Training Research in the Wild

Melinda K. Seibert, Frederick J. Diedrich
Jean MacMillan
Aptima, Inc.
Woburn, MA
mseibert@aptima.com, diedrich@aptima.com
macmillj@aptima.com

Gary E. Riccio
The Wexford Group International
Columbus, GA
griccio@thewexfordgroup.com

ABSTRACT

Training researchers may need to make a choice between conducting studies in a laboratory environment or a real-world operational training environment. There are pluses and minuses to each approach related to internal validity, external validity, and application. To help address these challenges, we have adopted a “use-centered basic research” (Stokes, 1997) approach to conducting operational training research. We strive to answer practical questions while simultaneously seeking to develop fundamental principles. This paper analyzes the challenge of studying training in the wild through the exploration of a variety of issues such as learning from each other in the context of the research, the need to leverage diverse expertise, the benefits of programmatic research, and the advantages of conducting use-centered research. Based on these issues, we discuss lessons learned from conducting research in various operational training environments – from field-based and classroom-based training in the Army to simulator-based training in the Air Force and Navy. In doing so, we focus on the critical meta-scientific issues that were revealed through this research rather than specific research results. In particular, we highlight key best practices for conducting high-quality, insightful, and practical research in operational training environments, focusing on interaction and collaboration within the operational community. We conclude that through this “use-centered basic research” it is possible to solve current problems facing Warfighters while conducting research in which theories with broader implications are developed.

ABOUT THE AUTHORS

Melinda K. Seibert is an Industrial-Organizational Scientist at Aptima, Inc. Ms. Seibert has experience in performance measurement, evaluating the effectiveness of training, identifying training needs, and conducting quantitative data analysis. Ms. Seibert holds a M.S. in Applied Psychology from Clemson University.

Frederick J. Diedrich is the Chief Operations Officer at Aptima, Inc. Dr. Diedrich focuses on performance measurement, as well as training and educational system design and development. He holds a Ph.D. in Cognitive Science from Brown University.

Gary E. Riccio is a Principal Scientist with the Wexford Group International. For 25 years, Dr. Riccio has worked primarily in trans-disciplinary teams and programs of research on training and education supporting human performance in hazardous or lethal environments. He holds a Ph.D. in Psychology from Cornell University.

Jean MacMillan is the Chief Scientist at Aptima, Inc. Her 25-year research career has spanned a broad range of topics in human-machine interaction and user-centered system design, including adaptive instructional design, team decision making, command center design, and human performance measurement in simulation environments. She holds a Ph.D. in cognitive psychology from Harvard University.

Training Research in the Wild

Melinda K. Seibert, Frederick J. Diedrich
Jean MacMillan
Aptima, Inc.
Woburn, MA
mseibert@aptima.com, diedrich@aptima.com
macmillj@aptima.com

Gary E. Riccio
The Wexford Group International
Columbus, GA
griccio@thewexfordgroup.com

INTRODUCTION

This paper summarizes our experience in conducting research in operational environments and shares the best practices we have learned in scientific collaboration within a community of practice that is focused on training and education. This work has been about building a particular kind of relationship over time, consistent with recent thinking in the sociology of science (see e.g., Gibbons, 1999; Flyvbjerg, 2001). The unique feature of these relationships is that scientists have continuous visibility, over a period of years, into the meaning that is made of their research and the consequences of the decisions influenced by their recommendations. Scientists involved in this type of collaboration enjoy accountability within a community of practice. This accountability guides their reasoning and informs their evolving program of research.

In the best examples of such collaboration, all stakeholders also have continuous visibility and opportunities for influence on the process of scientific inquiry (Gibbons, 1999). We have benefited from this "use-driven" research that presents unique opportunities for discovery within a context of application (see Stokes, 1997). Moreover, we have also observed that the operational community benefits from this knowledge production process because it is open and transparent and they have a true role in shaping knowledge that addresses their concerns. The traditional distinction between research, development, and transition is blurred. Instead, useful and relevant insights developed with the Warfighter occur early, often, and throughout the process. This collaborative research is "socially robust" (Gibbons, 1999) because there is a balance of choice and responsibility within and across all participants.

Equal Partners in Knowledge Production

An essential characteristic of the relationships we seek to establish and sustain is that the perspectives of all stakeholders are well understood, appreciated, and integrated with our own perspectives. Only in this way can subject matter expertise really be understood for what it is and be used in the right way. Subject matter experts' (SME) backgrounds should represent a wide range of relevant recent experiences, and SMEs must be on equal footing with researchers. SMEs of varying background often provide differing and at times conflicting opinions and these differences actually help us understand the "context of implication" for our research (Gibbons, 1999). Our method therefore, is to work with individuals from varied backgrounds, experience, and specialties (Goertzen, 2010; Quinlan, Kane, & Trochim, 2008).

We refer to the "context of application" as the setting and events that motivated the research and within which its products and recommendations would be used. A deeper understanding of context of application generally leads to a fuller appreciation of the implications of any changes that the research may motivate. In research on training, the implications include, for example, the downstream impact of training on behavior and performance on the job, or the collateral activities of training support elements or other related curricula. Following Gibbons (1999), we refer to this as the "context of implication." A deeper understanding of the context of implication for research generally reveals other contexts of application that can benefit from the research.

