Equal but Different: 5 Research Strategies for Improving Conclusions Drawn from Novice Populations

Stephanie J. Lackey, Lauren E. Reinerman-Jones, Julie N. Salcedo
University of Central Florida, Institute for Simulation & Training
Orlando, Florida
slackey@ist.ucf.edu, lreinerm@ist.ucf.edu, jsalcedo@ist.ucf.edu

ABSTRACT

Translating results from laboratory-based research studies conducted with novice participants (e.g. university students) to real-world applications represents a critical challenge facing researchers. Risk mitigation cannot wait until data is collected and analyzed, rather it must permeate every phase of the research process. Empirical evidence suggests that prudent application of fundamental human factors and training principles support experimental findings that equate to relevant recommendations for expert populations regardless of sample population experience. This paper presents five compelling strategies for conducting human participant research using novice populations that facilitate empirically sound insights for expert operators. Specifically, (1) designing experiments, (2) distilling skills into core components, (3) scaffolding, (4) proficiency testing, and (5) interpreting results will be discussed. The methods described represent the best practices in ongoing research efforts impacting highly specialized expert populations: Warfighters and nuclear power plant operators. The recommendations provided illustrate the potential that interdisciplinary experimental methods offer quantitative researchers.

ABOUT THE AUTHORS

Dr. Stephanie J. Lackey earned her Master's and Ph.D. degrees in Industrial Engineering and Management Systems with a specialization in Simulation, Modeling, and Analysis at the University of Central Florida (UCF). Her research focused on prediction, allocation, and optimization techniques for digital and analog communications systems. Dr. Lackey conducted high-risk research and development aimed at rapid transition of virtual communications capabilities to the Field and Fleet as a computer engineer with the United States Naval Air Warfare Center Training Systems Division (NAWC TSD). She joined UCF Institute for Simulation and Training's (IST) Applied Cognition and Training in Immersive Virtual Environments (ACTIVE) Lab in 2008, and assumed the role of Lab Director in 2010. Dr. Lackey leverages her experience in advanced predictive modeling to the field of human performance in order to develop methods for improving human performance in simulation-based training environments and human-robot interfaces.

Dr. Lauren E. Reinerman-Jones is an Assistant Research Professor at the University of Central Florida where she has extensive experience working with ARL, NRC, ONR, NAWCTSD, AFOSR, DoT, Florida Hospital, and the John Templeton Foundation. Her research centers on using physiological measures (EEG, ECG, TCD, fNIR, and eye tracking) for understanding, improving, and predicting human performance. She is Founder and CEO of DUJO, which applies cutting-edge science for skill assessment and improvement. Lauren has published and presented internationally in the fields of Psychology, Engineering, Philosophy, and Business. She serves on the editorial board of Theoretical Issues in Ergonomics Science (TIES) and previously as the HFES Augmented Cognition Technical Group Chair and on the HFES Board of Executive Officers.

Ms. Julie N. Salcedo joined the Applied Cognition and Training in Immersive Virtual Environments (ACTIVE) Lab as a Graduate Research Assistant in 2009. She holds a Bachelor's in Education, a Master's in Modeling and Simulation, and a Certificate in Instructional Design for Simulations all from the University of Central Florida (UCF). She is currently pursuing a Ph.D. in Modeling and Simulation from UCF. A former public school teacher, Ms. Salcedo leverages her education and instruction background to investigate learning and instructional design in simulation-based training systems.

Equal but Different: 5 Research Strategies for Improving Conclusions Drawn from Novice Populations

Stephanie J. Lackey, Lauren E. Reinerman-Jones, Julie N. Salcedo
University of Central Florida, Institute for Simulation & Training
Orlando, Florida
slackey@ist.ucf.edu, lreinerm@ist.ucf.edu, jsalcedo@ist.ucf.edu

INTRODUCTION

Research results produced in controlled laboratory studies suffer from a perceived lack of confidence in the ability to generalize findings from a novice population to expert populations situated in real-world environments. Criticism of the methods used to perform and control laboratory experiments typically stems from a concern that the participant pool excludes operational experts and/or a misunderstanding of experimental strategies available to account for differences between novice and expert participants. Scientific rigor requires researchers to vigilantly protect the integrity of experimental and analysis methods. However, this constraint does not inhibit the ability to bridge the gap between laboratory-based experiments and real-world environments.

