological data corroborated the self-report data. As far as the participants' emotional experience of the role-play was concerned, there was not a substantial measurable difference between the virtual human and the live human.

The physiological data parallel the self-report findings. Together, they suggest that the virtual and live role-players affect their counterparts in very similar ways.

## EXPERIMENT 2

Experiment 2 was designed to examine the overall training impact of the entire INOTS experience. We used a pretest-posttest design to evaluate learning at the first three levels of Bloom's taxonomy of cognitive skills (Anderson & Krathwohl, 2001). The lowest level of Bloom's taxonomy is *knowledge*, which is memory for previously learned information. We measured knowledge with questions that asked about components of the I-CARE/LiSA-CARE framework. For example:

> Please indicate (yes/no) whether each behavior suggests that a person is listening with the goal of understanding another person's problems.
>
> ___ When the speaker finishes describing the problem, the listener suggests a reasonable solution.
>
> ___ The listener has a neutral expression on his face.
>
> ___ The listener occasionally interrupts the speaker in order to help the listener focus his thoughts.
>
> ___ The listener summarizes what was said to make sure that he understands the problem.

The next two levels of Bloom's taxonomy are *comprehension* and *application*. A learner at these levels is able to transfer and apply knowledge to novel situations. We measured comprehension/application with *situational judgment test* (SJT) prompts. These prompts presented a scenario and then asked the learner to rate the quality of possible responses. Each response was linked to one or more of the components of the I-CARE/LiSA-CARE framework. An example SJT item:

> You ask a Sailor about his excessive phone use during work hours. The Sailor responds by saying that his time on the phone is spent helping other workers accomplish their tasks. You ask the Sailor questions to get more information, such as: Who are the other Sailors? What tasks are involved? When did they ask you to help them with the tasks? Please rate (not appropriate, somewhat appropriate, very appropriate) the following actions you could take after having gathered this information.
>
> ___ Tell the Sailor you will get back to him.
>
> ___ Tell the Sailor that he needs to let the other Sailors know that they should be seeking assistance from their immediate supervisor instead.
>
> ___ Set up a meeting with the Sailor to discuss this information.

## Method

### Participants

The participants were 142 students in ODS Newport. They were mostly male (66%) and mostly white (70%). The participants' average age was 29.1 (*SD* = 5.20); 38% of them held bachelor's degrees, 24% held master's degrees, and 38% held doctoral degrees.

### Design

The experiment used a between-subjects design with a single independent variable: *training type*. There were three levels of this variable: practice, no-practice, and control. The participants in the *practice* group completed the entire INOTS experience, including two scenarios with the virtual human. The participants in the *no-practice* group also completed every aspect of the INOTS experience—except the practice role-play session with the virtual human. The participants in the *control* group did not complete any of the INOTS experience (up-front instruction, demonstrations, homework, or role-play exercise). Instead, they received Navy-mandated leadership training.

### Procedure

The participants provided consent to participate and then completed the pretest. The pretest included several knowledge and comprehension/application questions (described above). It also included several self-assessment items (e.g., "Please rate your confidence in your current ability to listen, with the goal of understanding, to help someone resolve personal issues").

Two days later, the participants were provided the homework assignment and were asked to complete it that evening. One day after that, the participants in the practice and no-practice conditions completed the INOTS experience as described above. The participants in the practice condition also participated in two scenarios with the virtual human. Meanwhile, the participants in the control group completed unrelated coursework.

One day later, all of the participants completed the posttest. The posttest included the knowledge, comprehension/application, and self-assessment questions from the pretest.

## Results and Discussion

Table 2 presents the data from the three types of items on the pretest and posttest. Three participants failed to follow instructions during the experimental procedure. Their data are omitted from all reported analyses.

