Special issue

The ShadowBox Approach to Cognitive Skills Training: An Empirical Evaluation

Gary Klein, MacroCognition LLC, and **Joseph Borders**, ShadowBox LLC

Unlike behavioral skills training, cognitive skills training attempts to impart concepts that typically depend on tacit knowledge. Subject-matter experts (SMEs) often deliver cognitive training, but SMEs are expensive and in short supply, causing a training bottleneck. Recently, Hintze developed the ShadowBox method to overcome this limitation. As part of the Defense Advanced Research Projects Agency's Social Strategic Interaction Modules, Klein, Hintze, and Saab adapted the ShadowBox approach to train large numbers of trainees without relying on expert facilitators. As part of this program, we used the ShadowBox approach to train warfighters on the social cognitive skills needed to successfully manage civilian encounters without creating hostility or resentment. ShadowBox training was evaluated in two studies. Evaluation 1 provided 3 hr of nonfacilitated, paper-based training to Marines at Camp Pendleton and Camp Lejeune (N = 59), and improved performance (i.e., match to the SME rankings) by 28% compared to a control group. Evaluation 2 provided 1 hr of nonfacilitated training, administered via Android tablet, to soldiers at Fort Benning (N = 30) and improved performance by 21%. These results, both statistically significant, suggest ways to use scenario-based training to develop cognitive skills in the military.

Keywords: decision making, field evaluation, military, naturalistic decision making, training

One of the major challenges to training cognitive skills (as opposed to procedures, perceptual-motor skills, or declarative knowledge) is the availability of subject-matter experts (SMEs). Ideally, SMEs would be able to provide one-on-one instruction and coaching, but few settings have enough SMEs for classroom sessions, let alone individual coaching. Moreover, cognitive skills often involve tacit knowledge, and even if SMEs are available, they may not be able to describe how they make decisions, size up situations, and/or notice subtle cues. SMEs have little or no pedagogical training. They are good at the skill they are teaching but may not be good at teaching that skill. Thus, the lack of SMEs who can provide useful instruction is a bottleneck in training cognitive skills.

The goal of this article is to describe a possible way around this bottleneck: the ShadowBox approach.

ShadowBox Training

Neil Hintze (2008), a battalion chief with the New York City Fire Department (FDNY), developed the initial strategy that was later termed *ShadowBox*. Hintze wanted to train firefighters to handle unusual situations such as earthquakes or terrorist attacks. He presented trainees with a scenario that included a description of a realistic, job-related challenge supplemented with visual aids (e.g., diagrams, maps, and images). The scenario was periodically interrupted by decision points that required the trainee to rank order a set of options (typically three to six options). The decision questions were which action to select, which goals to prioritize, which cues to monitor more carefully, or what type of information to seek. Once the trainee prioritized the alternatives, he or she wrote a rationale explaining the reasons for their rank ordering.

Next, Hintze's method added a novel component to the training by incorporating carefully prepared narratives behind the decision-making process (e.g., mental model) provided by domain

Address correspondence to Gary Klein, MacroCognition LLC, 120 West Second St., Suite 1210, Dayton, OH 45402-8522, USA, gary@macrocognition.com.

Journal of Cognitive Engineering and Decision Making
201X, Volume XX, Number X, Month 2016, pp. 1–13
DOI: 10.1177/1555343416636515

experts. Hintze arranged for a panel of experts to work through the same scenarios and rank the alternatives. They were also asked to describe their rationale for the choices they made. In doing this, the SMEs conveyed their personal approaches and mental models of the situation. Hintze synthesized the rankings and rationale statements of the experts. When the experts disagreed, Hintze tried to resolve any discrepancies, but if he could not, he added a minority view to show the trainees that there was not merely one correct answer.

After the trainees completed their rankings and recorded their rationale, they were shown how the SMEs ranked the options, and they got to see the reasons that the SMEs provided. Trainees were often eager to discover if their rankings and rationale aligned with the SME panel. There is enough competition and desire for mastery to motivate the trainees to try to make rankings that more closely match with the experts. However, the match between the trainee/expert rankings is just the hook to create excitement and enthusiasm. The real learning comes from asking the trainees to study the rationale provided by the experts and to compare it to their own rationale. They reflect on what the SMEs noticed that they, the trainees, had not.

Therefore, the ShadowBox approach, as it was later called, lets trainees see the world through the eyes of the experts. And most importantly, the experts do not have to be present. The trainees learn how the experts make sense of situations, what they pay attention to, and why they make their choices. Trainees are exposed to the mental models of the experts without ever hearing the term *mental model*. The experts' choices and rationale reflect their mental models, but all the material is within the context of the scenario rather than as an explicit statement of the experts' mental models.

Hintze evaluated ShadowBox training using 14 SMEs (FDNY officers with at least 15 years of fire department experience). Twenty-nine New York State fire officers, promoted to the rank of lieutenant within the previous 12 months, participated in the evaluation study. The fire officers were split into two groups: experimental and control. The experimental group of 14 New York fire officers completed four ShadowBox scenarios and received SME feedback, whereas the control group (15 fire officers) completed four scenarios without receiving SME feedback. Both the experimental and control groups completed the scenarios in 1 day. Hintze compared the experimental and control groups on the fourth and final scenario, measuring how closely the rankings matched those of the SMEs. After a single day of training, the experimental group received a mean score of 86.9, and the mean control group score was 73.6, a difference of 18% that was significant ($F = 14.09$, $p < .001$). Hintze did not take the participants' starting scores into account to show relative improvement of the experimental group compared to the control group.

