Identification and Evaluation of Simulator System Deficiencies

Justin H. Prost
Lumir Research Institute
Tempe, AZ
Justin.Prost@lumirresearch.com

Brian T. Schreiber
Lumir Research Institute
Tempe, AZ
Brian.Schreiber@lumirresearch.com

Winston Bennett, Jr.
Air Force Research Laboratory
Mesa, AZ
Winston.Bennett@mesa.afmc.af.mil

ABSTRACT

The importance of fidelity in simulation training has been established as a key factor in training effectiveness. Design and development of simulation systems is dealt with commonly as an iterative procedure in which the goal is to improve on the training effectiveness of the system. Improvements are made based on methods of identifying and resolving deficiencies. Improvements must then be considered relative to available resources to determine how to prioritize the multiple options. The current paper presents a flexible methodology for evaluating the relative effects of multiple deficiencies on a simulation system, utilizing information from warfighters about the effectiveness of a Deployable Tactical Trainer (DTT). First, subject matter experts were asked to evaluate the fidelity of the current DTT system across a set of 199 experiences and emergency procedures (EP). Next, the same subject matter experts identified deficiencies of the current system. Then the warfighters evaluated the 199 experiences and EPs for which deficiency had the most adverse affect on training. A composite score was computed for each deficiency, using weighted variables accounting for 1) the amount of improvement possible and 2) the importance of training for each experience. The deficiency scores provide a means of comparing the relative effect of each deficiency on warfighter training. The impact of utilizing multiple sources of information is presented through a comparison of the different decisions that would result from using partial information or alternative weighting methods. The proximal implications of using the proposed methodology to have the greatest impact on improvement are discussed. The distal implications of the impact the improvements have on pilot perceptions, and ultimately, on objective pilot performance measures, are also discussed. Also discussed is the versatility of the methodology for incorporating information from various sources and weighting systems to include alternative decision-making factors.

ABOUT THE AUTHORS

Justin H. Prost is a Research Scientist with Lumir Research Institute. He completed his Ph.D. in Developmental Psychology at Arizona State University in 2001. Recently, he has worked on current simulation research at the Air Force Research Laboratory.

Brian T. Schreiber is CEO and Senior Scientist with Lumir Research Institute in support of the Air Force Research Laboratory, Warfighter Readiness Research Division, in Mesa, AZ. He completed his M.S. in Human Factors Engineering at the University of Illinois at Champaign-Urbana in 1995.

Winston Bennett, Jr. is a Senior Research Psychologist and team leader for the training systems technology and performance assessment at the Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division, in, Mesa AZ. He received his Ph.D. in Industrial/Organizational Psychology from Texas A&M University in 1995.

Identification and Evaluation of Simulator System Deficiencies

Justin H. Prost
Lumir Research Institute
Tempe, AZ
Justin.Prost@lumirresearch.com

Brian T. Schreiber
Lumir Research Institute
Tempe, AZ
Brian.Schreiber@lumirresearch.com

Winston Bennett, Jr.
Air Force Research Laboratory
Mesa, AZ
Winston.Bennett@mesa.afmc.af.mil

INTRODUCTION

Simulators have been found to be an effective training alternative to live-fly training (Bell & Waag 1998; Gehr, Schreiber, & Bennett, 2004; Schreiber & Bennett, 2006). The constraints and costs of training in aircraft makes simulator training an attractive alternative. The increasing sophistication of the simulators allows for training of a wider variety of skills, for example the ability to engage in distributed missions training due to the increased capabilities of linking simulator systems. As technology and simulation capabilities continually change, one of the keys to continually improving the training capabilities of the simulator systems is the *process* of improving the systems. Generally, improvements are always made to our simulation systems, but not executed according to a scientific decision making process.

In order to improve a simulator system, one must first look at the goals for the simulator system being developed. In the current study, the Deployable Tactical Trainer (DTT) under development at the Air Force Research Laboratory (AFRL) is under consideration. The DTT is a simulator system that is being developed with the goal of providing the best F-16 simulator training possible, while remaining capable of transportability and efficient maintainability. As many piloting skills do decay, there is great utility in having a deployable trainer available to provide skills maintenance and mission rehearsal at remote locations. In an effort to accomplish this goal, AFRL is working to maximize the fidelity and functionality of the DTT system, while still remaining deployable and cost effective. The identification and evaluation of system deficiencies most critical to providing quality training and mission rehearsal is essential to this process. This allows a prioritization of addressing deficiencies based upon the scientific, documented degree of impact each deficiency has on training utility.