### Knowledge

As can be seen in Table 2, there was a reliable increase in the participants' knowledge-question scores from pretest to posttest: $F(1, 136) = 11.53$, $p = .001$. This increase did not differ across conditions: $F(2, 136) < 1$, *ns*. Thus, the increase in knowledge for all three groups was roughly equivalent. Because all three conditions received doctrine-based instruction about interpersonal skills and leadership, it is unsurprising that a measure of knowledge did not detect reliable differences among them. INOTS is not designed to provide additional information, but rather to help trainees organize and understand how to *apply* that information.

### Comprehension/Application

As can be seen in Table 2, there was a reliable increase in the participants' comprehension/application-question scores from pretest to posttest: $F(1, 136) = 44.38$, $p < .001$. This increase differed across conditions; there was a reliable effect of training type: $F(2, 136) = 3.48$, $p = .033$. However, the improvements in the practice condition and no-practice condition were not reliably different ($p = .11$). The virtual human appears not to have made a detectable difference in the participants' ability to apply what they learned.

This lack of difference could have been due to the limited amount of practice; the total duration of the two

meetings with the virtual human totaled approximately 15 minutes. Alternatively, perhaps the effects of the role-play session are simply not detectable on an immediate posttest—a common pattern in the practice-effect literature (e.g., Roediger & Karpicke, 2006). Finally, based on our observations of the training sessions, the instructors for the no-practice condition were more effective than the instructor for the practice condition (please see the Limitations section, below). Any one of these issues may have obscured the benefit of the virtual-human role-play session.

Because the pretest-posttest improvements in the two experimental conditions were not reliably different, we combined them and compared them to the improvement in the control condition. Improvement on the SJT in the experimental conditions was reliably greater than improvement in the control condition: $F(1, 137) = 4.28$, $p = .04$. This result suggests that the INOTS instructional approach was effective in improving the participants' learning at the comprehension/application level of Bloom's taxonomy.

| | Knowledge | | | | Comprehension/ Application | | | | Self-Rated Confidence | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pretest | | Posttest | | Pretest | | Posttest | | Pretest | | Posttest | |
| Training Type | M | SE | M | SE | M | SE | M | SE | M | SE | M | SE |
| Practice | .66 | .01 | .71 | .01 | .56 | .02 | .63 | .02 | 6.13 | .08 | 6.42 | .06 |
| No-Practice | .67 | .02 | .71 | .01 | .54 | .02 | .65 | .02 | 6.19 | .09 | 6.43 | .08 |
| Control | .70 | .01 | .71 | .01 | .58 | .02 | .61 | .02 | 6.06 | .09 | 6.45 | .07 |

**Table 2. Pretest-Posttest Data (Means and Standard Errors) from Experiment 2.**

As discussed above, standard doctrine and instruction about interpersonal skills focuses on what the skills are. INOTS focuses also on *when* and *how* to use those skills. The results of Experiment 2 suggest that this approach was successful. The improvement in interpersonal-skills *knowledge* was roughly equivalent in all three conditions, but the participants in the INOTS conditions became differentially better able to *apply* that knowledge.

**Self-Assessment**

As can be seen in Table 2, there was a reliable increase in the participants' confidence in their ability from pretest to posttest: $F(1, 136) = 54.56$, $p < .001$. This increase did not differ across conditions: $F(2, 136) = 1.10$, $p = .335$. Thus, the increase in confidence for all three groups was roughly equivalent. Because the participants' knowledge and comprehension/application scores also increased from pretest to posttest, as well, this result appears to indicate that the participants' estimates of their ability were well calibrated with their actual ability.

Unfortunately, the scale of the participants' ratings makes this interpretation difficult to support. On the pretest and posttest, the participants' average ratings of their own ability were more than 85% of the maximum value (i.e., at least 6 out of 7). However, their knowledge and comprehension/application scores on both tests were considerably lower. Further, their self-ratings are not consistent with what the SMEs told us about how underprepared even seasoned Officers may be to help struggling subordinates. Thus, although the above results suggest that INOTS is an effective trainer, it does not address the participants' apparent overconfidence in their interpersonal skills and counseling abilities.

**Limitations**

There were several instances in which the logistics of Experiment 2 detracted from its empirical rigor. First, we were restricted to conducting the study using pre-existing Officer Training Command classes. It was therefore impossible to randomly assign the participants to conditions.