Traditional research in the area of novice performance focuses on theoretical frameworks characterizing novice populations (Benner, 1984). For example, Benner (1984) presents a five-level paradigm defining and describing expertise. Such models benefit the practitioner when classifying individuals, tracking development, and developing training. Additional work focused on training evaluation (Salas, Milham, & Bowers, 2003), training implications (Carnahan, Lickteig, Sanders, Durlach, & Lussier, 2004), and technology design tradeoffs (Blavier, Gaudissart, Cadiere, & Nyssen, 2006) for novice population aid in stakeholder decision-making and system development. Finally, improving qualitative research measurement instruments (Hao & Houser, 2010; Kitzinger, 1995) advances the ability to elicit value-added insight from the general population. However, these efforts fail to address issues related to generalizing experimental results from novice populations to experienced or expert populations.

Benner's (1984) defines a novice as an individual possessing no experience with a given task. This perspective represents an elegant way to differentiate a novice from other expertise levels and offers a foundation upon which to target the strategies presented below. In human participant research, participation may be compensated (e.g., monetary compensation or class credit), but is completely voluntary. The influence of individual differences, including previous experience, can be addressed by the strategies presented.

A vast array of research investigating experts and expert performance in fields such as chess, medicine, (Ericsson & Lehmann, 1996) and sports (Ward & Williams, 2003) provide foundational work to be leveraged by other domains. Ericsson & Lehmann (1996) explain that experts devote significant portions of their lives (e.g., 10 years or 10,000 hours) to enhancing their performance of highly constrained tasks/skills and define expert performance as, "consistently superior performance on a specified set of representative tasks for a domain." Due to the limited number of individuals with 10,000 hours of deliberate practice in a particular task area (Ericsson, Krampe, & Tesch-Romer, 1993), some researchers rely on peer nomination or achievement awards to select expert participants. This is not appropriate for some domains such as combat or Nuclear Power Plant (NPP) operations. Furthermore, pursuing an expert population as defined by Ericsson & Lehman (1996) may, in fact, be the wrong choice for studies focused on Warfighter and NPP operators.

Researching kinetic combat tasks involving marksmanship, room clearing, or robot-aided Intelligence, Surveillance, and Reconnaissance (ISR) presents two critical obstacles: (1) the inherent safety risks of these roles jeopardizes the ability to acquire 10,000 hours of experience without interfering factors coming into play, and (2) next-generation robot platforms and interfaces do not yet exist in the field – thus, reaching the 10 year/10,000 hour threshold is not possible. Similarly, human factors research in the NPP domain suffers from a lack of expert operators for emerging control room interfaces. Existing plants were built decades ago. A dozen new nuclear power plants are scheduled for development in the near future and the control room human system interfaces represent a glaring departure from

existing plants. Even when expert operators are available, their previous experience interferes with experimentation focused on emerging displays representative of new control room configurations.

Alternatively, relative experts and relative novices serve as a participant selection and recruitment mechanism (Parikh, 2011). Such measures are a step toward leveling the proverbial playing field, but much more can be done to assure the internal and external validity of laboratory-based experiments aimed at informing real-world endeavors. Sadish, Cook, and Campbell (2002) summarize the importance of surface similarity, ruling our irrelevancies, and discriminating between laboratory and field research. Efforts to advance these concepts resulted in a framework by Fincannon, Keebler, and Jentsch (2012) to investigate hypotheses from real-world situations within laboratory settings by considering interpolation, extrapolation, and the importance of causal explanation for finalizing theoretical development. This paper specifies research strategies and practices that leverage the work of Saddish, et al. (2002) and Fincannon, et al. (2012) to improve empirical human participant research.

