2005

Program Update and Prospects for In-Flight Simulation Upset Recovery Training

Janeen A. Kochan

James E. Priest

PROGRAM UPDATE AND PROSPECTS FOR IN-FLIGHT SIMULATION UPSET RECOVERY TRAINING

Janeen A. Kochan
James E. Priest
General Dynamics
Roswell, NM

The Flight Research Training Center, established in 2002 in cooperation with the Federal Aviation Administration, focuses on improving the safety of commercial air transportation through the reduction of the loss-of-control events, which continue to be the leading cause of fatal commercial air carrier accidents (Boeing Commercial Airplanes Group, 2004). The primary research purpose of this program is the optimization of in-flight simulation based upset recovery training. The goal of the training is to have a beneficial impact on the loss-of-control accident and incident rate. The program is designed to collect research data through an extensive training program offered to commercial airline pilots. To date, more than 235 commercial pilots have completed the integrated two-day program which includes classroom, aerobatic aircraft, and advanced in-flight simulation aircraft training on how to best respond to a variety of upset situations. This paper presents the results of the data collection and analysis effort for the FAA-Upset Recovery Training (URT) program for the twenty-four month period from August 8, 2002 through July 30, 2004.

Introduction

Program Background

The fundamental goals of the FAA-URT project are (a) to conduct research to optimize in-flight simulation (IFS) based upset recovery training, (b) meet the pilot training needs of commercial air carriers, (c) to design and develop IFS technology and systems specifically for the URT role, and (d) to have a beneficial impact on the loss-of-control accident, incident, and event rate. This results in a program which is hybrid in nature creating two complementary, yet independent, activities: operational training and an empirical study. Quasi-experimental field studies inherently present a host of research obstacles (Shadish, Cook, & Campbell, 2002) which are amplified in this blended training and research arena. A balance had to be reached between the need for efficient training operations and the need to collect sufficient data to provide a basis for the optimization activity.

Original Training Protocol

The structure and deployment of the training protocol has been continually monitored and revised during the course of the study based on participant feedback. For the first eighteen months of the study (August 8, 2002 through December 31, 2003) the training protocol was composed of three modules and conducted over a two day period. The first module was a classroom lecture where participants received instruction in causes of upsets, aerodynamic fundamentals, and recovery techniques. The second module was usually a training flight in the Aerobatic Bonanza. The Bonanza flight exposed the pilot to general aircraft characteristics, G-force awareness, slow flight and stall awareness, limited aerobatics, and unusual attitude recoveries. The third module was the IFS Learjet aircraft training where the participant experienced real-world upset events and practiced various recovery techniques. This module began with a flight rehearsal session using a ground-based (non-motion) simulator

Protocol Modifications

Changes in the program protocol based on feedback from the first 201 participants of the program were made beginning in January, 2004. Evaluation forms from the participants and instructor comments regarding the structure of the program and the usefulness of each of the training elements were reviewed. Based on the high frequency of comments regarding the order of events in the training, the structure of the program was changed. This involved separating the classroom briefings into two sessions; one prior to the Bonanza flight and one prior to the ground simulator and Learjet flight. Strict adherence to flying the Bonanza prior to the Learjet was also implemented in response to comments from the participants. These modifications to the protocol changed the order of the presentation of elements; however, the content of each module was not substantially altered.

In-Flight Simulator Learjet Flight

The URT protocol is an integrated, multi-part training event. However, the majority of the measures during the first 24-months of the program have focused on the efficacy of the In-Flight Simulator

Learjet training flight. A typical flight consisted of five phases: (a) familiarization exercises, (b) beginning evaluation exercises, (c) “g” awareness and confidence maneuvers, (d) upset recovery practice events, (e) ending evaluation exercises. The Learjet IFS aircraft, pre-programmed with upset events, is used to teach actual upset recoveries. The events programmed into the simulation system range from atmospheric effects and a wake turbulence encounter to extreme control failures and control surface hardovers. The simulation was of a light-to-medium size transport aircraft that is near max gross weight so that the inertias produce near worst case handling qualities.