Previous research has examined the fidelity of the DTT system and compared the fidelity of the deployable simulation system to a "high-fidelity" simulation system (Display for Advanced Research and Technology, DART) (Schreiber, Gehr, & Bennett, 2006) The previous work used a set of Mission Essential Competencies (MECs) and Emergency Procedures (EPs) to evaluate the ability of each simulation system to train F-16 skills. The findings from the study showed that the "high-fidelity" system provided better training than the "low-fidelity" DTT system. The research on the fidelity of the system is one part of the development process of the DTT system. The current paper presents the next stage of the development process for the DTT system, the identification and evaluation of the deficiencies in the system.

Current Work

The current study was developed to generate a methodology for quantifying the identification and evaluation of deficiencies. Furthermore, the methodology was designed to incorporate information from multiple sources in an effort to have a more comprehensive mathematical model for differentiating the relative impact of deficiencies. The identification and evaluation of deficiencies in simulators is a critical step in the process of improving the system. Subjective determinations—without quantifiable impact on utility—are typically the methods employed for deciding which simulator deficiencies to address or improve upon. The current work aims to provide a scientific, quantifiable method where previously there was none. The goal was to define a process that yields a priority order of deficiencies that need to be addressed, in terms of their greatest impact on training utility. The methodology presented in the current study is also designed to be flexible, allowing models to be constructed that incorporate specific information, as

well as multiple models based on different criteria to allow for comparison.

The current work is focused on the identification and evaluation of the deficiencies present in the most current version of the DTT. This identification and evaluation involved a multi-method approach including: focus groups, open ended surveys, fidelity ratings on the current system, deficiency evaluation surveys, and previously collected evaluations in training gaps. The current model will provide scores for each deficiency across all 199 identified MEC experiences and emergency procedures. The sum of the scores for each deficiency across the 199 items will provide an overall value representing the impact of addressing each deficiency relative to each other.

METHOD

Evaluators

Seven F-16 subject matter experts (SMEs) independently completed an evaluation of the fidelity of the system for 199 items. Additionally, the same SMEs provided feedback regarding the deficiencies of the most current DTT simulation system. Evaluators participating in the current project have all previously been exposed to the survey process of evaluating the DTT system for fidelity. All evaluators were given detailed documentation of the system components, capabilities, and limitations, and were provided multiple opportunities to fly various missions in the system before ratings on the system were given.

The seven DTT evaluators were all male, all recently retired (3 years on average) F-16 pilots from the United States Air Force (USAF), and had an average number of operational F-16 flight hours of 2,119 (average total flight hours of 3,495).

Surveys

There were two surveys administered: a fidelity survey and deficiency survey. Both surveys were administered to evaluate the current DTT system.

Each of the surveys used consisted of 199 total items to be rated: 72 air-to-ground (A/G) MEC experiences, 55 suppression of enemy air defense (SEAD) MEC experiences, 45 air-to-air (A/A) MEC experiences, and the 27 bold face EPs for the F-16.

The fidelity survey was administered in a consistent fashion with previous survey administration of the fidelity of the system. The scores obtained from the fidelity survey will be transformed for the final model presented in this paper. The transformation will be to subtract the average for each skill and divide by 5, to produce a proportion of the improvement potential for each skill. For the fidelity survey, each item was evaluated using the following rating scale:

0= N/A Capability to experience does not exist
1= Capability to experience exists, but is very poor
2= Capability to experience exists, but is poor
3= Capability to experience exists, but is marginal
4= Capability to experience exists, and is good
5= Capability to experience exists, and is very good

The deficiency survey was developed through a process of open-ended questions, focus, groups, and the ability to write-in additional deficiencies during the final survey. The process of identification of the deficiencies yielded 17 specific deficiencies and three additional options. The additional options included an option for an unspecified software issue, an option for an unidentified hardware issue, and an option for skills for which there was no deficiency that could be specified.

For the deficiency survey, each skill was evaluated based on which of the following were the primary, secondary, and tertiary deficiencies:

1= No EP instructor/operator station
2= No laser guided bombs (LGB) self designation capability
3= No SEAD (Block 50) capability
4= Limited out-the-window (OTW) capability
5= Realism and ease/difficulty in UFC functions on touch screen
6= No chaff elbow switch
7= No helmet mounted sight capability
8= Realism and ease/difficulty in OSB functions on touch screen
9= Location, size, or clarity of instruments on touch screen
10= Multifunctional display (MFD) replication on touch screen
11= Cursor control, slew rates, and sensitivity to slews for the radar are not the same as aircraft
12= Location of touch screen relative to pilot seat and sitting position
13= Requirement to select individual side panels on the touch screen
14= No night vision goggle (NVG) capability

15= No electronic counter measure (ECM) capability
16= No capability to experience g-force
17= No lighting
18= Other software related issue (e.g., not in the database, lack of weapon system capability, etc.)
19= Other hardware related issue (e.g., no ejection handles, etc.)
20= Not specified, system simply not capable of providing the experience as configured.