As the examples in this paper will illustrate, in our work on training and education we seek to establish collaboration within a community of practice where various complementary roles and perspectives are equally valued and understood (Goertzen, 2010; Watanabe, 2010). Such appreciation and comprehension are essential to knowledge creation. This goes far beyond simple "knowledge elicitation" and contrasts with the objectives of many approaches and research methods that seek to respond to requirements from outside their immediate community. As Figure 1 shows, the traditional method of inquiry involves the scientist responding to requirements and developing operational tools by pushing an insufficiently-contextualized product or idea to the context of application. The primary interaction between scientists and Warfighters is decontextualized.

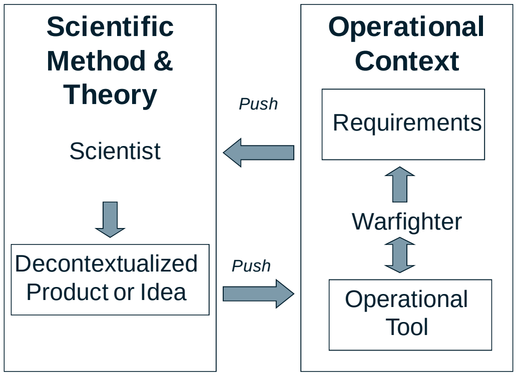

Figure 1. Traditional method of scientific inquiry.

In contrast, our intent in collaborating with SMEs and key stakeholders, as shown in Figure 2, is not to document knowledge based on the experiences of the operators, but rather to *create knowledge*, through the collaborative process. There is a collective knowledge development in which the whole is greater than and different from the sum of the parts and in which the perspectives of all participants in the collaboration can be changed by the experience (Mâsse, Moser, Stokols, et al. 2008; Smythe & McKenzie, 2010).

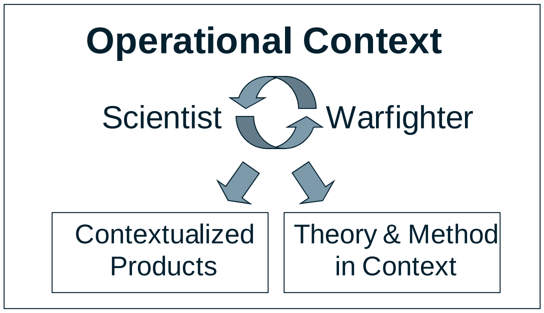

Figure 2. Collaborative, use-centered approach to scientific inquiry.

Adapting to Unknowns in the Wild

Our use-driven research (Stokes, 1997) often occurs in settings in which we have little or no control over the evolution of events (i.e., "in the wild"). Our "participants" are therefore true participants in the sense that their own intentions and engagements within the situation influence the setting and the evolution of events as much as the setting and events influence them. The investigators in our research are often not simply passive observers. They may actively intervene by making contributions to the research setting that, in turn, influences their observations. This of course is taken into account in the interpretation of our observations and in the identification of implications. More importantly, such embedding of scientists in a natural settings often leads to a focus on details about the context of implication that would be missed without such presence, or that would be considered insignificant without the grounded theory to which presence generally leads (Riccio et al., 2010).

Opportunities for Research the Wild

In this paper, we present our approach to collaborative, use-centered research in the context of four specific examples of our training, education, and assessment work with the U.S. Army, Navy, and Air Force. We emphasize two key themes in these examples: (a) collective knowledge development through focused collaboration within a diverse community over time, and (b) developing scientific theory in context with multiple stakeholder perspectives to solve operational needs of the Warfighter while contributing to the understanding of basic principles. These themes deliberately blur the distinction between research, development, and transition, in a manner that provides immediate value to *both* the scientific and the operational communities.

Through the examples outlined below we develop several recommendations regarding best practices for conducting training research in the wild that is truly use-driven and collaborative, with a focus on meaningful interaction with SMEs. This approach benefits the Warfighter by providing useful products and traceable conclusions. It also provides longer-term value by enabling theoretically coherent and empirically driven research in which Warfighters are inextricably linked with the products and conclusions of the research.

EXAMPLES

In this section, we describe four varied experiences conducting use-centered, collaborative research in training and education contexts. Our experiences cut across the Services and provide clear examples of how this approach can be applied in a variety of domains. Specific benefits to the Warfighter as a result of the application of this approach are noted.

Learning from Each Other: Battle Command Decision-Making Expertise

In military training settings, understanding and measuring the nature of the expertise to be trained is critical in designing training and assessing its effectiveness. A significant investment is constantly being made in training exercises, yet there is often no widely accepted reliable, valid way to measure the effects of these exercises. This is especially true for more senior personnel and thus there is often no way to determine return on the investment. A research project sponsored by the Army Research Institute, (Serfaty, MacMillan, Entin, and Entin, 1997) directly addressed this challenge of developing a reliable and valid quantitative measure of *battle command decision making expertise*. This is defined as the ability to make and implement decisions in a timely, efficient, and effective manner, most often with limited information, in a fluid and multidimensional battlespace.