Quantifying the Training Effectiveness

To optimize the in-flight simulation based upset recovery training, we needed to be able to measure how much the participant’s ability to recover, from a variety of upsets, improved during the training. We also needed to assess the value of the various events to the participant. Our initial research questions were (a) how much did the participant learn from the URT experience, and (b) what elements did the participant find most useful and why?

Recovery Ratings

Measuring a pilot’s ability to recover is a difficult task. Unfortunately, the seemingly straightforward concept of measuring performance parameters such as reaction times, maximum bank or pitch angles, etc. do not provide an accurate measure of a pilot’s ability to recovery from any given unexpected event. The essence of the difficulty in assessing human performance in this task is captured in the following example.

Consider that a single driver has two cars; car “A” steers poorly, car “B” steers like a dream. If you follow each of these cars for 10 miles, with the same driver, you may not be able to tell which car drives the best. When car “A” is driven, the driver pays strict attention to the task and rarely strays from the center of the lane. When car “B” is driven, the driver’s attention may wander to other things resulting in straying further from the center of the lane than occurred with car “A”. To the outside observer, using quantitative measures, it might appear that car “A” handles better. However, the most expeditious (and perhaps accurate) way of finding out which car drove the best is to ask the driver who will be able to tell you unequivocally about the (a) mental and physical workload, (b) level of apprehension and or stress, and (c) confidence experienced in performing the task. In this example, driver opinion would say car “B” performed better. Thus, in the long run, it may be much more cost effective and accurate to ask the driver to provide the performance evaluation.

Measuring the quality of a pilot’s recoveries to upset events presents a problem similar to the driving task. We must consider both the perceptual-motor performance, physical and mental workload, and the level of confidence one has in responding to an upset. Flight test organizations around the world have adopted the Cooper-Harper rating scale to facilitate quantifying aircraft handling qualities (Gawron, 2000). The Cooper-Harper scale incorporates performance and workload measures to assist an evaluation pilot in determining a single rating of the handling qualities of a particular aircraft.

The original Cooper-Harper Handling Quality Scale has been adapted to fit the needs of the URT program. The Recovery Rating Scale (RRS) is administered near the beginning of the flight to obtain a “beginning” rating and then again after practice near the end of the flight for an “ending” rating. These “beginning” and “ending” ratings did not intend to measure what had been learned from the entire course. Instead, the purpose of the scale was to help determine how much the participant learned specifically from the in-flight simulation upset recovery practice.

During this initial phase of the program, no effort was made to measure the amount learned in the entire course, or how much each element contributed to the overall program. However, participants did have an opportunity to comment on their perceptions of the elements and the benefit of the overall course as part of the post flight evaluation form.

Participant’s Evaluations of the Program

The second question of interest, how valuable was the course to the participant, was addressed by a postflight evaluation form. This form contained specific, liker-scaled and open-ended questions regarding the participant’s perception of each element of the training protocol.

Research Questions

Our initial research questions were (a) how much did the participant learn from the URT experience, and (b) what elements did the participant find most useful and why? To address these questions, we posed the following specific questions to guide our initial analysis of the data:

- Is there any significant improvement in the Ending RRS scores over the Beginning RRS scores?
- What is the relationship between total flight time and the Beginning RRS scores?
- What is the relationship between total flight time and the Ending RRS scores?
- What other factors influence the Beginning RRS scores, Ending RRS scores, or the magnitude of the difference between them?
    - What effect does military training have on the RRS scores?
    - What effect does previous aerobatic experience have on the scores?
    - What effect does being an instructor pilot have on the scores?
- What are the participants' perceptions of the URT program?

Method

Participants

The participants to date were 248 volunteers recruited by direct contact to airline training departments, website solicitation, and word of mouth. Participants were also informed of the study through numerous articles written about the project and published in aviation journals and trade magazines. Program contact information for participants was often included in the reports and articles.

