An Empirical Evaluation of the ShadowBox™ Training Method

Gary KLEIN[a], Joseph BORDERS[a], Corinne WRIGHT[a] and Emily NEWSOME[a]
[a]*MacroCognition LLC*

ABSTRACT
The ShadowBox training method was evaluated in two studies aimed at helping warfighters gain skills in working with civilians. The first study provided three hours of training to Marines at Camp Pendleton and Camp Lejeune, and improved performance by 28% compared to a control group. The second study provided an hour of training, administered via Android tablet, to soldiers at Ft. Benning and improved performance by 21%. These results, both statistically significant, suggest that ShadowBox training may be a useful way to develop perceptual and cognitive skills in the military.

KEYWORDS
Expertise; Military; Decision Making; Learning and Training

INTRODUCTION

The ShadowBox training method is an instructional strategy that enables trainees to see the world through the eyes of experts — without the experts being there (Klein, Hintze & Saab, 2013). Therefore, it avoids the bottleneck of limited access to Subject Matter Experts (SMEs), and can scale up to simultaneously train large numbers of warfighters.

ShadowBox Training Method

ShadowBox training is a scenario-based method that requires trainees to respond to decision points inserted within a scenario. Consistent with other scenario-based training methods such as Situational Judgment Tests (SJTs; McDaniel & Nguyen, 2001) and Tactical Decision Games (TDGs; Schmitt, 1994), decision points present challenges that require the trainee to choose between alternative actions. Unlike other training programs, ShadowBox also uses Cognitive Task Analysis (CTA; Crandall, Klein & Hoffman, 2006) principles to target cognitive skills and abilities. In addition to action-based questions, trainees are asked to prioritize alternative goals, select alternative cues to monitor, and identify types of information to seek. Typically, a decision point will present 3-5 alternatives. The trainees rank these (e.g., most preferred action, 2nd most preferred) and provide a rationale for their ranking.

ShadowBox training also expands on SJT and TDG methodologies by allowing trainees to compare their responses and rationale to those of a panel of experts. These experts have been polled in advance, given the same scenario and the same decision points, and have provided their rankings and their rationale. The developers synthesize the responses of the panel of experts and represent these synthesized responses for the trainees. By collecting and synthesizing the expert rankings and rationale on the front end, ShadowBox training can function without the need for instructors and facilitators to be present during the training session. Here, expertise is being captured and conveyed in a functional way, moving beyond simple rules and procedures.

Each trainee compares his/her rankings and rationale to the synthesized feedback from the experts. It is at this moment that the trainee discovers how the expert would respond. The feedback from the panel of experts specifies the experts' ranking but also the cues they were noticing and thinking about, and the reasons for the rankings. The discovery is grounded in the specific scenario, rather than in generalities. And the trainee cannot say, "Oh, I was thinking the same thing," because if it wasn't written down as part of the trainee's rationale, it doesn't count. Finally, the trainee consolidates the lessons learned by writing what he/she wants to take away: What did the experts choose and explain that the trainee finds new and valuable?

Strategic Social Interactions Modules (SSIM)

The Defense Advanced Research Projects Agency (DARPA) became interested in exploring the ShadowBox method further because of its potential for providing cognitive training that scales up for large numbers of trainees. The DARPA program, Strategic Social Interaction Modules (SSIM), was a 42-month effort that sought to help military and police become "good strangers" — building the social skills for having encounters with civilians that promote cooperation and voluntary compliance rather than coercion and resentment. In the past, the Army and Marine Corps have established semi-realistic environments (e.g., Afghan Village) equipped with role players to train combat etiquette and to help warfighters acquire social skills before rotating into combat zones. These environments appear to be useful, but they are expensive to create and are not generalizable to

multiple cultures. SSIM, nicknamed the "Good Strangers" (GS) program, recognized these constraints and explored different strategies for generating cost effective social skills training. As part of the GS program, we performed several ShadowBox training tasks, including: 1) configuring scenarios to train warfighters how to have more effective interactions with civilians; 2) evaluating the ShadowBox training; and 3) developing a software application of the ShadowBox method.