Influences on ShadowBox

Any new approach can be traced back to a wide variety of precursors and previous work. In the case of ShadowBox, some of the most important influences are cognitive transformation theory, accelerated expertise, scenario-based methods such as the Situational Judgment Test (SJT), tactical decision games (TDGs), and the work of Bloom and Broder on expertise.

Cognitive transformation theory. Klein and Baxter (2009) developed cognitive transformation theory to account for the acquisition of expertise as a step-wise rather than a smooth performance curve. The claim was that experts would develop powerful mental models but then would fixate on these mental models rather than discarding them in order to improve further. Only when something traumatic occurred, such as a failure, would the experts reexamine their mental models and replace questionable aspects. Cognitive transformation theory advocates training that emphasizes sensemaking and the improvement of mental models and includes unlearning flawed mental models. Wiltshire, Neville, Lauth, Rinkinen, and Ramirez (2014) assessed cognitive transformation theory and found that its recommendations matched the training strategies used by highly experienced air traffic control instructors. Wiltshire et al. concluded that "Klein and Baxter may be unique and are at least rare in their center-stage placement of the mental model and in the comprehensive way their theory draws together practice, diagnosis, feedback, and learning

objectives to guide mental model development" (p. 221).

But how can researchers and practitioners facilitate this type of successive transformation, and can they expedite the replacement of mental models? The ShadowBox approach may provide a means to let trainees discover flaws in their mental models and to shift to more effective mental models. Thus, ShadowBox training can be seen as a way to implement the training recommendations of cognitive transformation theory. Cognitive transformation theory definitely influenced the way we adopted and adapted the ShadowBox approach.

Accelerated expertise. The training strategy Hintze used, now called ShadowBox training, is consistent with the accelerated expertise program (Hoffman et al., 2014). ShadowBox training can be seen as a platform for achieving rapidized training, higher levels of proficiency (accelerated proficiency), better transfer (rapidized transposition), and facilitated retention. ShadowBox training is a way to achieve the "tough case time compression," recommended by Hoffman et al. (2014).

Scenario-based training. There are specific precursors to ShadowBox from the tradition of scenario-based training (see Burns, Cannon-Bowers, Salas, & Pruitt, 2006, for a review of scenario-based training approaches). It may be instructive to trace the ways that ShadowBox training builds on previous scenario-based approaches and also to examine how the ShadowBox strategy differs from these precursors.

SJT. ShadowBox training is consistent with SJT (McDaniel & Nguyen, 2001), which was developed a half century ago for personnel selection. Bruce and Learner (2006) described a method for using scenarios to assess supervisors. The SJT presents realistic scenarios and has the respondents identify the action they would most likely perform. SJT can be presented using different modalities such as paper and pencil and video, similar to ShadowBox. Thus, we can consider ShadowBox as a variant of SJTs.

Like SJTs, the development of ShadowBox training relies on critical incident elicitation with SMEs. ShadowBox places more emphasis on using cognitive task analysis (CTA) methods (e.g., Crandall, Klein, & Hoffman, 2006; Klein, Calderwood, & Macgregor, 1989) to capture critical incidents, generate scenarios, and formulate cognitive and behavioral-based decision points and response options. SJTs tend to focus on action-based questions. ShadowBox is not restricted to choosing between courses of action but also incorporates more cognitive-based decisions such as assigning priorities, monitoring cues, and gathering information. ShadowBox elaborates on the SJT methodology by having trainees provide a rationale for their ratings and by presenting SME feedback in the form of rankings and synthesized rationale. ShadowBox also has trainees compare their rationale statements with the SME rationale and describe what is different—that is, what the SMEs noticed and considered that the trainees had not considered. Thus, we suggest that there are several ways that scenario-based training can be enhanced in order to train cognitive skills.

TDGs. ShadowBox training is also an elaboration of TDGs (Schmitt, 1994). TDGs are scenario-based and designed to train individuals or small groups. The *Marine Corps Gazette* has published TDGs allowing readers to respond individually and send their responses to the magazine. The following month, the *Gazette* publishes the best responses they have received, so readers can compare their responses to the ones that are published. ShadowBox differs from TDGs by using CTA methods to identify a set of response options for each decision point, by having the trainees rank these options, and by synthesizing the rankings and rationale of the panel of experts as a point of comparison. ShadowBox does not claim that there is an absolute correct answer or a right way to rank the options.

Our experience is that TDGs work best when run in a small group with a skilled facilitator. Although we do not have data, we believe that under these conditions, TDGs provide better training than ShadowBox. However, we also have seen TDGs administrated by mediocre facilitators, with disappointing results. Furthermore, many organizations do not have the funding or departmental resources to construct and deliver this exhaustive training; therefore, use of skilled facilitators and small group exercises will not easily scale up to reach large quantities

of trainees. Skilled TDG facilitators are critical to the training sessions, as they can create excitement, tension, and discoveries. Mediocre or untrained facilitators are left on their own because TDGs do not systematically describe how to evaluate the trainees' responses or how to query the trainees about their reasons. ShadowBox addresses these problems using the panel of SMEs as a standard/point of reference. Trainees may disagree with the experts, but at least they have to review what the experts were thinking, including the cues on which the experts relied. Using a video format, ShadowBox can introduce more subtle cues and perceptual discrimination and address some aspects of tacit knowledge that TDGs usually do not cover.