RESEARCH STRATEGIES

Designing Experiments

Experiments are employed to systematically investigate a phenomena. Of utmost importance to experimentation is variable control. This is critical for ensuring data quality, drawing reliable and valid conclusions within a given scope, and planning future research. However, the phenomenon in question is often within or from a complex, dynamic environment with many factors and variations. This conundrum has led to a divide in research camps: basic and applied. Basic research has traditionally been associated with theory driven investigations and sometimes viewed as more academic. Basic research is thought of as science for science sake. In contrast, applied research is perceived as problem focused with the intention of applying experimental findings sooner than later. However, the best science does not operate in a vacuum. Researchers often glean ideas from their own interactions in life and therefore, the questions that arise are applied. However, methodological approaches for gaining insight into their research questions is the point of contention. Understanding that difference is key to bridging the gap between laboratory research and the real-world.

The benefits and limitations of each camp might seem obvious, but bringing to light common arguments from each stance is helpful. Basic researchers argue that without absolute control, nothing can truly be known. Applied researchers counter with the notion that too much control creates an artificial circumstance in which the phenomenon would not occur and thus, the findings from basic research are not useful. To which basic researchers accuse that experiments without systematic rigor do not contribute to the literature and thus, the wheel is reinvented each experiment by discarding theory that was built on numerous studies. The inflammatory reply is that basic researchers are not concerned with anything other than publishing and could care less about “real” problem solving. Each perspective has some truth to their assertions, but often fail to listen to the other side to resolve the differences and execute synergistic research. Theory has a purpose, but so does application. Application has a purpose, but so does theory.

Researchers need to seek understanding from theory and the real-world with which the phenomenon typically occurs. In other words, a researcher should identify a problem that needs to be solved or understood from real-world experience. Then, he or she should seek input from subject matter experts, specifically those actually performing the task, supervising the performers, and/or setting policy or regulation. Those inquiries provide insight about the phenomenon at every organizational level - performer, management, and investor- thus, informing the full tasking environment. Simultaneously, the researcher should review literature available on the topic. The literature review should include journals, proceedings, technical reports, standards documents, magazines, newsletters, and other media. The objective for that initial review is exploratory to refine the question for investigation. At this stage, the level of complexity involved in executing the task should be identified. Complexity can guide the researcher to determining the approach to take for experimentation because he or she will have an idea of the tradeoffs for the particular problem, thus enabling an informed selection to control all variables and completely simplify the task, control some variables and partially reduce the complexity of the task environment, or retain realism. However, the sample available for data collection needs to be considered. Experts are often limited and expensive to attain for experimentation. Utilizing a novice population like students, though, necessitates careful planning so that the cognitive and physical requirements of the task tested match those that would be induced in experts. Retaining the realism requires a balance between reducing environment and task complexity, but only as much is necessary to

allow the novice participants the opportunity to attain a level of proficiency to complete the task (for an example see Reinerman-Jones, Guznov, Mercado, & D'Agostino, 2013). The motto is different, but equal; the experimental environment is simplified compared to the real-world tasking environment, but elicits the state and performance responses equivalent to the experts for the domain. That is accomplished by first acknowledging that novice participants will not acquire the breadth or depth of knowledge of the experts, but will be able to learn a basic understanding of the tasking environment and domain, and the skills to execute a specific task or piece of a procedure. Therefore, distilling the skills into core components is critical for quality data and results that enable concrete, ecologically sound conclusions.

Distilling Skills into Core Components

Distilling skills into core components requires task analysis and cognitive task analysis. A task analysis addresses actions and physical task execution, whereas a cognitive task analysis addresses mental and affective states. The task analysis can be completed by determining the steps required to complete a task in the real-world. This might be the number and order of mouse clicks, time elapsed between events, or systems and technologies used. It is beneficial to be in the actual environment for the initial task analysis. If pictures or video are permitted, then those are beneficial for revising and confirming the initial task analysis. Screenshots or a simulation replicating the task and the environment are also useful for conducting the task analysis. It is ideal to have an expert assist with the task analysis.