Good Stranger Mindset

We selected an unconventional type of training objective for our ShadowBox project. Instead of trying to teach specific social skills and abilities, we elected to try to change the warfighters' professional identity – their mindset. As part of the DARPA SSIM program, G. Klein, H. A. Klein, Lande, Borders & Whitacre (2014a; 2014b) conducted a CTA study of 41 police officers and warfighters who were identified by supervisors and colleagues as GSs. The police officers were included because the military believed it had a lot to learn from the police about being a GS, given that civilian interactions are central to their work. In contrast, the military has often treated civilian encounters as peripheral to the mission of succeeding in combat against enemy forces.

The CTA study attempted to identify the knowledge, skills and abilities that go into working effectively with civilian populations. This study found that the GSs had a different mindset than other police and military personnel. They sought to build trust during their encounters with civilians without making unnecessary risks to their security. This mindset shaped how they conducted themselves, and encouraged them to gain rapport, to take the perspective of the civilians, to try to get voluntary instead of coercive compliance, and to de-escalate conflicts where possible.

We speculated that the training for cooperative civilian encounters could be more efficient and effective if it tried to impart a GS mindset of seeking to build trust where appropriate, instead of trying to teach specific skills or abilities or impart specific knowledge. We wanted to use ShadowBox training to change the warfighters' professional identity so that, in addition to the ways that they usually frame situations (e.g., maintain security, accomplish the mission, follow regulations), they would also have a GS mindset.

Mobile Application ShadowBox Training (MAST)

Military personnel have limited time to participate and engage in training exercises. That is why there is a value in being able to make training systems work on mobile technologies so that warfighters can train independently during their downtime. Our SSIM ShadowBox effort included a task to develop a mobile training application of the ShadowBox method that would operate on Android tablet devices. Similar to the paper-based training, MAST includes challenging scenarios, decision points, and expert feedback as part of the trainee experience. MAST goes beyond the pen-and-paper experience in several ways: the use of touch screen, presentation of scenarios using images, video and audio, and a timer feature. Most importantly, MAST was designed to train at the individual level so it could scale up to teach large numbers of trainees.

Effectiveness of ShadowBox Training Method

Neil Hintze, a Battalion Chief with the New York Fire Department (now retired from the FDNY) developed the ShadowBox method. Hintze (2008) evaluated the method by comparing the responses of trainees to those of a control group and found that the rankings of the group receiving ShadowBox training were 18% closer to the panel of experts than the control group, a difference that was significant beyond the .01 level.

One of our tasks on the SSIM program was to evaluate whether ShadowBox training could be effective in the military context. Our goal was to train warfighters to acquire a GS mindset, measured by their responses to the training scenarios. Our evaluation was different from the one Hintze had performed. He emphasized facilitated discussions as part of the ShadowBox experience. But if the method is to scale up, it will have to do so without any discussions. The Expert Feedback conditions (participants receiving SME rankings and rationale) used in the following experiments were not allowed to engage in any discussions. We compared the Expert Feedback group to a control group that did not receive such feedback. We also wanted to know how much is lost by restricting the facilitated discussions, so we also ran a few pilot training sessions that did include facilitated discussions. The purpose of this paper is to describe how we conducted the training and evaluation, and to present the results of the evaluation study of ShadowBox.

EXPERIMENT 1

METHOD

Participants

We collected data at two United States Marine sites. Commissioned and non-commissioned US Marines at Camp Pendleton, CA and Camp Lejeune, NC (N = 59) participated in the ShadowBox experiment. All participants were male.