Apparatus: Current DTT system

The DTT system evaluated consisted of the following major hardware/software functionalities (Figure 1):

Hardware: F-16 cockpit shell with three out-the-window 30-inch displays, the actual F-16 stick/throttle, and simulation of all cockpit displays and switch functions on a high resolution 23" interactive touch screen display. Image generator was an SDS International AAcuity® PC-IG system. Brief/debrief includes SmartBoard and two 50-inch displays for Head-Up Display, Radar Warning Receiver, and Multi-Function Display.

Software: The system uses classified Block 30 (SCU 5p) actual operational F-16 software. One database is currently installed and available. Debrief software has the ability to link and time-synchronize video recordings from multiple players. It also has the ability to network through Distributed Interactive Simulation (DIS) or High Level Architecture (HLA) standards. It has chaff/flare capability, but no ECM. Some classified weapons systems available, such as Joint Direct Attack Munition (JDAM).

Figure 1. Two DTT cockpits (brief/debrief hardware not shown).

Procedure

For the current study, the approach required gathering information through the use of two independent surveys. Additionally, previous work has been done regarding the development and evaluation of the MECs during the development of the competency skill sets. The Summary Reports produced during the development of the MECs include an evaluation of the skills for which there was a training gap present. The training gap analysis is used to provide the training gap data used in the study. The three sources of information (deficiency evaluation, fidelity evaluation, and training gap evaluation) about the current DTT system were analyzed to identify the relative importance of deficiencies in the system.

The analyses of the deficiencies are based on the proportion of the total value that is accounted for by each deficiency. The proportions are examined across all 199 MEC experiences and EPs and for the four sub-areas (Air-to-Ground experiences, Air-to-Air experiences, SEAD experiences, and EPs) separately. Only the top ten deficiencies for each set of skills are presented for brevity. The proportions presented are based on the following three models.

Frequency Model: In model one, the proportions presented represent the total frequency of responses for a deficiency, across all items, divided by the total number of responses.

Improvement Model: In model two, the proportions presented are the sum of the weighted scores for each deficiency divided by the total score across all deficiencies and items. The weighted scores in model two are computed by multiplying the frequency of a deficiency response by the proportion for improvement potential.

Improvement and Training Gap Model: In model three, the proportions presented are the sum of the weighted scores for each deficiency divided by the total score across all deficiencies and items. These weighted scores are computed by multiplying the frequency of a deficiency response by the proportion for improvement potential; and, then multiplying that product by a weight identifying the level of training gap present for each skill.

There were three factors considered in the models presented. The factors were defined as follows:

Deficiency Evaluation was based on the frequency of response for each deficiency by the evaluators for each skill.

Fidelity Evaluation was based on the proportion of improvement possible for each skill, calculated as five minus the average fidelity rating for the skill divided by five.

Training Gap Evaluation was based on previous work identifying the training gaps for the skills. The weightings were 1 for no gap, 2 for potential gap, and 3 for gap.

RESULTS

The analyses of the deficiencies are presented for all MECs and EPs and for each of the four skill areas independently. The results for each set of items includes a table identifying the top ten deficiencies in each area, based on the frequency of identification and summary statistics of the factors involved in the models.

All Items:
The analysis across all items found that the average fidelity rating was 1.25, reflecting the ability of the DTT to currently provide a training experience that is between poor and marginal across all items on average. The average proportion for improvement across all items was .75. The frequency of training gaps was 123 skills with no training gap, 23 with a potential training gap, and 53 with a training gap. The order of importance of the top five deficiencies is not consistent across the three models. As can be seen in the table with model 3 as compared to model 1, the SEAD capability and Emergency Procedures Instructor trade position as the second and third greatest deficiency having an impact.

Table 1. Proportion of total score attributable to each of the top five endorsed deficiencies across all items.

	Model 1	Model 2	Model 3
4	0.33	0.25	0.26
1	0.19	0.24	0.16
3	0.15	0.16	0.19
18	0.10	0.12	0.13
7	0.05	0.04	0.04

Air-to-Air Items:
The analysis for the air-to-air items found that the average fidelity rating for the items was 2.08, reflecting the ability of the DTT to currently provide a training experience that approximately "marginal" across air-to-air items on average. The average proportion for improvement across air-to-air items was .58. The frequency of training gaps was 32 skills with no training gap, 5 with a potential training gap, and 8 with a training gap. As was seen with the analysis for all items, the air-to-air items were impacted by deficiencies differently based on the model considered. It is clear for the air-to-air skills that the limited out the window capabilities have the greatest impact; however, the lack of ECM capabilities becomes the second most important deficiency when we consider the model that includes more factors.