Similarly, a cognitive task analysis should be completed. Determining the types of information processing requirements needed to complete each step in the task analysis is important for matching cognitive experience when designing the experiment. The opportunity to ask experts about their thought process, scanning strategy, or task flow is priceless. However, some of this information can be derived from publically available documents like journals, logs, and accident reports. Research literature on similar tasks and domains are beneficial and theoretical constructs are the driving force behind this analysis.

Once the task analysis and cognitive task analysis are complete, the tasking environment can be simplified for experimentation and training can be developed.

Scaffolding

Scaffolding has been related to Lev Vygotsky's Zone of Proximal Development (ZPD) theory (Cazden, 1979). ZPD is considered the optimal realm of development between a learner's present level of understanding and his or her potential level given instructional support (Sanders & Welk, 2005; Vygotsky, 1978). Maintaining ZPD during instruction is often accomplished through scaffolding by providing dynamic instructional assistance and an appropriate level of challenge for the learner's abilities (Hirumi, Appelman, L., & Van Eck, 2010).

In instructor-learner settings, scaffolding involves adapting the level of instruction (i.e., increase or decrease) to support the learner's gradual acquisition of a skill (Puntambekar & Hubscher, 2005). The skill is decomposed into its critical sub-skills and the learner is guided to achieve his or her own understanding and mastery of each component (Carroll, Milham, & Champney, 2009; van de Pol, Volman, & Beishuizen, 2010). Instructional support is gradually withdrawn until the responsibility to conduct the skill is transferred entirely to the learner (van de Pol, Volman, & Beishuizen, 2010).

Scaffolding is also applicable for experimental task training in experimenter-participant interactions. As discussed previously, the experimental task should be distilled into its core components. Using a scaffolding model for task training, participants gradually become familiar with the core components until the experimental task may be executed proficiently. As the participant becomes more adept at the skills, the experimenter should taper the support and only offer prompts that guide the participant to the correct response or procedure. After task training is complete, the experimenter should not offer guidance during the actual experimental scenarios.

As core components are progressively combined and the complexity of the task increases, experimenters should check for understanding and provide opportunities for skill practice. Whether these checks are completed verbally or within an experimental testbed, the participant's responses should be logged in a handwritten or electronic format. These short evaluations provide valuable insight into individual differences in task understanding and level of task

mastery, which may explain variations in performance or outliers. One method to check for understanding is proficiency testing.

Proficiency Testing

Proficiency testing involves assessing a participant's ability to successfully execute the experimental tasks or subtasks. The purpose of proficiency testing is to check that the participant has a general understanding of how to execute the task and/or recall critical information from the task training. For complex experimental tasks, it may be appropriate to incrementally assess proficiency throughout task training as core components are combined and task complexity increases. For simpler experimental tasks, a single proficiency test upon completion of all task training may be sufficient.

It is highly recommended that in order to avoid priming effects, the task during the proficiency test should be similar to the experimental task, but not identical. For example, if participants must select scenario objects in a specified order for their response to be valid, then the proficiency test may focus on the selection procedure and does not require objects that mirror those in the experimental trials. Likewise, if an experimental task requires the participant to recognize images, then the proficiency test may simply present a series of the images for participant recall and does not require the exact presentation method of the experimental scenarios. Many experimental tasks involve a combination of procedural and cognitive skills in which case the proficiency test may involve a modified version of the experimental environment.

To ensure all participants in a single experiment are evaluated objectively, it is highly recommended that scoring methods provide a numeric value to indicate levels of proficiency. Objective proficiency measurements are particularly useful when more than one experimenter may administer proficiency tests during the entire period of data collection because there is little to no risk of bias in the interpretation of the result. An example of objective proficiency scoring is calculating the total number or percent of correct responses.