Data analysis for this report was completed using data sets from 185 qualified air carrier pilots representing 27 different U.S. Part 121 air carriers. The additional participants (not included in these analyses) were from government organizations (e.g., FAA and NTSB), universities, research facilities, and private organizations (e.g. Airline Pilots Association, National Business Aircraft Association, etc). The exclusion of these data facilitated a focused look at the representative air carrier pilot.

Study participants included three females and 182 males. Approximately one-third (68) of the participants had military training and 121 participants had experience as either military or civilian instructor pilots. All participants held at least an FAA Commercial pilot certificate with an Instrument Rating, although the majority (157) held an Airline Transport Pilot certificate. All participants maintained a current FAA Medical Certificate.

Data Collection

Data was collected by the program administrator and Safety Pilots through forms, questionnaires, and instructor notes. All materials containing study materials and data were kept in secure quarters, accessible only to the study principals and researchers.

Results

Data Screening

All data in this study were next reviewed for accuracy of input into the SPSS file by checking for (a) out-of-range values, (b) plausible means and standard deviations, (c) univariate outliers, and (d) missing data. Pairwise plots for nonlinearity and heteroscedasticity were also reviewed when necessary for the statistical method used. When data was found to be missing, it was random in nature and therefore posed little threat to the validity of the results. Unless otherwise specified, all analyses were conducted with alpha level set at $p < .05$. Analyses were conducted using the Statistical Package for the Social Sciences v. 11.5. All graphic scales depicting data analysis are scaled identically for ease of comparison.

Inter-rater Reliability

The study was conducted and data were collected by six different instructor pilots (safety pilots). Analyses were conducted and no significant differences were found between Beginning or Ending RRS scores from participants of different instructor pilots $F(5, 144) = 2.175$, $p = .060$.

Upset Recovery Rating Score Differences

A paired-samples t-test was conducted to evaluate the impact of the in-flight simulator Learjet training on responses on the RRS scores. There was a statistically significant decrease (improvement in perceived performance) from the beginning rating scores ($m = 6.29$, $SD = 2.06$) to the ending rating ($m = 2.87$, $SD = 1.13$), $t(149) = 24.13$, $p < .0005$. The eta squared statistic (.80) indicated a very large effect size.

Effects of Total Flight Hours on Upset Recovery Rating Scores

A repeated-measures univariate analysis of variance (ANOVA) was conducted to investigate the effects of total flight hours on Beginning and Ending RRS scores. The number of participants per cell was not equal, *n* ranging from 25 to 35 per cell as depicted in Table 1.

Table 1. *Mean and Standard Deviations of RRS scores Participants by Total Flight Hours*

Flight Time	n	Beginning RRS (*SD*)	Ending RRS (*SD*)
< 5,000	35	7.03 (2.24)	2.69 (1.20)
5,001 – 10,000	42	6.24 (2.07)	2.69 (1.20)
10,001 – 15,000	45	5.98 (1.92)	2.82 (1.03)
> 15,000	25	5.76 (1.76)	2.56 (0.87)

A significant difference was found for the main effect of flight time on the RRS scores, F (3,143) = 3.11, $p < .05$ with a moderate effect size (partial Eta squared = .06). Estimated marginal means and standard errors were evaluated post-hoc (*a posteriori*) for differences using the Least Significant Differences pairwise multiple comparison test. No significant interaction effects with the dependent variable (RRS scores) were found. There were significant differences ($p < .05$) in the means of the beginning scores of the lowest time pilots as compared to each of the other three groups (5,001 – 10,000; 10,001 – 15,000; > 15,000). A graphical depiction of the effects of total flight time experience is displayed in Figure 1.