Materials

We created four ShadowBox scenarios based on incidents identified during the CTA interviews as well as several previously designed scenarios reflecting the GS mindset. We interspersed three or four decision points within each scenario. The decision points presented different actions that could be taken, different priorities that could be emphasized, or different cues that could be monitored. The participants ranked their options for each decision point. The scenarios included a range of common kinetic and non-kinetic situations. (A) *Stolen Knives* is about managing cultural differences in a foreign dining facility in the Marshallese Islands. (B) *Film at 11:00* placed the trainee in an uncertain warzone, unable to discern friendly and enemy territory, and having to maintain security and rapport with locals. (C) *Adad's Cafe* created a conflict between protecting a civilian who is a potential source of information versus reacting to the presence of suspected insurgents. (D) *Shots into a Crowd* required the trainee to choose between ensuring security versus gaining the trust of the community.

Subject Matter Expert Synthesis

After developing the ShadowBox scenarios, we asked eight current and former warfighters, nominated as SMEs, to review each scenario and complete the decision points (i.e., ranking options and providing their rationale). We discovered that several members of the expert panel were not calibrated with the GS concepts. Therefore, we reduced the SME panel to the three SMEs who recognized the importance of being a GS and appreciated the skills and mindset required. Not surprisingly, these were the only SMEs who were actively participating in the SSIM program. Even these SMEs did not completely line up with each other in their ratings, which we expected because Hintze (2008) never found unanimous agreement between his SMEs. We did not exclude any of the responses of the dissenting SMEs; we included minority views as part of the SME feedback, allowing trainees the opportunity to consider various perspectives and acknowledge that one could approach these problems in multiple ways.

Procedure

The participants were told that they were in a decision making study, but were not told about the topic of GSs or the importance of working effectively with civilians without antagonizing them. They were given individual booklets containing ShadowBox materials. The booklets contained an instructions page and the four military-based scenarios. An instructor led the group of participants through each scenario in a synchronized fashion. The scenario and decision points were read aloud. The participants worked through the four scenarios using a pen-and-paper format. We assured participants that their responses would be anonymous.

Each scenario contained three or four decision points that assessed cognitive challenges and actions. Participants ranked the alternatives from most important to least important. All participants ranked alternatives in two ways; 1) indicating their own ranking, using a column at the left of the page presenting the options, and 2) predicting how a panel of experts would rank the options, using a column to the right of the options. The instructions stated, "In the left column, rank the options to align with how you would handle the scenario. In the right column, try to predict how a panel of experts would respond. Rank the options (1 = best, 4 = worst)."

We also clarified the nature of the expert panel by providing a brief explanation: "Who are the experts? They are highly experienced and respected military personnel. Some are Marines; others are Army soldiers (Special Forces). But what makes them experts for this study is their skill in working with civilians to get voluntary compliance without making people angry. They are aware of the need for security but know how to gain cooperation without provoking antagonism. Thus, they may be different from experienced warfighters you have seen in action. Keep that in mind when you try to predict their responses."

At each site, participants were randomly grouped in 15 person cohorts (Control or Expert Feedback), and completed the experiment with their cohort. The control group (n = 31) consisted of Marines at Camp Pendleton (n = 15) receiving an ABCD order of the four scenarios, and Marines at Camp Lejeune (n = 16) receiving a counterbalanced DCBA order of the scenarios. Twenty-nine (n = 29) Marines from Camp Lejeune comprised the Expert Feedback condition (ABCD, n = 15; DCBA, n = 14).

Control Condition

Participants in the control condition provided both rankings (own and perceived expert) and their rationale for their own rankings. This group did not receive SME feedback. The four scenarios took approximately three hours to complete.

Expert Feedback Condition

Participants in the Expert Feedback condition received SME feedback after each decision point in the form of text that described the experts' rankings and rationale. We used PowerPoint slides to display SME rankings and their rationale. The instructor also read this information aloud after the participants had recorded their rankings and rationale. After seeing the SME rankings and rationale, participants were asked to compare their responses with those of the SME panel, and write down any lessons learned from the decision point. The four scenarios took approximately three hours to complete.