Proficiency testing provides a relative baseline of the participant's ability to implement a task. During data analysis, this baseline may explain individual differences in performance variables. Experimenters can also utilize proficiency results to exclude individuals who do not meet minimum skill execution requirements for the task. Depending on the task, experiences from previous experimentation indicate that a proficiency level of 75% accuracy or more is often sufficient to obtain a representative population. Expecting 100% accuracy is often unrealistic as even expert populations may make mistakes.

Interpreting Results

If the above recommendations are adhered, then interpretation of results and conclusions drawn should be optimistic in the ability of the findings to extend to the population and domain for which the experiment was designed, but careful to not overgeneralize. The context and original intent for which the data was collected are important to interpreting results in scope. Conclusions resulting from this approach will be the most valuable.

EMPIRICAL EXAMPLES

Two case studies illustrate the benefits employing these five strategies offer to human participant research efforts. The first use-case, focused on Soldier marksmanship training, demonstrates clear ties between Soldier and college student performance outcomes and perceptions. The statistical results and findings lend credibility to the strategies presented above. Next, specific strategy implementations from a nuclear power plant human factors experiment demonstrate how the research strategies are successfully implemented.

Use Case 1: Marksmanship Training

Previous experimentation comparing the performance of college students and active duty Soldiers demonstrates the ability to effectively apply the strategies presented. The comparison involved marksmanship tasks performed within the U.S. Army's Engagement Skills Training 2000 (EST 2000) (see). The EST 2000 is a projector-based virtual reality simulator that provides collective training in marksmanship and discriminatory firing. This experiment involved 144 Soldier participants ranging in age from 19 to 26 years with an average of 23 (SD=4.8). All Soldiers

had prior training with the M16 rifle and M240B machine gun. The 72 non-military participants, of varying backgrounds, ranged in age from 18 to 63 years with a mean of 24 (SD=8.2). None of the non-military participants had previous experience with the M16 or M240B weapons. Neither group had experience using a Remote Weapon System (RWS).

The study compared marksmanship performance outcomes between a standard four-person fire team configuration and an alternate configuration that replaced the gunner role player with a RWS. Participants were divided into four-person teams; each consisting of a one Gunner, and three riflemen (one rifleman was assigned role of team leader).

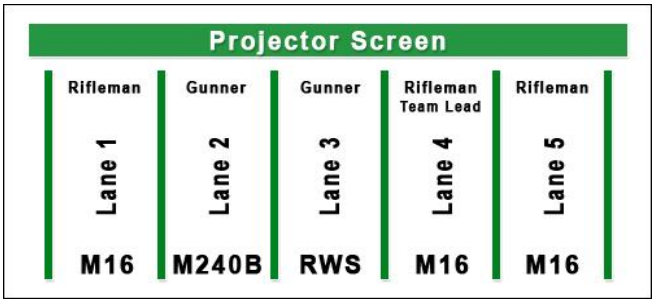

Figure 1. EST 2000 Experimental Configuration (Ortiz, et al., (2010))

In the standard condition, all participants were co-located in the EST 2000 simulator, and in the RWS condition, the RWS device was operated by the team's gunner from a different room. Both groups received training in the operation of the M16 and M240B weapons modified for the EST 2000. A training scenario provided an opportunity to practice marksmanship tasks within the EST 2000 prior to the execution of the experiment. Each of the fire teams, performed a predetermined set of tasks in the EST 2000 across simulation scenarios in the standard and RWS configurations.