Our research began in the usual manner by interviewing SMEs. In this case the SMEs were considered "superexperts"—retired general officers identified by their peers as extremely effective decision makers who were also serving as Senior Mentors in Army training exercises. Although the SMEs agreed that Army officers varied in their level of battle command decision making expertise, that this expertise was developed over time, and that military rank alone was not a reliable indicator of such expertise, they agreed on little else. Each of them described the nature of expertise in a different way, and they were highly skeptical that a quantitative measure could be developed that would reliably assess an individual's expertise level for such a complex ability.

At this point, the research became both more innovative and more collaborative, as described by Gibbons (1999). The cognitive scientists on the team believed that reliable and valid measures could be developed even in a domain where there was not an agreed upon definition of the concept to be measured. The SMEs on the team were skeptical about the measurement, but they did believe that battle command decision-making expertise existed and was observable. Based on their years of experience as Senior Mentors responsible for training officers, they believed that they "knew it when they saw it."

The method used to develop and validate expertise measures started with the development of a hypothetical scenario in which participants were presented written materials and a map that described a situation set in the Persian Gulf—a scenario developed in close collaboration with an extremely experienced retired officer. Four tactical situations were developed based on the general scenario. Acting as Division or Brigade commanders, participants were asked to recommend appropriate actions in each of the tactical situations. Participants were permitted to ask questions about the situation, and they then prepared a written statement of commander's intent as well as orders/messages to subordinates. Participants then explained the rationale for their chosen course of action (COA), which we recorded on videotape.

The next step was to determine whether the panel of "superexpert" judges could reliably distinguish different levels of expertise among the experiment participants—46 Army officers ranging in rank from Captain to General—based on their chosen course of action in the scenario. It is important to note that the scenario was a difficult one, and that there was no "textbook solution" or single right answer in developing the COA.

The panel of judges was asked to independently rate the expertise of each of the participants on a seven point scale (from 1 = novice to 7 = expert) based on (1) the written statement of intent and orders, and (2) the videotape in which the participant explained his (all participants were male) course of action. After viewing and rating both the written materials and the videotape, judges were asked to rate the participants' overall expertise. By design, the judges were not supplied with a definition of expertise, but were asked to explain their ratings using any factors they felt were relevant.

The judges' expertise ratings were remarkably consistent. Coefficient alpha (Cronbach, 1970) was .81 (out of a maximum of 1.0) for the overall expertise rating. Judges gave very similar independent ratings of the expertise of each participant. Their explanations for *why* they were giving the ratings varied widely, however. They did indeed "know it when they saw it," although they could not consistently describe what exactly they saw. The judges' ratings were not only consistent, but they discriminated successfully across a range of expertise levels.

Using the expert judges' ratings as the "gold standard" for determining a participant's level of expertise, the cognitive scientists were then able to test a variety of hypotheses from a theoretical framework based on the expertise literature (Serfaty et al., 1997). For example, a number of measures supported the hypothesis that experts build and use a richer and more complex mental model of the situation, and that the experts use this mental model to visualize and predict outcomes, and to act under uncertainty.

As this example illustrates, the effort was a true collaboration between two groups with very different backgrounds and perspectives. The experience proved to be an interesting one for the judges because the systematic approach of the scientists provided them with insight they would not otherwise have had into the skills and knowledge of the Army officers that they were helping to train. The judges typically worked with small groups of officers in a variety of different exercises, and they had never seen a large group of officers all try independently to solve the same problem. Although there was no one right answer for the COA, the judges noted that they saw areas where, to their surprise, many of the participants consistently lacked knowledge. The judges resolved to focus more intensely on those areas in future training exercises.

The cognitive scientists benefited as well because none of the findings related to explanations of expertise would have been possible without the measure that the expert judges were able to provide. In addition, the cognitive scientists were surprised that the simple diagnostic technique of giving everyone the same problem and seeing where people had common difficulties was innovative for the judges. It was a learning experience and a mutually beneficial exchange for both groups. Through this back and forth approach of learning and applying creative techniques to understand a challenging construct, we showed that decision making training provided to Army officers can be based on both theoretically sound and operationally relevant knowledge and skills.

Leveraging Diverse Expertise: Evaluating Effects of Interventions in Submarines

As part of a research program sponsored by the Naval Sea Systems Command, we developed measures of individual actions and team interactions in order to understand the impact of interventions (technology or training) on overall operator effectiveness (for a more complete treatment, see Jackson, Woods, Durkee, O'Malley, Diedrich, Aten, Lawrence, & Ayers, 2008).

Our approach to measure development focused on employment of the Competency-based Measures for Performance Assessment Systems (COMPASS) process (MacMillan, Entin, Morley, and Bennett, in press). The method employs an intensely interactive process in which operators work directly with scientists to identify what needs to be measured. In the case addressed here, the ultimate outcome of three years of collaboration with SMEs was a comprehensive set of approximately 200 measures focusing on the behaviors of Fire Control Technicians (FTs) and the Officer of the Deck (OOD) across selected mission applications.

Our work began with the formation of the team used to create the measures. We employed a combination of active-duty and retired personnel, working with scientists and engineers from Aptima and the Naval Undersea Warfare Center Division Newport. Hence, a key element of our approach was the utilization of personnel from multiple specializations who interacted across roles. For instance, when our focus was on the OOD during the measure development process we included input from FTs and Sonar. Similarly, when investigating FT duties we included input from the OOD and Sonar. The reason for this was to address teamwork by purposely putting Officer and NCO perspectives as well as different specialties in opposition. In other words, our selection and utilization of a particular diversity of SMEs was principled and traceable to their interdependence in operational tasks.