We also conducted a pilot test using skilled facilitation along with the expert feedback. We identified two exemplary scenario facilitators and added two additional training sessions, Facilitated Expert Feedback groups, all using the ABCD order of scenarios. One group, at Camp Pendleton, had 13 Marines, the other, at Camp

Lejeune, had 14 Marines. These sessions lasted approximately four hours because of the time needed in the discussions.

RESULTS

We measured performance by comparing each participant's top ranking to that of the SME panel for each decision point. Participants from both sites (Camp Pendleton and Camp Lejeune) performed similarly across all scenarios, $t(60) = 1.49$, $p > .05$. Therefore, we collapsed site and investigated potential condition differences (control group vs. Expert Feedback group).

We examined the improvement of the Expert Feedback group from Time 1, the first scenario they received, versus Time 4, the last scenario they received. The Expert Feedback condition showed significant improvement from Time 1 ($M = .46$, $SD = .26$) to Time 4 ($M = .59$, $SD = .23$), $t(29) = -2.77$, $p < .05$. Over time, participants' choices better matched those of the SME panel by 28%.

We also compared the control group to the Expert Feedback group. We found a significant interaction effect between scenario performance (Time 1 vs. Time 4) and condition (control group vs. Expert Feedback group), $F(1, 58) = 9.01$, $p < .01$ (see Figure 1). No performance differences were observed between Expert Feedback and control conditions at Time 1, $t(58) = 1.10$, $p > .05$. However, at Time 4, the Expert Feedback condition performed 28% better than the control condition, $t(58) = 2.52$, $p < .05$.

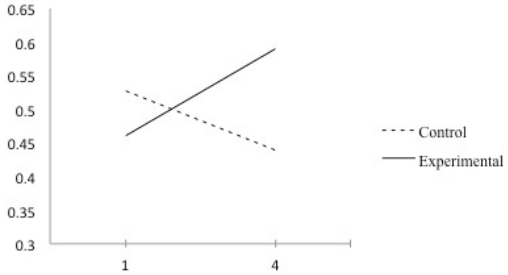

Figure 1. Experiment 1: Interaction and Learning Effect.

The two pilot sessions with skilled facilitators (Facilitated Expert Feedback condition) also improved performance. These groups, which provided the expert feedback plus a discussion led by highly skilled facilitators, showed improved performance. The alignment with the experts on the fourth scenario was 6% higher for the facilitated discussion group compared to the group that received expert feedback but without any discussion, 58% versus 54.8%. This difference was not statistically significant, $t(56) = .80$, $p > .05$.

Next, we explored whether warfighters trained using ShadowBox on an Android tablet would show comparable performance improvements to the pen-and-paper results described above.

EXPERIMENT 2

METHOD

Participants

We collected data from Army officers at Ft. Benning, GA. The participants were 2nd Lieutenants ($N = 30$) who recently completed IBOLC (Infantry Basic Officer Leadership Course). All participants were male.

Materials

MacroCognition LLC and SoarTech developed an Android based application, Mobile Application ShadowBox Training (MAST). The same four scenarios, as well as the decision points and SME feedback used in Experiment 1 were integrated into MAST and used for Experiment 2. Participants were issued a Samsung Galaxy Tab S 10.5, running Android version 4.4.

Procedures

We used many of the same procedures in Experiment 2 as described in Experiment 1, but with important differences. Experiment 2 used the Android tablet instead of pen-and-paper booklets and each participant had his own tablet computer. In Experiment 2 the feedback from the expert panel was presented individually, on the Android tablet, rather than to the entire group via PowerPoint. In Experiment 2 each participant worked at his own pace. In Experiment 1 the groups were synchronized, the scenario was read aloud and the expert feedback was presented at the end of each Decision Point.

We conducted two sessions ($n = 14$, $n = 16$), both using the same ABCD scenario order.

RESULTS

The groups using the software application MAST showed a 21% improvement from Time 1 (M = .48, SD = .21) to Time 4 (M = .58, SD = .19), $t(29) = -2.10$, $p < .05$. Though we did not run a counterbalanced design, we had found in Experiment 1 that the scenarios used at Time 1 and Time 4 had obtained approximately the same scores.