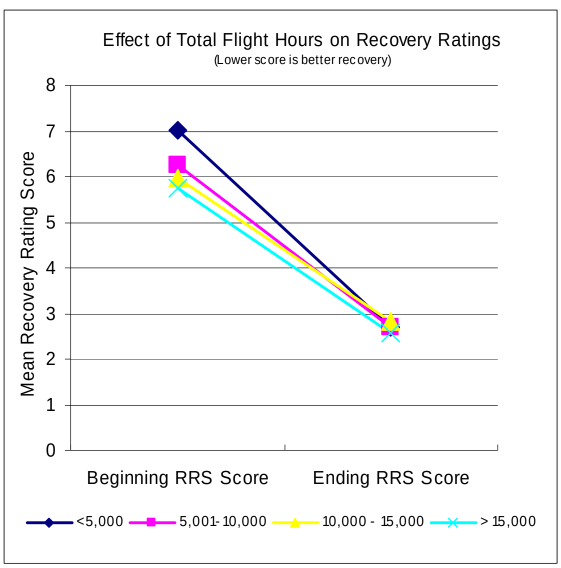

Figure 1. Effect of total flight hours on upset recovery improvement.

Effects of Type of Training on Upset Recovery Rating Scores

A repeated-measures univariate analysis of variance (ANOVA) was conducted to investigate the effects of military training on Beginning and Ending RRS scores. The number of participants per cell was not equal as displayed in Table 2. A significant difference was found for the effect of type of training on the RRS scores, F (1,145) = 4.41, $p < .05$ with a small effect size (partial Eta squared = .04) as presented in Table 2.

Table 2. *Mean and Standard Deviations for Different Types of Training*

Type of Training	*n*	Beginning RRS (*SD*)	Ending RRS (*SD*)
Civilian Only	93	6.46 (2.21)	3.01 (1.23)
Military and Civilian	54	5.93 (1.71)	2.55 (0.79)

Effects of Aerobatic Experience on Upset Recovery Rating Scores

A repeated-measures univariate analysis of variance (ANOVA) was conducted to investigate the effects of previous aerobatic training on Beginning and Ending RRS scores. No aerobatic training was indicated by “None” while recreational or minimal aerobatic training was designated “Some”. “Extensive” aerobatic training was either (former) military pilots or those who had performed in airshows.

A significant Levene’s statistic ($p < .05$) was found for the Beginning and Ending RRS scores; therefore ,corrections to the alpha level were made for these analyses. The means and standard deviations for each aerobatic experience group are shown in Table 4.

A significant difference was found for the main effect of previous aerobatic training on the RRS scores, F (2,146) = 4.71, $p = .01$ with a moderate effect size (partial Eta squared = .06). Estimated marginal means and standard errors were evaluated post-hoc for differences using the Least Significant Differences pairwise multiple comparison test. No significant interaction effects with the dependent variable (RRS scores) were found. There were significant differences ($p < .005$) in the estimated marginal means of the pilots with no aerobatic experience, and those with extensive aerobatic experience for beginning RRS scores.

Table 3. *Mean and Standard Deviations of RRS scores for Participants by Aerobatic Experience*

Aerobatic Experience	*n*	Beginning RRS (*SD*)	Ending RRS (*SD*)
None	36	6.89 (2.35)	3.33 (1.49)
Some	51	6.29 (2.14)	2.88 (0.99)
Extensive	62	5.90 (1.71)	2.54 (0.84)

Effects of Flight Instructing on Upset Recovery Rating Scores

A repeated-measures univariate analysis of variance (ANOVA) was conducted to investigate the effects of flight instructing on Beginning and Ending RRS scores. No significant difference was found for the effect of flight instructing on the RRSs, F (1,148) = .317, p = .57 as shown in Table 4.