Table 2. Proportion of total score attributable to each of the top five endorsed deficiencies across air-to-air items.

	Model 1	Model 2	Model 3
4	0.54	0.47	0.47
7	0.10	0.09	0.09
1	0.09	0.13	0.09
18	0.06	0.08	0.07
15	0.05	0.08	0.15

Air-to-Ground Items:
The analysis for the air-to-ground items found that the average fidelity rating for the items was 1.36, reflecting the ability of the DTT to currently provide a training experience that approximately "marginal" across air-to-ground items on average. The average proportion for improvement across air-to-ground items was .73. The frequency of training gaps was 38 skills with no training gap, 9 with a potential training gap, and 25 with a training gap. As seen with the previous two sets of items, the rank order of the deficiencies flips depending on the model being considered. Of additional interest is the finding that a separation of the top two deficiencies for model 1 of 21 percent, drops to a separation of only 10% in models two and three. The top deficiency for the air-to-ground items is the out-the-window visual capabilities. The second deficiency pertains to software issues, primarily identified as a database deficiency. Implications of this finding will be discussed further in the discussion paper.

Table 3. Proportion of total score attributable to each of the top five endorsed deficiencies across air-to-ground items.

	Model 1	Model 2	Model 3
4	0.43	0.37	0.37
18	0.22	0.27	0.27
7	0.08	0.06	0.07
20	0.07	0.08	0.04
1	0.06	0.06	0.06

SEAD Items:

The analysis for the SEAD items found that the average fidelity rating for the skills was 0.96, reflecting the ability to of the DTT to currently provide a training experience that approximately "marginal" across SEAD skills on average. The average proportion for improvement across SEAD skills was .81. The frequency of training gaps was 26 skills with no training gap, 9 with a potential training gap, and 20 with a training gap. A switch in the order of impact of the top five deficiencies can be seen when comparing the different models. The overwhelming issue with the SEAD skills is the lack of Block 50 capabilities, while the other four deficiencies in the top 5 are much closer together as far as impact, particularly when the weighting is considered for amount of improvement and training gaps. Implications of these findings are discussed below.

Table 4. Proportion of total score attributable to each of the top five endorsed deficiencies across SEAD items.

	Model 1	Model 2	Model 3
3	0.52	0.55	0.57
4	0.17	0.14	0.13
5	0.08	0.08	0.08
1	0.05	0.06	0.05
14	0.04	0.05	0.07

Emergency Procedure Items:

The analysis for the emergency procedure items found that the average fidelity rating for the skills was 0.14, reflecting the ability to of the DTT to currently provide a training experience that was approximately "marginal" across air-to-air skills on average. The average proportion for improvement across air-to-air skills was .97. There was no training gap information available for the emergency procedure items, therefore all skills were given a 1, indicating no training gap. The results were clear for the emergency procedure items. There was one deficiency identified for all EP items, unanimously across all pilots. The deficiency was the lack of an EP instructor (deficiency number 1).

Table 5. For Emergency Procedure skills, all pilots indicated deficiency number 1 for all skills.

	Model 1	Model 2	Model 3
1	1	1	1

DISCUSSION

The results of the current study provide evidence of the utility of a quantitative model for the process of evaluating the deficiencies in a simulation system. The models provided a means of quantifying decision-making that occurs during the improvement of the simulation system. The findings from the comparative analyses provide very clear evidence for the importance of including multiple sources of information in the models evaluating deficiencies. The order of importance of the deficiencies was found to be inconsistent across the models. This result provides the strongest evidence that the models reveal different information about the relative importance of deficiencies of the system.

Two important factors to consider in interpreting the findings of the current study are the current developmental stage of the system being evaluated and the limited number of decision-making factors included in the models presented. The current system is in the early stages of development and has been designed with some serious constraints to the available technologies. The deployability of the system has presented constraints on using technologies that would increase the fidelity of the system. As an example, the constraints have lead to the most pervasive deficiency identified for the system, the out-the-window capabilities. Previous research illustrates this point.

Schreiber, Gehr, and Bennett (2006) provided evidence that the DART system provides a higher-fidelity training environment than the DTT. This previous finding combined with the current study and knowledge of the specifications of the two systems makes it clear that the deficiency in the DTT system accounting for most of the difference in the two systems is the OTW capabilities of the DART system. In the current study, the OTW capabilities accounted for from one third to over one half of the deficiency of the DTT system in the three areas other than emergency procedures. The need for the DTT system to remain deployable (i.e., transportable)

constrains the system from using the full field visual displays utilized by the DART system.