Table 1. Soldier and Student Hit Percentage and Data Analysis (adapted from Ortiz, et al., (2012)

Soldier and College Student Hit Percentages									
		Soldiers		Students		Paired-T-Test		95% Confidence Interval	
Condition	Scenario	Mean	SD	Mean	SD	t	p-value	Lower	Upper
Standard Fire Team	Desert Setting	7.30	1.08	9.84	3.89	2.65	0.02	0.52	4.56
	Quarry Setting	7.97	1.47	10.57	3.07	3.50	0.00	1.03	4.16
Fire Team w/ RWS	Desert Setting	9.12	1.84	8.94	3.05	-2.13	0.83	-1.92	1.57
	Quarry Setting	10.88	3.05	9.09	3.41	-1.37	0.19	-4.54	0.97

The number of targets hit during two scenarios (i.e., desert setting and quarry setting) served as the primary performance metric. In the standard, fully manned, condition team performance was found to be significantly different based on paired sample t-tests. The confidence intervals reveal a very close proximity to zero, bringing into questions the practical significance of this finding. Furthermore, team performance showed no significant difference in the RWS condition. Results from a Cognitive Load Questionnaire (CLQ) reveal several similar mental load patterns between the Soldier and non-Soldier participants. The team leader scores and combined CLQ scores reported were significantly higher for both groups in the RWS condition during the desert scene scenario. For the gunner position, CLQ scores associated with the desert scene scenario were significantly higher for both groups in

the standard configuration. The similarities in performance and perception of Soldiers and non-Soldiers during this experiment provide confidence in the research strategies summarized above.

Use Case 2: Nuclear Power Plant Operations

These five recommendations were employed in an experiment for NPP operation. Research in the NPP domain has primarily taken an applied research approach in which expert, licensed NPP operators serve as participants on full physics-based system simulators. These experiments are largely proprietary and unpublished or written as non-distributable technical reports. O'Hara and colleagues (2010) were able to derive four categories or types of tasks from applied research: monitoring and detection, situational assessment, response planning, and response implementation. However, workload associated with each task type had not been investigated.

In an effort to understand workload associated with task type in the NPP domain, investigators gathered information from researchers who conducted studies in the NPP domain. They also talked with subject matter experts from the Nuclear Regulatory Commission (NRC) and former and active operators. The use of expert operators as participants was cost and time prohibitive, not to mention the sample size would be small as is the case with existing NPP research published in the literature. In order to complement the applied research completed and underway, the present effort sought to fill the more basic research piece, but balanced with realism.

To do this, investigators conducted a task analysis and a cognitive task analysis. Both analyses began at the NRC's Technical Training Center (TTC) in concert with three former NPP operators and trainers. These three SMEs provided a high level overview of a NPP and the main control room (MCR). They helped the investigators through common emergency operating procedures (EOPs) and investigators took notes about the types of controls and instruments used and frequency. Then, popularly used workload measures were administered to these operators as they served as an example crew executing EOPs. Following each EOP, investigators interviewed operators about their perceptions, strategies, decisions, and opinions on the measures administered. That experience enabled investigators to identify three overlapping and integrated stages in which task types occur: 1. Three-way communication, 2. Navigation, and 3. Task execution. The three stages and four task types were further scrutinized. Monitoring and detection was determined to really be two separate tasks and were thus renamed checking and detection (Reinerman-Jones, Guznov, Mercado, & D'Agostino, 2013). Reading the literature available on workload and tasks in other domains that are similar to the four task types guided investigators to design a baseline experiment to determine workload levels and types associated with checking, detection, and response implementation task types.

The next step was to identify an EOP that would be suitable for novice participants to accomplish, yet maintain the fidelity enough for ecological and external validity. Working with SMEs, the EOP that contained the most and the most equivalent steps for each task type was identified. This enabled a second level task analysis to occur on the EOP and the control panels associated with that EOP. That task analysis was organized in an Excel file for another former NPP operator to compare while performing the EOP in a computer-based generic pressurized water reactor (PWR) simulator, the same that would be used for experimentation. That simulator included the full physics and controls of a real PWR. The operator made notes next to items in the Excel workbook and provided screenshots for reference with inserted notes.