Contrast to experts. Another precursor of ShadowBox training is the work of Bloom and Broder (1950). They contrasted college students who were successful at handling difficult multiple-choice tests versus students who were struggling by having the students provide a think-aloud protocol as they worked through multiple-choice items. The low-performing students tended to read the problem and judge if they knew the answer. If they did, they would provide it. Otherwise, they would skip the problem or guess randomly. In contrast, the successful students approached a multiple-choice test as a problem-solving exercise. If they did not know the answer, they would try to eliminate options that seemed wrong. They gleaned whatever they could from the information they were given and also drew on any related information they might have. They searched for ways to value some options over others. In this way, they might reduce a four-item question to two plausible items and then guess, with odds of about 50% rather than 25%, assuming they were successful in filtering out the wrong answers. Bloom and Broder took this exercise further. They provided the low-performing students with access to the think-aloud protocol records of successful students but did not tell the low-performing students how to do better. Bloom and Broder reasoned that for the lessons to stick, the low-performing students had to make their own discoveries. If the researchers had tried to impose a new problem-solving strategy, the low-performing students might not understand it or feel comfortable with it. In this way, Bloom and Broder successfully boosted the scores of the low-performing students. We think that ShadowBox training takes advantage of this finding by having the trainees define for themselves what the panel of SMEs had noticed that was missing from their own rationale statement. The trainees flag what they are noticing—what is in their zone of proximal development. Through reflection, trainees can incorporate new information (e.g., what the expert recorded that they missed) into their existing knowledge base.

Thus, ShadowBox training is consistent with several different instructional strategies that use scenarios and seek to help trainees gain tacit knowledge in order to build richer mental models. ShadowBox appears to be an advance over the existing methodologies because of the way it uses the rationale behind decisions and makes use of a panel of SMEs who have gone through the same scenarios. That is how ShadowBox enables trainees to see the scenario—and the world—through the eyes of experts without the experts having to directly participate in the training.

Hintze previously demonstrated that when he facilitated the discussions, ShadowBox training resulted in significant improvements in performance, measured as the match between the trainee responses and those of the experts. But the question remained as to whether the method could scale up and improve performance without any facilitator.

Applying ShadowBox Training With Warfighters

The ShadowBox approach grew out of the Hintze (2008) research shortly after that project was completed. The first author met Hintze in September 2008, just as Hintze was completing his research study. They conducted a decision training workshop together in Seattle in January 2010, which was the first time the first author had a chance to observe Hintze's method in action. The opportunity to evaluate the effectiveness and scalability of ShadowBox training arose in 2011 as part of the Defense Advanced Research Projects Agency's (DARPA) Strategic Social Interactions Module (SSIM; U.S. Department of Defense, Defense Advanced Research

Projects Agency, 2011). SSIM, nicknamed the "Good Strangers" program, was designed to teach social skills to warfighters. The unofficial term Good Strangers was somewhat controversial within the SSIM project; some warfighters felt the term misrepresented their work with civilian encounters. However, no other descriptor ever emerged, so we use that term in this article.

The difficulties warfighters have faced working with civilians in Afghanistan and Iraq motivated the SSIM program. Warfighters who were very effective in combat often struggled during their encounters with civilians. They frequently intimidated civilians, treated them with contempt, antagonized them, and escalated conflicts unnecessarily. They alienated civilians who might have been willing to cooperate and angered civilians who might have been content to stay neutral. In response, the Army and Marine Corps established mock villages to provide warfighters with realistic cultural training. Although these mock villages seemed quite valuable, they were very expensive to maintain. They required large numbers of role players and long lead times to arrange. So in future theaters, with different cultures, the military would be hard-pressed to set up comparable villages in a short time. The SSIM program sought a faster and less expensive strategy for teaching warfighters the necessary social skills to work with civilians to gain cooperation rather than rely on coercion.

The SSIM program incorporated ShadowBox as a means of training some of the Good Stranger skills that were identified through a separate project using CTA interviews with police and military personnel regarded by their peers and supervisors as Good Strangers (Klein, Klein, Borders, & Whitacre, 2015). The purpose of the CTA study was to understand the skill set necessary for effective social encounters with civilians and particularly the cognitive aspects of these skills. Klein, Borders, Wright, and Newsome (2015) wanted to uncover the ways that Good Strangers made sense of situations and how their sensemaking differed from colleagues who were much less successful in handling civilian encounters. Thus, the DARPA program supported the CTA work to clarify the cognitive training requirements for being a Good Stranger in parallel with the development of the ShadowBox approach as a means for training these cognitive skills.

The CTA study consisted of interviews with 41 participants—19 police officers and 22 military personnel with overseas experience. This study resulted in a sensemaking account of Good Strangers (Klein, Klein et al., 2015). Good Strangers frame civilian encounters as opportunities to gain the trust of civilians. They use this trust-building frame in addition to other frames for carrying out the mission and ensuring their own safety and security. Warfighters and police officers who did not qualify as Good Strangers seemed to lack this trust-building frame. The Good Stranger frame of trying to increase trust from the beginning to the end of an encounter served to organize how the Good Strangers viewed situations, what they noticed, what opportunities they seized and also how they managed their other goals of safety and mission accomplishment. Additionally, the Good Stranger frame guided the ways they sought to gain rapport, to take the civilians' perspective, to gain voluntary compliance, and to deescalate conflicts.

When the ShadowBox task was added to the SSIM program, we decided to apply ShadowBox training to the overall Good Stranger sensemaking frame: seeking to gain the trust of civilians during encounters. We prepared three law enforcement scenarios that highlighted this frame, using the incidents gathered during the CTA study, and conducted a pilot study of these ShadowBox materials with 16 police officers.