More problematically, we were also unable to counterbalance the instructors with the conditions; one instructor taught the practice group and other instructors taught the no-practice groups. This confounding variable may have affected our results. Compared to the instructor for the practice condition, the instructors for the no-practice condition provided better explanations for the I-CARE/LiSA-CARE

framework and facilitated more discussion among the students about using the skills. The instructors for the no-practice condition also provided additional concrete examples (from their own experiences as Officers) of how to use the skills. On the contrary, the instructor for the practice condition did not accurately describe the skills and neglected to discuss the examples in the handout. To the extent that these differences affected our measures of comprehension/application, the true effect of the virtual human role-play session cannot be perfectly determined from the present data. Thus, although instructor facilitation adds great flexibility and educational value to INOTS, it has also presented the greatest challenge to this and our other summative assessment efforts.

Our measures of learning, too, require scrutiny. SJTs, for example, are well established ways of evaluating training (Legree & Psotka, 2006). The SJT items we created were vetted by SMEs, and the correct responses were provided independently by other SMEs. However, our SJT has not yet been subjected to any psychometric analyses. As a result, the items are not yet optimally constructed to measure learning. For example, two items were answered correctly on the pretest and posttest by all of the participants; the items were too easy. Such items deflated pretest-posttest gain, inflated scores, added statistical noise, and impaired our ability to detect between-group differences.

A related potential limitation of our experimental design is that the pretest and posttest versions of the SJT are identical. It could be argued that the participants learned from taking the test itself. In constructing the test, we used several established techniques to reduce this likelihood (Asher, 2007). We also chose to use an SJT in part because its answers are ratings (rather than selections of unique and potentially informative answers, as in multiple-choice tests). The differential improvement of the INOTS groups versus the control group suggests that we were at least somewhat successful in our efforts, but multiple counterbalanced versions of the SJT would be more empirically sound.

## DISCUSSION AND FUTURE DIRECTIONS

In two experiments, we demonstrated that a virtual-human-based training experience can help prepare Navy Officers to counsel subordinates. In large part, this success is due to the solid foundation of cognitive psychology and educational research that went into the design.

### Iterative Design is Critical

**Discussion-Driven Improvement**

This is not to say that INOTS cannot be improved. Indeed, as development progressed and we began to use INOTS with actual Sailors, we found that the initial design required revision. For example, the instructors reported that many students in the classroom failed to vote with their clickers during the first role-play session. According to the instructors, the students were so interested in how Cabrillo responded, moved, and spoke that they forgot to participate in the training activity. In future efforts, we will consider adding an initial demonstration scenario so that the novelty of the virtual human does not compete with the primary training goals.

We could not have known to build an introductory scenario without exposing users to the unfinished prototype training experience and speaking to them and the instructors about what worked well and poorly. We also supplemented these discussions with formative assessments (e.g., Experiment 1). The result is an iterative design process that ensures the training experience is constantly improving.

**Data-Driven Improvement**

We can also use the data we have collected to further this iterative design process. For example, we intend to review the physiological data from Experiment 1 to evaluate the virtual human's response quality. Across participants, if some utterances elicited strong emotional responses, we may be able to determine how to improve other utterances. On the other hand, if there are some utterances that caused the participants to become less engaged, we can revise them in a directed way.

We also intend to more deeply examine the pretest-posttest data from Experiment 2. Rather than a by-participant or by-item analysis, a by-component analysis would reveal whether there were any I-CARE/LiSA-CARE framework components on which participants generally failed to improve—or on which they developed misunderstandings. Similar analyses of the participants' clicker votes should allow us to examine the up-front instruction, vignettes, and the virtual human role-play exercise to make sure that the training is as effective as possible.