Investigators were empowered to begin experiment development upon completion of the task and cognitive task analyses. All steps in the EOP that were not associated with three task types of interest were eliminated. The steps were then organized into the task types, such that all checking steps were grouped, all detection steps were grouped, and all response implementation steps were grouped. The order of steps in each grouping were maintained according to the order that would be completed in the EOP, thus considering the physics of an NPP. To equate the number of steps in each task type, former NPP operators assisted in identifying steps that operators would do throughout most operating procedures and when those would occur in relation to the existing steps. Additionally, the experiment was designed to partially counterbalance the task types such that checking always preceded response implementation because an operator would never act on the controls before checking the state of the plant. Finally, complexity of the panels was reduced to account for the fact that while expert operators have extensive training approximately five times a year of the location of controls and are aided to the locations by a comprehensive knowledge of the NPP, training novice participants to perform to the same level as the experts with a limited knowledge-base and time was not feasible (Reinerman-Jones, Guznov, Mercado, & D'Agostino, 2013). Table 2 demonstrates this reduction in

complexity. The approach to designing the experimental tasks and modifications to the panels sought to retain ecological and external validity.

Table2. Example of one panel and its associated reduction in complexity

A2 Panel				
	Original Panel		Modified Panel	
Controls	Number of specific controls	Percent reduction needed	Calculated reduction of specific controls	Number of specific controls
		-43%		
Number of gauges	108		61.95	62
Number of switches	80		45.89	46
Number of light boxes	4		2.29	2
Number of status boxes	0		0	0
Other controls	5		2.87	3
Number of total controls	197		113	113

Investigators then turned to the procedure for the experiment. The three aforementioned stages were included to match the real-world environment as closely as possible. Participants were required to complete three-way communication, navigation, and task execution for each step of each task type. Therefore, scaffolding was used to train participants over two and a half hours. Training began with the basics of NPPs and moved to using 3-way communication to clearly relay critical information. Participants practiced the skill and completed a proficiency test, scoring 80% or better before training to navigate within the simulator to locate and read status indicators. Again, practice and a proficiency test were administered with an 80% pass. Participants then trained to respond appropriately to a simulated NPP system warning by following standardized procedures and completed with practice and a proficiency test. After achieving an 80% or better on each stage, a practice session combined all components. The training guide, PowerPoint slides, and proficiency tests were finalized from a review by a former NPP operator and SMEs.

Participants were joined by a confederate to complete the experimental session. The confederate previously received training and seven practice sessions on the entire experimental session. The confederate did not interact with the participant, but completed three-way communication with the experimenter who served as the supervisor reactor operator. Thereby, the realism of experiencing a crew was preserved. The confederates' data was collected for later comparison to absolute novices. Thereby, the experiment essentially consisted of an experienced participant sample and a novice participant sample. Data analyses are in progress, but preliminary findings are under review and look promising (Mercado, Reinerman-Jones, Barber, & Leis, under review). The results for the novice participants are showing limited differences in workload between task types. The results for the experienced participants are showing that WL differences were found among the different task types, but not sessions (Leis, Reinerman-Jones, Sollins, Barber, & Mercado, under review). The performance data has yet to be examined, but the preliminary workload findings indicate promise for using novice or experienced participants to investigate problems in a highly complex expert task.

DISCUSSION

Previous frameworks developed by Saddish, et al. (2002) and Fincannon, et al. (2012) aim to reduce external validity violations as a consequence of inappropriately extrapolating results from a sample population. For some real-world situations, access to experts is infeasible due to inherent constraints found in high-risk environments and/or the novelty of interfaces and devices under investigation. The strategies described target five critical threats to experimental validity and recommend best practices for performing human participant research. Ongoing research experimentation applying these methods continues to refine the process presented. Suggested extensions to this

framework include advancement of subjective measurement instruments (e.g., self-report surveys), inclusion of objective measures (e.g., physiological sensors), and evaluation of non-parametric analysis methods.

REFERENCES

Benner, P. (1984). *From novice to expert.* Menlo Park: American Journal of Nursing.

Blavier, A., Gaudissart, Q., Cadiere, G. B., & Nyssen, A. S. (2006). Impact of 2D and 3D vision on performance of novice subjects using da Vinci robotic system. *Acta Chirurgica Belgica, 106*(6).