We also used the pilot study to formulate the ShadowBox approach so that ShadowBox training did not depend on a facilitator with domain knowledge. We worked out the procedure for injecting the SME panel results during the training, and we settled on the term *ShadowBox training* to describe the method because the task required the trainees to write down their assessments and reasoning in a small box, forcing them to prioritize what they thought was important (Klein, Hintze, & Saab, 2013).

Once we had formalized the procedures for presenting ShadowBox training without relying on a skilled facilitator, we were ready to evaluate how effective the training was. We conducted two evaluation studies with military participants.

METHOD

Evaluation 1

Participants. We collected data at two U.S. Marine Corps sites, using commissioned officers (lieutenants and captains) and noncommissioned officers (staff sergeants) at Camp Pendleton, California, and at Camp Lejeune, North Carolina (N = 59). Most of the participants were between 25 and 30 years of age and had less than 6 years of active duty experience with at least one overseas deployment. All participants were male.

Materials. We generated four ShadowBox scenarios, all involving military–civilian interpersonal encounters. The scenarios included a challenge for managing workers from a foreign culture in preparing food in a military mess hall, taking control of a large Iraqi village that contained a militia that might possibly be hostile, trying to gather information from a civilian despite possible threats to the civilian's life (see Appendix), and deescalating a situation in which Marines accidentally discharged weapons into a nonhostile crowd, injuring three children. Each scenario contained three to four decision points presenting options about actions to be taken, cues to be monitored, goals to be prioritized, information needed, or anticipating various outcomes.

Procedure. Prior to the training interventions, a panel of SMEs provided their rankings and rationale for the four scenarios. We started with eight SMEs but determined that five of them needed to be excluded because their responses were not consistent with the Good Stranger sensemaking frame: seeking to gain the trust of civilians during encounters. Despite their extensive overseas combat experience, this group of SMEs had little experience working cooperatively with civilians. As expected, the remaining three SMEs did not always agree, so we included a minority view for several of the decision points.

The participants were told they were taking part in a decision-making study. We did not explain the concept of Good Stranger or provide any information that might influence their choices (we did conduct an extended debrief after each data collection session). A facilitator distributed the booklets and monitored the compliance of the Marines but did not lead any discussion, nor was discussion between the participants permitted. Early in our SSIM work, we identified a possible source of confusion for Marine participants. They tended to understand "experts" as warfighters who had no tolerance for risk and had little or no interest in fostering good relationships with civilians. Marines had learned this approach in their previous overseas deployments, and it confused them to receive feedback from SMEs who were skilled at managing civilian cooperation. Therefore, prior to the start of the training exercise, we found it necessary to explain who our experts were with the following description:

> The experts are highly experienced and respected military personnel. Some are Marines; others are Army soldiers (e.g., 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.

At each site, Camp Pendleton and Camp Lejeune, the Marines were randomly assigned to a "no feedback" or "SME feedback" condition and completed the session in their cohort. All participants worked individually within classroom settings, filling out their rankings and the rationale for their rankings in a paper booklet. The no feedback group consisted of 31 Marines at Camp Pendleton (n = 15) and Camp Lejeune (n = 16), receiving a counterbalanced ABCD or DCBA order of the four scenarios. We subsequently assessed whether there were differences between the Camp Pendleton and Camp Lejeune participants and did not find any, $t(60) = 1.49$, $p > .05$. The no feedback group worked through the scenarios, ranked the options, and filled in their rationale, but never received any feedback about the choices and reasons of the panel of SMEs. They required approximately 3 hr to complete the four scenarios. We refer to them as a "no

feedback group" rather than a "control group" because they did prepare rationale statements to explain their rankings, and this type of reflection may have conferred some training benefits.

The SME feedback group consisted of 29 Marines from Camp Lejeune, both commissioned and noncommissioned officers, run in two cohorts. One cohort received the scenarios in ABCD order (n = 15) and the other in DCBA order (n = 14). This group received SME feedback after each decision point in the form of PowerPoint slides displaying the SME choices and rationale. After seeing the SME ranking and rationale, the Marines were asked to compare their own responses to those of the SMEs and write down any lessons they learned about what the experts noticed that they had not. There was no class discussion. The four scenarios took approximately 3 hr to complete.

RESULTS

To assess performance, we compared participants' top-ranked options to the choice of the expert panel for each decision point. In cases where the SMEs disagreed, we used the majority choice. We analyzed performance scores for each scenario by calculating the number of times the participant's top ranking matched the top ranking of the SME panel, for all the decision points in a scenario, and dividing this by the number of decision points in the scenario. We conducted two different comparisons.

First, we conducted a within-subjects t test to see if the group receiving expert feedback showed improvement—a closer match to the expert panel's top rankings from Time 1, the first scenario they received, to Time 4, the fourth and last scenario. Over time, with successive scenarios, the Marines' top rankings better matched those of the experts by 28%. The difference between Time 1 (M = .46, SD = .26) and Time 4 (M = .59, SD = .23) was significant, $t(29) = -2.77$, p = .01 (see Figure 1). Furthermore, Cohen's effect size value (d = .51) suggested a moderate practical significance.

Second, we conducted an independent-samples t test to investigate if the group receiving expert feedback outperformed the no feedback group that worked through the same scenarios and generated rationale statements but never received any SME feedback. These two groups did not differ at Time 1, $t(58) = .98$, $p > .05$. However, they did significantly differ at Time 4, as the SME feedback group (M = .59, SD = .23) performed 28% better than the no feedback group (M = .44, SD = .24), $t(58) = 2.45$, p = .02 (see Figure 1). The Cohen's effect size value (d = .63) suggested a moderate practical significance.