The importance of keeping the relatively new stage of development of the DTT system in mind is that the results show that there are one or two very pervasive deficiencies with the system. Future analyses of the deficiencies will provide a more fine-grain investigation into the deficiencies that exist. For example, in the results for the air-to-air items, if we look at the proportion of accountability for the 3rd, 4th, and 5th deficiencies, some very interesting results occur. Looking at the results of model one, the proportions are .09, .06, and .05. The results for model three are .09, .07, and .15. The difference in the two models is very significant at this level of differentiation of the deficiencies. The first model and the third model have a difference in magnitude of importance of a multiple of three. A deficiency that only accounts for 5% in model one, accounts for 15% in model three, moving from being at the last position in a list to the top position.

The second critical factor to keep in mind when interpreting the results of the current study is the use of only two decision-making factors. The models presented in the current project are not comprehensive of all factors that should be considered, but represent models incorporating some of the most fundamental issues impacting decision-making. First, what are the current deficiencies of the system and how do they impact the different skills being trained on the system. Second, how much improvement would a solution to a deficiency affect for each skill. Third, how important is improving the skill for the current system, i.e., is there a training gap that needs to be addressed by the simulation system. These are some very fundamental issues for considering the relative importance of the deficiencies to the system. In future work, there is the intention of including additional factors; such as, the costs involved in providing solutions to the various deficiencies and comparisons of the importance of the capabilities of the four different skill areas. Additionally, the next stage of work will consider different weighting systems for the factors included in the models, to reflect different degrees of importance for multiple levels of a factor.

Overall, the methodology presented in the current paper should be viewed as a significant first step in creating quantifiable models to assist in the decision-making process during the improvement of simulation systems. The complexity of simulation systems, when combined with the diverse functions the systems are being asked to perform, provide a difficult task for the teams developing the systems. Utilizing a method that provides for quantification of the impact of deficiencies to a simulation system will allow the improvement of the system to meet the goals of development more efficiently. As the sophistication of the model presented in the current paper grows, so too can the utility of the methodology.

ACKNOWLEDGEMENTS

The authors wish to thank the USAF Air Combat Command for sponsoring this research and development work. This research was conducted under USAF Contract No. F41624-97D-5000. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

Allen, J., Buffardi, L., & Hays, R. (1991). *The Relationship of Simulator Fidelity to Task and Performance Variables* (ADA238941). Fairfax, VA: George Mason University..

Bell, H. H. & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology.* 8 (3), 223-242.

Colegrove, C. M. & Alliger, G. M. (2002, April). *Mission Essential Competencies: Defining Combat Mission Readiness in a Novel Way.* Paper presented at the NATO RTO Studies, Analysis and Simulation Panel (SAS) Symposium. Brussels, Belgium.

Dennis, K. A. & Harris, D. (1998). Computer-based simulation as an adjunct to ab initio flight training. *International Journal of Aviation Psychology,* 8, 261-276.

Gehr, S. E., Schreiber, B. T., & Bennett, W. Jr. (2004) Within-Simulator Training Effectiveness Evaluation. In *2004 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) Proceedings*. Orlando, FL: National Security Industrial Association.

Glass, G. V. & Hopkins, K. D. (1995). *Statistical Methods in Education and Psychology, 3rd Ed*. Needham Heights, MA: Allyn and Bacon.

Schank, J. F., Thie, H. J, Graf, C. M., Beel, J., & Sollinger, J. (2002). *Finding the right balance: Simulator and Live Training for Navy Units*. Santa Monica: RAND Corporation.

Schreiber, B. T. & Bennett, W. Jr. (2006). D*istributed Mission Operations within-simulator training effectiveness baseline study: Summary report.* (AFRL-HE-AZ-TR-2006-0015-Vol I). Mesa AZ: Air Force Research Laboratory, Warfighter Readiness Research Division.

Schreiber, B. T., Gehr, S. E., & Bennett, W. Jr. (2006). Fidelity Trade-Offs for Deployable Training and Rehearsal. In 2006 *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) Proceedings.* Orlando, FL: National Security Industrial Association.

Schreiber, B. T., Rowe, L. J., & Bennett, W. Jr. (2006). *Distributed Mission Operations within-simulator training effectiveness baseline study: Participant utility and effectiveness opinions and ratings.* (AFRL-HE-AZ-TR-2006-0015-Vol IV). Mesa AZ: Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division.