The participants in this experiment took an average of 47 minutes to complete the four scenarios, much less than the three hours in Experiment 1. The results of Experiment 2 were similar to the pen-and-paper evaluation, even though they used the software application and took much less time.

DISCUSSION

The evaluation data demonstrates a strong learning effect with minimal amounts of practice. The participants' responses aligned much closer to those of the expert panel on the final scenario compared to the first scenario, a 21% improvement in Experiment 2 with the Android tablet version, and a 28% in Experiment 1 with the pen-and-paper version. The Expert Feedback group in Experiment 1 was 28% more aligned with the experts than the control group. These findings support the significant improvement that Hintze (2008) found in his work with firefighters. In Experiment 1, the alignment with expert ratings on the fourth and final scenario was 59%. We have achieved a very strong level of alignment with less than a half-day of ShadowBox training.

How does ShadowBox training without any facilitated discussion compare to having a skilled facilitator conducting the training? We were not able to collect controlled data on this issue, but we did run several sessions using highly skilled facilitators along with providing expert feedback. We found that the facilitated discussion did improve performance slightly over the non-facilitated expert feedback condition, but the increment was relatively small and non-significant.

The GS training we provided using ShadowBox differs from efforts to teach social skills by describing ways to gain rapport (e.g., Damari & Logan-Terry, 2015), or ways to influence people and gain voluntary compliance (e.g., Cialdini, 1993; Thompson & Jenkins, 1993/2004). Our approach was aimed, not at the level of individual skills and behaviors, but at a cognitive level: the way people make sense of situations. We speculate that by adding a GS frame to a person's sensemaking repertoire, we may be altering their professional identity.

The ShadowBox training method appears to have the potential to be a fruitful addition to military training, especially for training perceptual and cognitive skills. ShadowBox training can be used in a classroom setting. This method also would work well outside of class, in situations where instructors want to preserve existing curriculum arrangements or increase the preparation of trainees prior to classroom sessions. In addition, ShadowBox should be an effective tool outside of a training rotation, as in field situations where warfighters have spare time. We have developed a web-based software version for situations in which trainees do not have Android tablets; this version runs on laptops and other computers.

Limitations of the Research

We continue to analyze the results to try to develop better measurement procedures beyond alignment with the expert panel for the top-ranked option. Our data analysis did not systematically examine the material entered in the rationale sections of the ShadowBox protocol, and we are exploring ways to capture these qualitative data. We did informally review the material in the rationale sections for insights about the thought processes of participants. Our assumption was that a participant's match with the expert's top ranking indicates the participant agreed with the SME panel and that as participants aligned more closely with the SME responses, the more they were adopting the experts' mental model. We did not have a way to independently capture the mental models of the participants or the experts.

Additional investigation is needed to validate the effectiveness ShadowBox training using organizational performance metrics. It is unclear how ShadowBox training performance (i.e., match with expert responses) translates to actual performance improvement in the wild. The ShadowBox method can be used to train specific skills and impart certain mindsets, therefore, we expect that individuals receiving the training would be more inclined to apply such mindset(s) in new situations. In the future, we plan to assess the relationship between ShadowBox performance and workplace performance.

The analysis in this study only included participants' own rankings for what they would do (the left-hand column in the answer sheet). We did not include their predictions of the expert panel, in the right-hand column, but we intend to explore this further in the future. Our pilot testing revealed that many participants mistakenly believed the SMEs used in this study were very security oriented. We addressed this by including a short description explaining that these experts were distinguished by their skills of gaining voluntary compliance from civilians without provoking hostility.