In addition to between-group performance differences, we identified the number of participants within each group that improved, stayed the same, or got worse from Time 1 to Time 4. Supervisors may be more interested in how many people changed than in an overall change in the proportion of agreement with the SMEs. Twenty out of the 29 participants in the SME feedback condition demonstrated improvement; their average rate of improvement was 26% from Time 1 to Time 4. Only eight participants in the SME feedback condition got worse, averaging an 18% decrease in alignment with the expert rankings (see Figure 2). In contrast, only 14 out of 31 participants in the no feedback condition improved, and their average improvement was 17%. The remaining 17 participants in the no feedback condition decreased in performance (responses matching to the SME panel), with an average drop of 30%.

METHOD

Evaluation 2

DARPA required that the SSIM program yield innovative, efficient, cost-effective, and field-ready training by the conclusion of the contract to reciprocate various military and police partners' time and effort expended on the program. ShadowBox training was one of the methods proposed because it is scalable in that it can be administered without training facilitators. The SSIM program also funded the following experiment, which involved presenting the scenarios and recording participant responses using mobile Android tablets rather than a pen-and-paper booklet. The software was developed by SoarTech and was labeled MAST (Mobile Application ShadowBox Training). The purpose of this effort was to create a scalable training solution that can extend the ShadowBox training approach and become more accessible

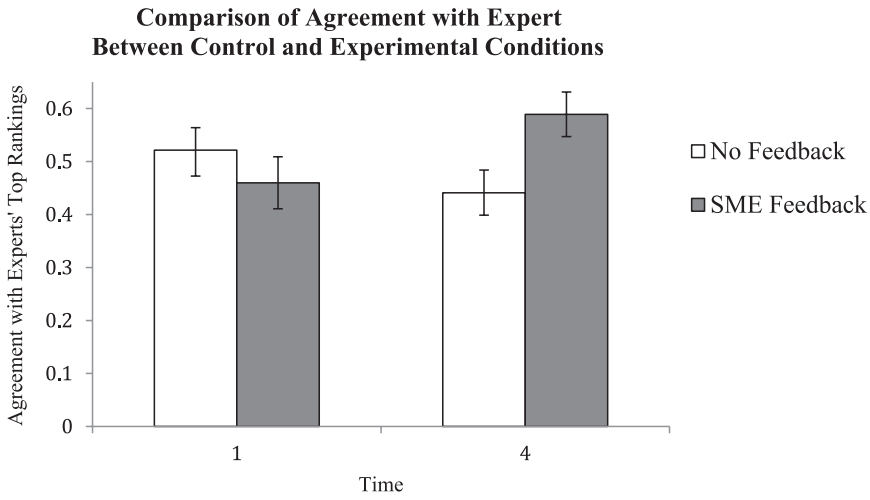

Figure 1. Mean agreement with expert for no feedback and SME feedback test conditions.

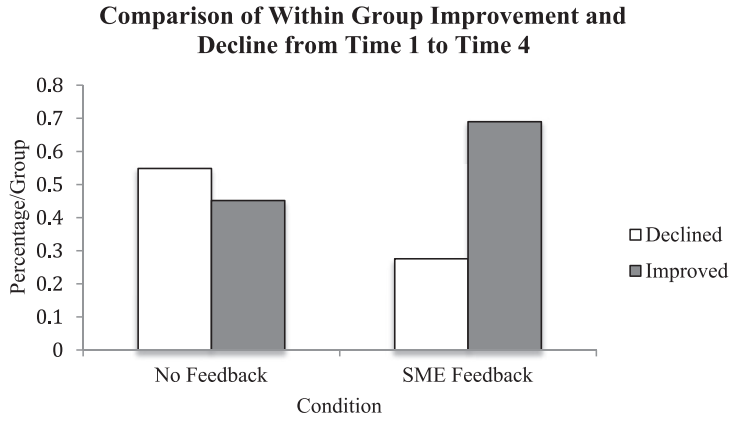

Figure 2. Comparing the percentage within groups that improved and declined from Time 1 to Time 4. One participant in the SME feedback group stayed the same from Time 1 to Time 4.

to warfighters in the field. MAST replicated the paper-and-pencil ShadowBox experience described in Evaluation 1 on a tablet platform and leveraged existing web-based technologies.

Participants. Thirty Army commissioned officers (second lieutenants; all male) at Fort Benning, Georgia, participated in the evaluation study. Most of the participants were recent college graduates, 25 years old or younger, and had less than 1 year of military experience. Nine participants were enlisted prior to becoming a commissioned officer.

Materials. We used the same four scenarios described in Evaluation 1. All scenarios were locally downloaded onto Android tablets, where participants read scenarios, ranked options, and wrote their rationale using the MAST user interface.

Procedure. Each participant was issued his own Samsung Galaxy Tab S 10.5 running Android version 4.4, and each participant completed the training scenarios at his own pace. Unlike Evaluation 1, which presented the SME panel feedback via PowerPoint to the SME feedback cohorts, in Evaluation 2 the expert feedback was presented on the Android tablet. We conducted two sessions ($n = 14$ and $n = 16$), both using the same scenario order (ABCD). Logistical constraints prevented

us from using a counterbalanced design, as we had done in the previous evaluation. We did not include a no feedback group. The participants completed the four scenarios substantially faster than in Evaluation 1, taking an average of 47 min, compared to 3 hours in Evaluation 1.