The opposition of Officer and NCO allowed us to explore not only local FT behaviors, but also how their behaviors are viewed and impacted by others within the team. We found that this was critical because in the discussions, there were numerous opportunities for the team to explore issues more thoroughly and work through disagreements. Although the objective of the workshop process was to develop measures, the military participants learned from each other by having to specify their mutual expectations regarding processes and outcomes. This clarity was intended to be for the scientists, but this exchange enabled the SMEs to learn about alternative views that came from the need to explain their thoughts to scientists who were naïve in the domain. For instance, FTs expressed situations in which their ability to support the OOD would be enhanced through different OOD behaviors, and vice versa. The assessment development process had the effect of creating knowledge for the participants as well as the scientists.

Through these workshop discussions it also became apparent that the whole was greater than the sum of the parts. The operators brought extensive domain expertise. The scientists brought expertise in measurement and in associated theoretical explanations of human performance, primarily in the area of teamwork (e.g., back-up and monitoring of teammates, pushing of information, etc.; see Smith-Jentsch, Johnston & Payne, 1998). Thus, while the measures developed were in the language of the operators with task and mission specific wording, theories of teamwork also permeated the measures. The resulting measures captured critical insights in the provision of information in a manner not previously realized by the operators. If the scientists alone had created the measures, they would have lacked domain specificity. If the operators alone had created the measures, they would have lacked a theoretical basis linked to team performance that may have lead to inconsistencies or gaps. In other words, theory was expressed in meaningful operational terms and operationally relevant observables were presented in a systematic and traceable framework. The process involved knowledge creation, not merely knowledge elicitation.

A second key element of our approach was a sustained involvement of the team in a collaborative process that lasted over a period of years. This work was in sharp contrast to some knowledge elicitation efforts in which SMEs are interviewed, knowledge is "extracted," and the collaboration ends. Instead, in the context of our work in undersea warfare, the collaboration began with the COMPASS process and included a test of the measures in a laboratory setting with active duty and retired operators (Jackson et al., 2008). This work then continued on to include measure revision, testing in an operational training setting with an actual crew, and the creation of additional measures. As Gibbons et al. (1994) note, this form of sustained collaboration evolving through interaction with additional individuals over time is characteristic of emerging research paradigms in which teams assemble to address problems, dissemble, and reassemble in such a way that knowledge is built up over time and dispersion of knowledge occurs as participants interact (see also, Flyvbjerg, 2001; Mâsse et al. 2008; Quinlan, Kane, & Trochim, 2008; Smythe & McKenzie, 2010).

The element that made our approach effective was a sustained collaboration over several years. Some key members of the team that created the initial measures also participated in data collection used to test and refine the measures (Jackson et al., 2008). Critically, interpretation of results was a joint process in which the operators helped to explain apparently ambiguous findings. As Jackson et al. (2008) noted, for instance, patterns of tool use showed that higher performing FTs cycled through various tools to refine solutions at a frequency that was much higher than lower performing FTs with the exception of one tool, which was used more frequently by the lower performing FTs. Initially, the scientists viewed this one instance as a potential outlier. However, the operators were able to explain why this particular tool might be used more in order to verify ambiguous solutions as the FTs struggled to obtain more certainty.

In the context of this effort, knowledge and understanding emerged from the sustained interaction of individuals with varied backgrounds. Scientists did not work alone as scientists, and operators did not work alone as operators. The result was a sort of local transition in which the participants both contributed to and gained from the interaction (Quinlan et al., 2008; Mâsse, et al. 2008). Moreover, this reciprocal coming to know was captured in assessments that can be transferred to learning environments in which they can be used to guide learning through theoretically and operationally sound insights.

Developing a Collaborative Research Program: Outcomes Based Training and Education (OBTE)

A habitual relationship between scientists and operators in a community of practice reflects a commitment to collective knowledge development that transcends the objectives and demands of specific projects. Such sustained relationships provide the ability to see the deeper context of implication for the collaboration (Gibbons, 1999; see also Flyvbjerg, 2001; Wenger, 1998). That is, participants discover issues and opportunities that were not anticipated and that most likely would not have been noticed without a diversity of perspectives. Such implications can influence concurrent and subsequent research as well as activities in the operational community. Thus, the research can become programmatic even in the absence of a unitary programmatic line of funding (cf., Quinlan, Kane, & Trochim, 2008).

Serendipitous influences on scientific inquiry and its impact within a context of application are not commonly documented, especially when the influence is interpersonal and meta-scientific (Mâsse et al., 2010; Smythe & McKensie, 2010; Watanabe, 2010). In this section, we describe some high-impact meta-scientific influences across three different research projects related to our interest in measuring values-based outcomes in training and education (Riccio et al., 2010). Making such meta-scientific influence more explicit is an emerging best practice in socially robust research. We believe it is an important aspect of what it means for research to be programmatic.