The scenarios varied somewhat in difficulty — some scenarios seemed to have higher scores than others for matching the expert rankings. This variation did not affect our findings because in Experiment 1 we counter-balanced the scenario order. We also determined that Scenarios A and D were roughly comparable in difficulty, so the non-counterbalanced order in Experiment 2 should not have been affected. If anything, Scenario A was easier than Scenario D, making it more difficult to find a difference between the first trial with Scenario A and

the fourth trial with Scenario D. Nevertheless, we are not satisfied that we sufficiently understand what determines a scenario's level of difficulty, and we are continuing to explore the data for clues.

The control group was not a true control group because these participants worked through four scenarios, generated rankings for decision point options, and pondered the rationale for their choices. Even though they didn't receive information about the choices and rationale of the experts, they had an opportunity for reflection and learning. If we had a chance to repeat Experiment 1, we would not ask these control participants to ponder the rationale for their choices.

The training objective was probably too difficult. We expect that ShadowBox training will provide better results with conventional training objectives. We were trying to change the professional identity of the Expert Feedback group, in the face of a security-oriented military mindset that rejects short-term risk. The premise of the GS mindset is that some small short-term risks may reduce long-term risks, as when a civilian populace comes to trust the warfighters and warns them of insurgent attacks and IEDs. We had not anticipated the extent to which the security-oriented military mindset was opposed to any unnecessary short-term risks. In retrospect, we might have chosen a less controversial and difficult training objective. We were fortunate that our intervention worked so well.

A final limitation was that the training regimen was somewhat inappropriate. Our training was conducted in less than a half-day, four scenarios one after the other, which was the same procedure Hintze (2008) used. This massed practice, three hours of ShadowBox scenarios in Experiment 1, generates fatigue; the control group received lower scores on the fourth scenario than the first. However, for pragmatic reasons we had to conduct the training in one long block. Ideally, the participants would have received only one or two scenarios a day over the course of a week. Nevertheless, even with a non-optimal training regimen, the ShadowBox method produced strong training effects.

ACKNOWLEDGEMENTS

This work was supported by The Defense Advanced Research Projects Agency (government contract 06-1825383). The views, opinions, and/or findings contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the SSIM program managers, the Defense Advanced Research Projects Agency, or the Department of Defense. We appreciate the assistance we received from: LTC John M. Grantz, Neil Hintze, Purush Iyer, Kenn Knarr, Brian Lande, Adele Luta, Kimberly Odam, Rebecca Reichardt, John Schmitt Leah Watson, and Jonathan Wender

REFERENCES

Cialdini, R. B. (1993). *Influence: Science and practice* (3rd ed.). New York, NY: HarperCollins.

Crandall, B., Klein, G. & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis.* Cambridge, MA: MIT Press.

Damari, R. R. & Logan-Terry, A. (2015). *Rapport Building in Military Role Play Training.* Manuscript in preparation.

Hintze, N. R. (2008). First responder problem solving and decision making in today's asymmetrical environment. Unpublished Master's thesis, Naval Postgraduate School, Monterey, CA.

Klein, G., Hintze, N. & Saab, D. (2013). Thinking inside the box: The ShadowBox method for cognitive skill development. In H. Chaudet, L., Pellegrin & N. Bonnardel (Eds.) *Proceedings of the 11th International Conference on Naturalistic Decision Making, Marseille, France, 21-24 May 2013.* Paris, France: Arpege Science Publishing.

Klein, G., Klein, H. A., Lande, B., Borders, J. & Whitacre, J. C. (2014a). The Good Stranger frame for police and military activities. *Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting*.

Klein, G., Klein, H. A., Lande, B., Borders, J. & Whitacre, J.C. (2014b). *Police and Military as Good Strangers*. Manuscript submitted for publication.

McDaniel, M. A. & Nguyen, N. T. (2001). Situational Judgment Tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment 9*(1/2), 103-113.

Schmitt, John F. (1994). Mastering Tactics: A Tactical Decision Games Workbook, Marine Corps Association, Quantico, Virginia.

Thompson, G. & Jenkins, B. (1993/2004). *Verbal judo: The gentle art of persuasion.* Colorado Springs, CO: Alive Communications.