RESULTS

Using a within-subjects *t* test, participants' ranking agreements with the SME panel's top rankings were compared at Time 1, their first scenario, and Time 4, their final training scenario. Ranking alignment with the SME panel was significantly higher at Time 4 ($M = .58$, $SD = .19$) than at Time 1 ($M = .48$, $SD = .21$), as participants showed a 21% improvement over the course of the training exercise, $t(29) = -2.10$, $p = .04$. Furthermore, the Cohen's effect size value ($d = .38$) suggested a low to moderate practical significance. Although we did not run a counterbalanced design in Evaluation 2, we found in Evaluation 1 that the scenarios used at Time 1, scenario A or D, yielded approximately similar scores. The results of Evaluation 2 matched those in Evaluation 1, despite the delivery method (paper-and-pencil vs. mobile tablet), group presentation of the SME rankings/rationale, and the time it took participants to complete the training.

Twenty out of the 30 participants in the experiment demonstrated improvement from Time 1 to Time 4; their average rate of improvement was 55%. The remaining 10 participants got worse, averaging a 36% decrease in alignment with the expert rankings.

DISCUSSION

These findings support Hintze's study, which obtained an 18% improvement for the experimental group (SME feedback) versus groups that did not receive expert feedback. In a relatively short period of time, 3 hr in Evaluation 1 and less than 1 hr in Evaluation 2, the group receiving expert feedback from the SME panel showed a significantly better match to the rankings of the experts.

The results of our research suggest that it is possible to train cognitive skills in a reasonably short amount of time and in a way that can scale up. Evaluation 1 relied on a facilitator to distribute the ShadowBox materials and present the rankings and rationale of the SME panel via PowerPoint slides. However, the facilitator did not engage in any discussion or even permit an in-class discussion. Despite the limited facilitation, on average, participants in the SME feedback condition demonstrated significant improvements matching with expert rankings over the course of four training scenarios. Facilitation was even more sparse in Evaluation 2. All participants completed the training at their own pace and relied on the MAST software application to receive expert feedback. Participants in Evaluation 2 also improved from the first to the fourth scenario.

Limitations

In the preceding studies, we defined performance by comparing the participants' response alignments with the SME panel's top-ranked option. We did not systematically account for the participants' rationale. We did informally review their rationale for insights about their thought processes, but we excluded this information from data analysis because it could not be easily integrated into the quantitative measures. Our assumption was that if a participant's top ranking matched the expert panel's top ranking, then the participant agreed with the expert panel. To prevent response bias, we asked participants to respond to each decision point in two ways. First, we asked participants to respond to the questions based on what they would do. These are the data we used in our study. Second, we asked participants to predict what the expert panel would do. Interestingly, we found that many participants did not agree with the expert panel. In our pilot studies, we found that many participants mistakenly believed the expert panel was security-oriented because this is what they experienced from their superiors. We addressed this confusion by including the brief expert description explaining that the SMEs involved in this project were skilled in the art of voluntary compliance.

Our goal was to investigate if the participants would be more inclined to apply the Good Stranger mindset as they progressed through the training scenarios. Although we are reporting training improvements, it is uncertain how

response alignment with SMEs translates to performance on the job. Warfighters acting as peacekeepers face dynamic and unfamiliar challenges where they must rapidly make sense of complex information and make effective decisions. We recognize that the ecological validity of ShadowBox training is limited because trainees read scenarios and rank order a list of preselected options. However, we believe that ShadowBox training can be a useful tool to expose trainees to a variety of challenging scenarios that may augment on-the-job experience and introduce targeted SME feedback that can improve sensemaking and decision-making capabilities.

Another limitation is that ShadowBox training is static—the trainee follows along a scripted scenario, rather than letting the trainee's choices alter the scenario. Future versions of ShadowBox may permit some branching. This is an area for further development.

Finally, the ShadowBox approach presents a preselected set of options, as opposed to an open-ended procedure. The preselected options provide strong prompts and are necessary in order to permit efficient scoring.

Formats

We learned that it is important to be careful of the way we summarize the comments provided by the SME panel. In his original study, Hintze led a classroom discussion and could quote from the material provided by the experts. In the paper-and-pencil version of ShadowBox, we needed to fit the SME feedback on a single PowerPoint slide for each decision point. But even this was too much content for the Android tablet to comfortably display. It was a challenge for some soldiers in Evaluation 2 to navigate through the material. Moving forward, it will be important to tailor the training content to the medium being used for content delivery.

We also have explored the use of image (e.g., graphic novel) and video formats in order to reduce the amount of text, and this has received favorable reactions from test groups. The production of reasonable quality graphics may make this format less feasible for some industries.

The video format seems much more suitable for ShadowBox training than a graphic novel format. Using the video format to present scenarios and collect user responses has multiple advantages. Participants have a more enjoyable experience watching a video instead of reading text. Also, this approach allows the test creator(s) to maintain control of the scenario and ensures that participants interpret the scenario content more similarly. Scenario writers can also insert decision points within the video clips. In an alternate version, participants indicate relevant cues within the video by clicking on them within the video screen. This action stops the video, and participants provide their reactions and reasoning for the click. At the end of the video clip, the participants can compare the time, location, and rationale of their clicks with the expert panel. In some settings where perceptual cues are critical, the video format may offer advantages over a text-based format.

Domains

The kind of cognitive skills training exemplified by ShadowBox training seems applicable to a variety of domains (Borders, Polander, Klein, & Wright, 2015). We have used the ShadowBox approach with console operators in petrochemical plants, with social workers in child protective services agencies, and with nurses preparing to complete their training and take hospital positions. The strategy of using cognitive probes inserted into scenarios, along with comparison of rankings and rationale with those of an expert panel, seems to apply easily in each of these different domains. A next step in our research would be to assess whether ShadowBox training improves performance in the field, using organizational performance metrics, and not simply increasing the trainee's match to the SME rankings. This is difficult to study in a military setting, but we are seeking to collect validation data in other domains, such as the performance of nurses as they move from university into a hospital environment.