In a project for the U. S. Army Research Institute (ARI), we demonstrated that it is possible to identify relationships between abstract values (e.g., "I will never leave a fallen comrade") and the concrete behavior of Soldiers in ostensibly mundane training activities (Brunyé, Riccio, Sidman, Darowski, & Diedrich, 2006; Riccio, Sullivan, Klein, Salter, & Kinnison, 2004). We were able to elaborate on the definition of Warrior Ethos in ways that were both verifiable in the context of Initial Entry Training (IET) and valid with respect to the implications of training for a future operational context.

In addition to various levels of expertise in the behavioral and social sciences, our collaborative inquiry in this work included participants who were involved in all levels of command and execution of IET at Fort Benning and Fort Jackson, as well as military experts who had relevant experience with the needs of commanders in the operational Army (i.e., the context of implication). These operational experts outside the immediate context of application helped us to appreciate other contexts of application that we did not anticipate prior to the project and that we were not able to address until a later project. Examples of a broader context of application for values-based measures are discussed below. This broader context of application enabled a more programmatic approach to values-based measures that extended beyond the initial project on Warrior Ethos.

The Warrior Ethos project focused on measuring values-based intangibles in trainee behavior (e.g., perseverance, adaptability, sense of calling). It also revealed the need for guidance to instructors that, unlike the focus of a typical training support package (TSP), helped them understand how they could have a more positive impact on the development of trainees with respect to intangibles. It was not until a later project with ARI, however, that we were able to develop formative measures (those designed to guide learning) of instructor behavior in the context of Army training (Sidman, Riccio, Semmens, Geyer, Dean, & Diedrich, 2009). In this latter work, we focused on instructors in the Basic Non-Commissioned Officer Course (BNCOC) at Fort Benning. We were able to utilize the COMPASS process for collaborative development of measures of instructor behavior with respect to the development of a different set of values-based intangibles that was being utilized in BNCOC (i.e., Warrior Leader, an Ambassador, a Critical and Creative Thinker, a Leader Developer, and a Resource Manager).

During our BNCOC work, one of the operational experts on our BNCOC research team was also working on a values-based approach to training (i.e., Outcomes Based Training and Education [OBTE]) with the Asymmetric Warfare Group (AWG). Through his participation in the COMPASS process for BNCOC, he immediately recognized how it could be applied in further development of OBTE. Ad hoc discussion surrounding this realization stimulated interactions between the AWG and BNCOC at Fort Benning with respect to the development of values-based outcomes in a training environment (e.g., confidence, initiative, accountability). As a result of these discussions, the AWG began to utilize the COMPASS process in the definition of OBTE, in field-based verification of its implementation, and in validation of its potential impact with respect to values-based requirements in the Army (Riccio et al., 2010).

There were discrete phases in our research on OBTE, and the break between phases enabled us to engage with the command chain at various sites where OBTE was being implemented (e.g., Fort Benning, Fort Jackson, Fort Sill). By sustaining our presence in the context of application for OBTE, we came to appreciate some important implications. We learned about the different meanings that leaders in various programs of instruction gleaned from research pertaining to their responsibilities. We view this as analogous to the way Army leaders, in general, interpret the intent of their higher commanders in somewhat idiosyncratic ways based on peculiarities of the situations within their span of control. Our research products and recommendations thus had to explicitly address such differences in interpretation and their implications for command decisions in particular programs of instruction. The measures we developed for intangible outcomes of training and education, for example, had to strike a balance between utility (e.g., concreteness and directness of application in planning and execution of training) and usability (e.g., flexibility in prioritization, selection, and interpretation of measures) (Riccio et al., 2010).

Balance between concreteness and flexibility is justifiable to the extent that there are multiple ways for subordinates to execute on the intent of a higher commander and given that there also are principled boundaries on such initiative (cf., Freeman, Jason, Aten, Diedrich, Cooke, Winner, Rowe, & Riccio, 2008). In our research on OBTE, we thus devoted a significant effort to developing a grounded theory for the measures and their use with respect to which changes could be negotiated given application to a particular context (Riccio et al., 2010). The development of grounded theory was not one of the initial objectives of the research. The necessity became clear as we delved deeply into the personal meaning OBTE holds for various individuals within the community of practice; that is, by allowing the research to be confronted with the context of implication.

The reciprocal influence between the evolving theory of values-based behavior and the findings about the various practices in training and education moved us systematically toward a theory of practice. In particular, our interactions with participants and stakeholders directly influenced the development of measures in our broader program of research. The findings from field-based observations of training and education, in turn, influenced the dialog with participants and stakeholders by providing increasingly clear and relevant questions. As a result, the research products and recommendations have had direct impact on a wide variety of programs of training and education in the Army, many of which are due to implications that go far beyond the original intent of the research (e.g., Riccio et al., 2010).

Conducting Use-centered Research: Fidelity Requirements for Training Simulators

Researchers have long been studying the impact of fidelity dimensions on performance in flight simulators (e.g., Bradley & Abelson, 1995; Cress, McMillan, & Gilkey, 1989; Winterbottom, Geri, Pierce, & Harris, 2001). At the same time, operational users have long been interested in knowing the appropriate level of simulator fidelity to implement in simulators. While researchers typically construct careful, controlled evaluations of the impact of differences in fidelity on performance, acquisition and training professionals generally turn to end-users to identify fidelity requirements for training simulators specific to their training context.