### Beyond Iteration

Beyond iterative changes, there may be other ways to alter the INOTS experience so that it harnesses yet more principles of cognitive psychology or instructional design. For example, recent research indicates that classrooms that use clickers can benefit in surprising ways from peer discussion (Smith et al., 2009). After

voting, if each student discusses his/her choice with one other student in the classroom and then votes again, the proportion of correct votes increases dramatically. Part of this effect, of course, is that students who selected the correct response convince students who selected the incorrect response to change their vote. But pairs of students who *both* answered incorrectly on the first vote also tend to converge on the correct response by eliminating each other's misconceptions (Smith et al., 2009). We are considering integrating a time to pause and discuss during each interaction. Although that may make the role-playing student's interaction less authentic, the potential pedagogical benefit for the rest of the class is more important. Ultimately, our goal is not to create a conversation simulator, but to train interpersonal skills as well as possible.

## CONCLUDING COMMENT

Interpersonal skills are critical for Naval Officers—and Warfighters in general—to be able to understand and also to apply. We supplemented doctrine with an instructor-led training experience that produced reliable learning gains. Adding a virtual-human-based practice environment to that training experience did not improve scores on an immediate posttest, but the virtual human interaction was as engaging and compelling as the same interaction with a live human role-player. Unlike a live human, however, a virtual human provides a cost-effective and consistent training protocol, with the added advantage of built-in assessment tools and instructor support. In this way, INOTS better meets the training needs of young Officers about to assume their first command.

We attribute the success of our efforts to theory-driven design and data-driven iterative refinement. We based our instructional approach on the principles of cognitive psychology and education research. We constructed the instructional materials and software so that every activity—down to each student's vote—can be mapped back to an identified interpersonal skill. Using this framework helps instructors effectively manage the training and also guides us in our efforts to improve it. We encourage the designers of other training systems to implement this structured approach, which we believe provides the best chance of discovering ways to support our Warfighters.

## ACKNOWLEDGEMENTS

The work discussed here was sponsored by the Office of Naval Research and Naval Service Training Command (DOD Contract #W911NF-04-D-0005). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. We thank the following people for their support in executing the above experiments: Dr. Ray S. Perez, Mr. John Drake, and Ms. Laura Major.

## REFERENCES

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational outcomes* (Complete ed.). New York: Longman.

Asher, H. (2007). *Polling and the public: What every citizen should know* (7th ed.). Washington: CQ Press.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Campbell, J. C., Hays, M. J., Core, M., Birch, M., Bosack, M., & Clark, R. E. (2011). *Interpersonal and leadership skills: Using virtual humans to teach new Officers.* Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL.

Clark, R. E., & Estes, F. (1996). Cognitive task analysis. *International Journal of Educational Research, 25*, 403-417.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347-362.

DOD. (2006). Report on predatory lending practices directed at members of the armed forces and their dependents.

Evans, C., & Gibbons, N. J. (2006). The interactivity effect in multimedia learning. *Computers & Education, 49*, 1147-1160.

Fernald, L. D. (1989). Tales in a textbook: Learning in the traditional and narrative modes. *Teaching of Psychology, 16*, 121-124.

Gaynor, P. (1981). The effect of feedback delay on retention of computer-based mathematical material. *Journal of Computer-Based Insruction, 8*, 28-34.

Holsbrink-Engels, G. A. (2001). Using a computer learning environment for initial training in dealing with social-communicative problems. *British Journal of Educational Technology, 32*, 53-67.

Legree, P., & Psotka, J. (2006). *Refining situational judgment test methods.* Paper presented at the Proceedings of the 25th Army Science Conference, Orlando, FL.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced laerning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.

Saleh, N. (2010). The value of virtual patients in medical education. *Annals of Behavioral Science and Medical Education, 16*, 29-31.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*, 475-522.

Simmons, D. (2008). A need to implement personal financial education as part of professional military education: Marine Corps Command and Staff College.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*, 122-124.

Stahre, M. A., Brewer, R. D., Fonseca, V. P., & Naimi, T. S. (2009). Binge drinking among U. S. active-duty military personnel. *American Journal of Preventive Medicine, 36*, 208-217.

Wagner, W. (2010) Navy Total Force. (April 2010 ed.).