Probes

Most scenario-based training, such as SJT variations and TDGs, seems to rely on probes about which action the trainee would take. We suggest a broader set of probes, to include prioritizing goals, identifying cues, and selecting types of information to gather. The video feature opens

up additional options for addressing perceptual skills. One of our training observations across domains is that action-based probes may be less effective than other cognitive and perceptual probes. Our work with child protective service caseworkers found that when social workers considered which action to select, they tried to conform to existing guidelines. But when they considered alternative priorities, and especially when they contrasted different cues to monitor, we found a much sharper difference between the highly skilled caseworkers and the journeymen (Newsome, Wright, Klein, Flory, & Baker, 2015). Therefore, we suggest that scenario-based approaches to cognitive training should try to move beyond response selection and into more subtle cognitive issues, much in the way that Endsley (1995) has found ways to measure changes in situation awareness during the course of training.

Platforms

Hintze used a paper-and-pencil version of ShadowBox, as we did in Evaluation 1, and this mode is certainly easy to prepare and deliver. However, we were surprised by the results of Evaluation 2, in which participants using Android tablets with no facilitation demonstrated a 21% improvement in performance. We were even more surprised by the brevity of the training—an average of 47 min versus 3 hr with paper and pencil. In our ShadowBox study with child protective services, we have watched the caseworkers massaging their wrists after writing rationale statements for several hours, and in our study with nurses, we heard the complaints about having to hand-write rationale statements. In response, we have developed a web-based ShadowBox training platform (beta) for desktops, laptops, and tablets, operating on common browsers (e.g., Chrome, Safari, Firefox). Computer-based versions offer many advantages for content delivery, user engagement, data entry (particularly the rationale statements), and data collection.

Applications

Hintze developed the ShadowBox method for personnel training. The computer-based platforms would also enable ShadowBox use with distance learning. So these are two obvious applications of the research.

As we work in different domains, we see the opportunities of using ShadowBox for personnel selection. This, of course, takes us back to one of the precursors of ShadowBox, SJTs, which were originally designed for selection and are still primarily used for that purpose. ShadowBox can move beyond action-oriented questions and target cognitive skills and strategies, which are also important in personnel selection and evaluation. Therefore, personnel assessment would be a third possible application, using the person's match to the ranking and rationale of SMEs as an indicator of the person's mastery of the task requirements.

We have received interest in using these kinds of scenario-based methods for capturing expertise and for knowledge management. Many organizations, particularly those facing large-scaled employee attrition to retirement and turnover, are seeking approaches to capture and transfer expertise efficiently and effectively. ShadowBox is designed to systematically capture components of expertise, such as tacit knowledge, best practices, and lessons learned, and package this information so that it can be saved and is made assessable to large quantities of trainees. This fourth potential application, knowledge management, would capture expertise in the form of the rankings and the rationale statements rather than as written material that may be difficult to use.

A fifth potential application is to use ShadowBox for leadership and supervision. Leaders can provide their own rankings and rationale, rather than relying on a panel of SMEs. Leaders would not claim that their responses were correct. Rather, they could use ShadowBox to help their subordinates anticipate how they, the leaders, would be likely to respond and to interpret situations.

Strategy

One of the most important lessons that we learned from this effort was that it might be possible to provide cognitive training for sensemaking and, particularly, for the frames people use in understanding situations. Klein, Moon, and Hoffman (2006a, 2006b) described a data/

frame model of sensemaking, and Klein, Klein et al. (2015) have amplified this model to suggest some of the ways that a sensemaking frame provides guidance: It influences attention, the cues people notice and ignore, expectations, the goals they pursue, and the actions they consider. Thus, a Good Stranger frame of building trust would differentiate someone who possessed the frame from someone who did not. We speculate that frames help organize a range of subskills (e.g., perspective-taking, rapport building, gaining voluntary compliance) that might otherwise be treated as separate training requirements. Therefore, in some situations, there may be advantages to identify and train sensemaking frames in order to increase the effectiveness of the training program and also to increase its efficiency.

In this research, we observed that a number of the trainees discovered limitations in their frames and mental models. (We are using these terms interchangeably here, even though each has its own research tradition.) They had not considered the value of a trust-building frame for social encounters until they studied how the SMEs viewed the scenarios and the options open to them. This shift in frames/mental models is what was expected from cognitive transformation theory—not a gradual elaboration of existing beliefs but a more abrupt shift. In this case, many of the trainees reduced their strong adherence to the frame of maintaining security and added the new trust-building frame to their cognitive repertoire.

We have explored the use of sensemaking frames in several other ShadowBox projects recently, relying on CTA methods to capture the frames that differentiate personnel who are working at a high level versus their journeyman counterparts. For child protective services caseworkers, we found that the journeymen concentrated on following the rules and procedures and handling each case in accordance with its requirements, whereas the highly skilled caseworkers moved beyond the official complaint/report and investigated a range of safety issues within a family situation (Newsome et al., 2015). For panel operators in a petrochemical plant, the journeymen responded to alarms and problems as they arose, whereas the elite possessed an "operator's mindset." Skilled operators could manage a variety of variables at once, and they anticipated problems before there were any clear symptoms. For nursing students who were preparing to manage medications with a geriatric population, the journeymen conceptualized their job and their role as "getting pills into patients" and following standard practice, whereas the highly skilled nurses tried to take the perspective of the patients to understand what might be causing the resistance and how the standard practice might need to be modified to take into account special needs of a different subgroup. Therefore, we speculate that a sensemaking approach to cognitive skills training may have general value and may be trainable using ShadowBox or other types of scenario-based methods. Of course, training needs to include details of the job, affordances, and other aspects of tacit knowledge, not just the overall frame, so the scenarios have to operate at several levels in parallel.