With years of research on the impact of fidelity on performance in simulators, why is it that there remains little guidance and no standard tool available to facilitate and support these decisions? We suggest that it may be in part due to scientist's lack of consideration for the contexts of application for simulators in military training. Thus, while providing valuable insights into which aspects of fidelity impact performance, scientists have been unable to—on their own—answer the questions asked by individuals who must employ simulators to meet operational needs. As scientists, we are often quick to develop solutions before we have taken the time and effort to fully understand the complexity of the operational problem. Before we can answer the complex questions, we must take the time to understand the training objectives, the knowledge and skills required to complete training objectives, and the full context in which operators use simulators in the training environments.

Our work defining simulator fidelity requirements illustrates our collaborative, multidisciplinary approach to solving a real-world problem while at the same time providing meaningful results that apply and extend the scientific community's current understanding of the impact of simulator fidelity on specific outcomes. This example also illustrates how "use-centered basic research" (Stokes, 1997) has facilitated knowledge production beyond that possible in more traditional research methods and approaches.

Aptima first worked with the Air Force to develop a standard approach and tool for matching training objectives to the training device with the most appropriate fidelity – from lower-fidelity simulators, to higher-fidelity simulators, to actual training in the aircraft. Working in close coordination with F-16 pilots and Air Force Researchers at the Air Force Research Laboratory in Mesa, AZ, we applied the RELATE (Relating Effective Learning to Attributes of the Training Environment) approach (See Estock, Alexander, Gildea, Nash, & Blueegel (2006) for thorough description). The RELATE approach combines fidelity requirements defined by end-users in accordance with training objectives and required knowledge and skills, existing theory and research about fidelity, and objective performance data from fidelity experiments to develop a predictive, computational model. Combining scientific theory with training objectives ensures that there is not only a strong theoretical basis but also relevance to operational missions. After conducting an extensive review of the scientific and technical literature on fidelity, we worked with F-16 pilots to map training objectives onto fidelity dimensions culled from the literature. We also identified key knowledge and skills required to meet training objectives and with F-16 pilot collaboration, developed hypotheses regarding the impact of fidelity on the ability of simulators to effectively train these knowledge and skills. We then conducted training effectiveness experiments comparing different levels of visual and cockpit fidelity at AFRL in Mesa, AZ (Estock, Alexander, Stelzer, & Baughman, 2007; Estock, Stelzer, Alexander, & Engel, 2009).

Because we took the time to understand the simulator context and dissect simulator fidelity, we provided a way for acquisitions and training professionals to develop an integrated strategy for employing both high- and low-fidelity simulators to meet training objectives. Throughout our entire process, our multidisciplinary team consisting of human factors scientists, software engineers, industrial-organizational psychologists, and mathematical modelers worked side by side with end-users and pilot subject matter experts to ensure the end result was not only theoretically sound but operationally relevant and useful.

During our work with the Air Force, the Navy was also experiencing a similar problem – they too wanted to know how to determine the appropriate level of fidelity of training simulators to achieve specified training objectives, while maintaining trainee acceptance, and fitting within budgetary constraints. We once again applied the RELATE approach, and worked side by side with F/A-18 pilots to ensure everything from our fidelity hypotheses to performance measures aligned with current Naval Aviation training objectives. By involving operators at each step in the process, we ensured that our solution would address the real questions underlying the Navy's need for a decision-making aid, which we only learned through close communication and a working relationship with operational users and end-user decision-makers alike.

We conducted training effectiveness research at NAS Lemoore during Air-to-Ground simulator training events with F/A-18 pilot trainees (see Figure 3). F/A-18 Instructor Pilots provided ratings on pilot trainee performance using customized measures designed to differentiate high and low performing aviators in Air-to-Ground missions that were developed using the COMPASS methodology.

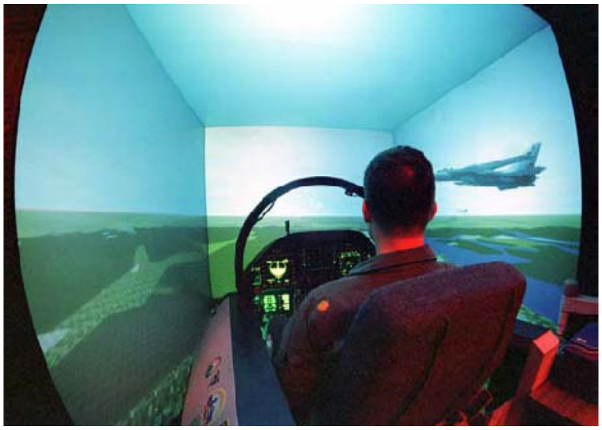

Figure 3. F/A-18 simulator at NAS Lemoore used in training effectiveness research.

This research provided a unique and valuable opportunity to observe and evaluate training "in the wild." We were able to conduct our empirical inquiry in an environment that cannot be completely and wholly contextually replicated in a lab environment (Gibbons, 1999), and we were able to obtain first-hand knowledge about the context of implications for our research. The data we collected in this training environment accounts for contextual variables like simulator malfunctions, instructor pilot differences, and time constraints. As a result, the data provides a realistic indication of the impact of simulator fidelity on training effectiveness as it occurs in a live training context.