Carnahan, T. J., Lickteig, C., Sanders, W. R., Durlach, P. J., & Lussier, J. W. (2004). Novice versus expert command groups: Preliminary findings and training implications for future combat systems. *(No. ARI-RR-1821) ARMY RESEARCH INST FOR THE BEHAVIORAL AND SOCIAL SCIENCES*.

Carroll, M., Milham, L., & Champney, R. (2009). Military observations: Perceptual skills training strategies. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)* (p. No. 9287). Arlington, VA: NTSA.

Cazden, C. (1979). Peekaboo as an instructional model: Discourse development at home and at school. Palo Alto, California: Stanford University Department of Linguistics.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology, 47*(1), 273-305.

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review, 100*(3), 363.

Fincannon, T., Keebler, J. R., & Jentsch, F. (2012). Examining external validity issues in research with human operation of unmanned vehicles. *Theoretical Issue in Ergonomics Science*, 1-20.

Hao, L., & Houser, D. (2010). Getting it right the first time: Belief elicitation with novice participants.

Hirumi, A., Appelman, B., L., R., & Van Eck, R. (2010). Preparing instructional designers for game-based learning: Part 1. *Tech Trend, 10*(4), 27-37.

Kitzinger, J. (1995). Qualitative research. Introduction focus groups. *BMJ: British medical journal, 311*(7000), 299.

Leis, R., Reinerman-Jones, L., Sollins, B., Barber, D., & Mercado, J. (under review). Nuclear power plant task workload across repeated sessions. Proceedings for the annual conference of the Human Factors and Ergonomics Society (HFES). Chicago, IL.

Mercado, J., Reinerman-Jones, L., Barber, D., & Leis,R. (under review). Investigating workload measures in the nuclear domain. Proceedings for the annual conference of the Human Factors and Ergonomics Society (HFES). Chicago, IL.

O'Hara, J. M., & Higgins, J. C. (2010). Human-System interfaces to automatic systems: Review guidance and technical bases. Human Factors of advanced reactors (NRC JCN Y-6529) BNL Tech Report No BNL91017-2010.

Ortiz, E. C., Salcedo, J. N., Lackey, S. J., Fiorella, L., & Hudson, I. L. (2012). Soldier vs. non-military novice performance patterns in remote weapon system research. *In Proceedings of the 2012 Symposium on Military Modeling and Simulation* (p. 5). Society for Computer Simulation International.

Ortiz, E., Lackey, S. J., Stevens, M. A., & Hudson, I. (2010). The impact of unmanned weapon system on individual and team performance. *In Proceedings of the 2010 Spring Simulation Multiconference* (p. 20). Society for Computer Simulation International.

Parikh, S. E. (2011). Characterizing expert and novice differences in problem solving in heat transfer. Stanford University.

Puntambekar, S., & Hubscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist, 40*(1), 1-12.

Reinerman-Jones, L., Guznov, S., Mercado, J., & D'Agostino, A. (2013). Developing

Methodology for Experimentation Using a Nuclear Power Plant Simulator. In Foundations of Augmented Cognition (pp. 181-188). Springer Berlin Heidelberg

Sadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized casual inference.* Wadsworth Cengage Learning.

Salas, E., Milham, L. M., & Bowers, C. A. (2003). Training evaluation in the military: Misconceptions, opportunities, and challenges. *Military Psychology, 15*(1), 3.

Sanders, D., & Welk, D. S. (2005). Strategies to scaffold student learning: Applying Vygotsky's zone of proximal development. *Nurse Educator, 30*(5), 203-207.

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*.

Vygotsky, L. (1978). *Mind and Society: The Development of Higher Mental Processes.* Cambridge: Cambridge University Press.

Ward, P., & Williams, A. M. (2003). Perceptual and cognitive skill development in soccer: The multidimensional nature of expert performance. *Journal of sport & exercise psychology*, *25*(1).