The training community has known for a long time about the value of scenarios. The contribution of this research is to suggest some ways to present scenarios that are enhanced with the reactions of experts and engage the trainees to actively compare their responses to those of the experts. Furthermore, the kind of scenario-based training we have described appears to scale up so that it can be broadly delivered without running into the bottleneck of relying on SMEs or skilled facilitators. Finally, our findings suggest a strategy for conducting cognitive skills training that tries to use sensemaking concepts in order to help the trainees acquire new ways to frame situations.

ACKNOWLEDGMENTS

This work was supported by the Defense Advanced Research Projects Agency (government contract 06-1825383). We would also like to thank Brian Lande, Bill Casebeer, Adele Luta, Kenn Knarr, LTC John Grantz, Corinne Wright, and Emily Newsome. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the aforementioned entities. Portions of this paper were presented at the 12th International Conference on Naturalistic Decision Making (Klein, Borders et al., 2015).

REFERENCES

Bloom, B. S., & Broder, L. (1950). *Problem-solving processes of college students*. Chicago: University of Chicago Press.

Borders, J., Polander, N., Klein, G., & Wright, C. (2015). ShadowBox™: Flexible Training to Impart the Expert Mindset. *Procedia Manufacturing, 3*, 1574-1579.

Bruce, M. M., & Learner, D. B. (2006). The supervisory practice test. *Personnel Psychology*, *11*(2), 207-216.

Burns, J. J., Cannon-Bowers, J. A., Salas, E., & Pruitt, J. S. (2006). Advanced technology in scenario-based training. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress* (pp. 365-374). Washington, DC: American Psychological Association.

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

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65-84.

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

Hoffman, R. R., Ward, P., Feltovich, P. J., DiBello, L., Fiore, S. M., & Andrews, D. H. (2014). *Accelerated expertise: Training for high performance in a complex world*. New York: Taylor & Francis.

Klein, G., & Baxter, H. C. (2009). Cognitive transformation theory: Contrasting cognitive and behavioral learning. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond. Volume I: Learning, requirements and metrics* (pp. 50-65). Westport, CT: Praeger Security International.

Klein, G., Borders, J., Wright, C., & Newsome, E. (2015). *An empirical evaluation of the ShadowBox™ training method*. Paper presented at the 12th International Conference on Naturalistic Decision Making, McLean, VA.

Klein, G. A., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. Systems, Man and Cybernetics, *IEEE Transactions on*, *19*(3), 462-472.

Klein, G., Hintze, N., & Saab, D. (2013, May 21-24). *Thinking inside the box: The ShadowBox method for cognitive skill development*. In H. Chaudet, L. Pellegrin, & Bonnardel (Eds.), *Proceedings of the 11th International Conference on Naturalistic Decision Making, 2013* (pp. 121-124). Marseille, France: Arpege Science.

Klein, G., Klein, H. A., Borders, J., & Whitacre, J. C. (2015). Police and military as good strangers. *Journal of Occupational and Organizational Psychology*, *88*(2), 231-250.

Klein, G., Moon, B., & Hoffman, R. R. (2006a). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, *21*(4), 70-73.

Klein, G., Moon, B., & Hoffman, R. R. (2006b). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, *21*(5), 88-92.

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.

Newsome, E., Wright, C., Klein, G., Flory, J., & Baker, A. (2015, October). *Using the ShadowBox™ method to detect the 'Investigator' and 'Proceduralist' mindsets in frontline social workers*. Poster presented at the Human Factors and Ergonomics Society Annual Meeting, Los Angeles, CA.

Schmitt, J. F. (1994). *Mastering tactics: A tactical decision games workbook*. Quantico, VA: Marine Corps Association.

U.S. Department of Defense, Defense Advanced Research Projects Agency. (2011). *DARPA-BAA-11-32: Strategic Social Interaction Modules (SSIM)*. Retrieved from https://www.fbo.gov/index?s=opportunity&mode=form&id=056e2ac62485ca5f2598e739e5761b01&tab=core&_cview=1

Wiltshire, T. J., Neville, K. J., Lauth, M. R., Rinkinen, C., & Ramirez, L. F. (2014). Applications of cognitive transformation theory: Examining the role of sensemaking in the instruction of air traffic control students. *Journal of Cognitive Engineering and Decision Making*, *8*, 219-247.

Gary Klein received his PhD in experimental psychology from the University of Pittsburgh in 1969. He was an assistant professor of psychology at Oakland University (1970-1974), a research psychologist for the U.S. Air Force (1974-1978), and founded Klein Associates (in 1978), which he sold in 2005. His books include *Sources of Power: How People Make Decisions* (1998) and *Seeing What Others Don't: The Remarkable Ways We Gain Insights* (2013).

Joseph Borders holds a BA in psychology from Wittenberg University. He is a research associate and project manager with MacroCognition LLC and ShadowBox LLC. He has assisted in the development of the ShadowBox approach and designed ShadowBox training evaluations for organizations including Center for Operator Performance (COP) and the Defense Advanced Research Projects Agency (DARPA).