RECOMMENDATIONS

Whether you are an Army officer seeking to understand how to train decision-making or adaptability, an analyst seeking to measure improvements in submariner performance in a diverse team, a simulator or training designer seeking to understand required levels of simulator fidelity for effective training, or someone else facing an entirely different operational need, one thing is certain: Conducting research in the wild will provide an environment that is complex, difficult to understand, and impossible to wholly replicate in a laboratory environment alone. Our varied experiences have shown that the key to working effectively in such an environment is meaningful habitual collaboration with a diversity of operational stakeholders. When done well, such an approach can lead to creative, effective, and novel solutions that address the needs of today's Warfighters. Our experiences have shown that this approach can also lead to challenges that are different from those in more controlled environments. In light of these challenges, we conclude by recommending several best practices for the conduct of use-driven, collaborative training research in the wild.

Recommendation #1: Collaborate with Warfighters.

In each of our examples, collaboration was critical to the success of the project and to ensuring maximum benefit to the Warfighter. Not only must scientists with varied backgrounds be represented in research, but also the operational users who interact in real life. It has often been through intentional, reciprocal influence of meaningful collaboration that our best solutions surfaced. From our work with undersea warfare and OBTE in particular, we offer the following suggestions to promote the investigation of training in the wild: First, those participating in the process of coming to know should be from varied backgrounds, carefully selected so as to identify a non-arbitrary range of implications relevant to the problem at hand. Second, every part of the project should be collaborative in the sense that emerging implications should be visible to all participants and all participants should have an opportunity to influence the direction of the collaborative inquiry, even if there is not a requirement to do so.

Recommendation #2: Think long-term.

To obtain a deep, complete understanding of the operational needs facing the Warfighter, sustained relationships with stakeholders over a period of time is often required. Even in the absence of a single funding source, it is possible to apply a programmatic approach. In our OBTE work, for example, we had several sources of funding for the application and extension of the same essential set of concepts and challenges across different local applications. More importantly, appreciation of the context of implication in one immediate context of application helped us realize opportunities in a different context of application and it had a fundamental impact on design of the research in that context. The same is true of our work identifying fidelity requirements.

Long-term thinking emphasizes research that is iterative and programmatic, to the extent possible. Visibility of assumptions, findings, and interpretations of the research provides participants with opportunities to influence the research. The consequence of such influence meaningfully situates the collaborative research in a context of implication. Participants in such collaborative research develop a much deeper sense of accountability for their perspectives, contributions, and impact. This sometimes requires a level of due diligence that goes beyond the initial objectives and scope of a research project. The necessity of this broader commitment is perhaps one of the unique requirements of "applied" research that aspires to be programmatic and scholarly. In our experience the effort is well worth it.

Recommendation #3: Consider the implications.

Use-centered research requires consideration of the implications of the research. If we fail to understand the real, underlying problems that may or may not be apparent in the original reason for the funding of our research, we will likely also fail to discover important implications. Further, failure to gain insight into implications as the research evolves could result in "solutions" that do not actually solve the underlying problems. For example, in our work with OBTE we were able to engage a spectrum of leaders and influencers in the chains of command in various programs of instruction and in the larger U.S. Army Training and Doctrine Command (TRADOC) organization. This was critical in helping us achieve a sufficient understanding of reactions and needs with respect to programmatic decision-making and actual implementation of OBTE. This context of implication shaped the work in terms of decisions regarding our approach to assessment, and over time, we gained increasing visibility into factors in the larger context that are in the critical path for implementing any approach to training and education in the Army. The key, therefore, is to understand the implications of the research, and to keep these implications continually in mind as the research progresses. This approach differs from simply identifying and responding to requirements, for it implies continuous engagement with the broad operational context of the work.

Recommendation #4: Conduct research in the wild, and document observations.

Anytime a researcher is unable to control the research setting, they would be remiss if they did not take copious notes regarding observations, challenges, possible interactions and confounding variables. We suggest that in unconstrained settings, researchers should document assumptions, observations, and interpretations of factors that emerge whether planned or not. We are not suggesting that empirical research in the wild always be as extensive as ethnography but more modestly that it should have ethnographic sensibilities. For instance, in our work on fidelity, we were unable to easily run fully controlled laboratory studies. However, by necessity, the data we collected addressed contextual variables like simulator malfunctions, instructor pilot differences, and time constraints—that is, the inconveniences of the real world. Rather than being viewed as a confound or compromised data, we believe that the often unexpected contextual observations provided a more realistic indication of the impact of simulator fidelity as it occurs in a live training context. Rigorous documentation of contextual findings, collected in the wild, makes this type of insight possible.

Recommendation #5: Be willing to adapt.

As in the expert decision-making example, the approach you begin with may not be the approach you end up with. By working with stakeholders to develop a workable approach, you may end up with a product far more useful and relevant than you anticipated. Because understanding operational users and needs is so important, we may need to shift our approach in the course of an investigation to suit the changing needs of operational missions and goals. Just as with approaches to military engagements, our research must adapt to different challenges we encounter along the way rather than respond only to a set of issues identified prior to sustained engagement.

ACKNOWLEDGEMENTS

We thank the many collaborators, including military personnel, government sponsors, academics, and industry scientists that contributed to the ideas and insights reflected here that came about through deep commitment and sustained partnership.

REFERENCES

Bradley, D. R., & Abelson, S. B. (1995). Desktop flight simulators: Simulation fidelity and pilot performance. *Behavior Research Methods, Instruments, & Computers, 27*, 152-159.

Brunyé, T., Riccio, G., Sidman, J., Darowski, A., & Diedrich, F. (2006). Enhancing warrior ethos in initial entry training. *Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society*, San Francisco, CA.

Cronbach, L. J. (1970). *Essentials of Psychological Testing* (3rd edition) New York: Harper and Row.

Cress, J. D., McMillan, G. R., & Gilkey, M. J. (1989).

The dynamic seat as an angular cuing device: Control of roll & pitch vs. the control of altitude & heading. *Proceedings of the AIAA Flight Simulation Technologies Conference* (pp. 94-100).

Estock, J. L., Alexander, A. L, Gildea, K. M., Nash, M., & Blueggel, B. (2006). A Model-based Approach to Simulator Fidelity and Training Effectiveness. *Proceedings of the 28th Annual Interservice/Industry Training, Simulation and Education Conference*, Orlando, FL.

Estock, J. L., Alexander, A. L., Stelzer, E. M., & Baughman, K. (2007). Impact of simulator fidelity on F-16 pilot performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 75-79). Santa Monica, CA: HFES.

Estock, J. L., Stelzer, E. M, Alexander, A. L, Engel, K. (2009). Is cockpit fidelity important for effective training? Perception versus performance. *Proceedings of the 31st Annual Interservice/Industry Training, Simulation and Education Conference,* Orlando, FL.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again.* Cambridge, UK: Cambridge University.

Freeman, J., Sidman, J., Aten, T., Diedrich, F., Cooke, N., Winner, J., Rowe. L., & Riccio, G. (2008). *Shared Interpretation of Commander's Intent (SICI).* Final Report to the Army Research Institute for the Behavior and Social Sciences, contract number W74V8H-06-C-0004.

Goertzen, J. R. (2010). Dialectical pluralism: a theoretical conceptualization of pluralism in psychology. *New Ideas in Psychology, 28,* 201–209.

Gibbons, M. (1999). Science’s new social contract with society. *Nature, 402, C81,* 11-17.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies.* London, UK: Sage Publications.

Jackson, C., Woods, H., Durkee, K., O’Malley, T., Diedrich, F., Aten, T., Lawrence, D., & Ayers, J. (2008). Tools for assessment of operator contribution to system performance. *Proceedings of the Undersea Human Systems Integration Symposium*, Bremerton, WA.

MacMillan, J., Entin, E. B., Morley, R, & Bennett, W. (in press). Measuring team performance in complex and dynamic military environments: The SPOTLITE method. *Military Psychology*.

Mâsse, L. C., Moser, R. P., Stokols, D., Taylor, B. K., Marcus, S. E., Morgan, G. D., Hall, K.L., Croyle, R.T., Trochim, W. (2008). Measuring Collaboration and Transdisciplinary Integration in Team Science. *American Journal of Preventive Medicine, 35(2, Supplement 1)*, S151-S160.

Quinlan, K. M., Kane, M., & Trochim, W. M. K. (2008). Evaluation of large research initiatives: Outcomes, challenges, and methodological considerations. In C. Coryn & M. Scriven (Eds.), *Reforming the evaluation of research. New Directions for Evaluation, 118,* 61–72.

Riccio, G., Diedrich, F., & Cortes, M. (Eds.) (2010). *An initiative in outcomes-based training and education: Implications for an integrated approach to values-based requirements*. Fort Meade, MD: U.S. Army Asymmetric Warfare Group.

Riccio, G., Sullivan, R., Klein, G., Salter, M., & Kinnison, H. (2004). *Warrior ethos: Analysis of the concept and initial development of applications*. ARI Research Report 1827. Arlington, VA: US Army Research Institute

Serfaty D., MacMillan J., Entin, E. E., & Entin E. B. (1997). The decision-making expertise of battle commanders. In C. E. Zsambok and G. Klein (Eds.), *Naturalistic Decision Making*. New York: Lawrence Erlbaum.

Sidman, J., Riccio, G., Semmens, R., Geyer, A., Dean, C., & Diedrich, F. (2009). *Reshaping Army institutional training: Current training.* Final Report to the Army Research Institute for the Behavior and Social Sciences, W74V8H-04-D-0047 DO 0010.

Smith-Jentsch, K.A., Johnston, J.H., & Payne, S.C. (1998). Measuring team-related expertise in complex environments. In Cannon-Bowers & Salas (Eds.), *Making decisions under stress: Implications for individual and team training*. Washington, DC: American Psychological Association.

Smythe, W. E., & McKenzie, S. A. (2010). A vision of dialogical pluralism in psychology. *New Ideas in Psychology, 28(2),* 227–234.

Stokes, D. (1997) *Pasteur’s quadrant: Basic science and technological innovation.* Washington, DC: Brookings Institution Press.

Watanabe, T. (2010). Metascientific foundations for pluralism in psychology. *New Ideas in Psychology, 28(2),* 253–262.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity.* New York: Cambridge University.

Winterbottom, M., Geri, G., Pierce, B., & Harris, N. (2001) Low-altitude flight performance as a measure of flight simulator fidelity. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting,* Santa Monica, CA.