Table 6. *Mean and Standard Deviations for Flight Instructors and Non-Flight Instructors*

Flight Instructor	*n*	Beginning RRS (*SD*)	Ending RRS (*SD*)
No	49	6.41 (2.20)	2.93 (1.18)
Yes	101	6.24 (1.99)	2.83 (1.10)

Participants' Perceptions of the Upset Recovery Training Program

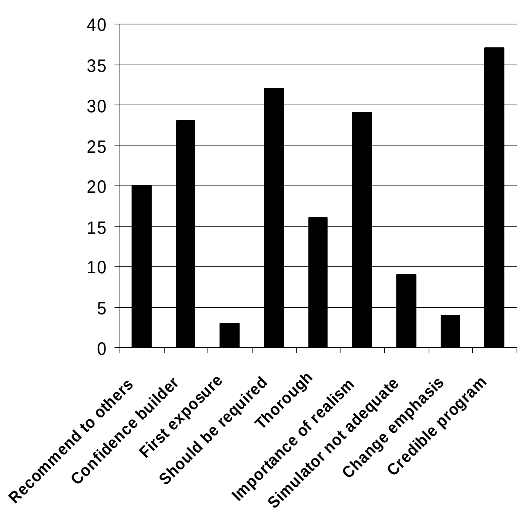

Figure 3. *Frequency and types of responses to open-ended question, "comments on course," from Upset Recovery Training course evaluations.*

Participants' perceptions of the URT program were evaluated through frequency analysis. A summary of these results are presented pictorially in Figure 2 which shows the frequency of overall comments from the participants in the study. In addition, the relative importance of each element of the in-flight simulator Learjet flight was determined by rank order (1 = fair to 5 = excellent) of the mean scores from the participants' course evaluation forms.

Discussion

Overall, these results suggest a strong positive influence of the Upset Recovery Training Program on a pilot's ability to respond to an inflight upset. Specifically, the RRS scores indicate a very strong training effect. It is interesting to note the effect of flight times on the Beginning and Ending RRS scores. Even though there is a significant main effect, the bulk of the variance between groups is with the lowest time pilots (< 5000 hours) and those pilots with > 5000 hours. This suggests that pilots of all experience levels (based on total flight time) gain essentially the same benefit from the training.

The effect of military training on the RRS scores is also worth noting. The effect size is particularly small and the majority of the variance was found to be in the beginning scores. Since many civilian pilots have not had the opportunity to perform aerobatics, it is not surprising that there is a significant effect of having experienced aerobatic flight on the RRS scores. Additional analyses were undertaken to determine how much of this effect was confounded by military training. When the effects of aerobatic experience were held constant, there were virtually no differences in the two groups.

The lack of a significant effect of experience as a flight instructor on RRS scores is not particularly surprising. This specialized, advanced airmanship type regimen is not currently taught to flight instructors, therefore it is also not taught *by* flight instructors (Federal Aviation Administration, 2002). Furthermore, most airline pilots no longer participate in instruction outside of the airline.

Although this adaptation of the Cooper-Harper Scale (RRS) has not been previously validated as a measurement instrument, these results offer some interesting insights into its usefulness. First, the six instructor pilots assisting in the use of the scale found homogeneity in their participants' Beginning and Ending RRS scores. This suggests that the scale is being used in a consistent manner across all users. Next, the RRS scores follow known trends in pilot expertise research for total flight time (Jensen, 1995), whereas the lower time pilots showed significantly higher Beginning RRS scores than the higher time pilots. Finally, the RRS scores followed the hypothesis that pilots with aerobatic training would have lower RRS scores (better performance) than those without any aerobatic training. Neither the total flight time, nor the aerobatic experience level of the participants was known to the instructor pilots before the training event minimizing experimenter bias. Therefore, one could conclude that the RRS is a valid measure of this task.

The participants' perceptions of the program appear very positive in these data; however, their scores reflect a ceiling effect on their ratings of the specific course elements.

Limitations of Current Study

There were numerous limitations in the current research avenue of this hybrid training-research program. The most salient issues (which we are already addressing) were:

- The study population was self-selected volunteers who, in many cases, held positions in the airline's flight training department or management. Even though there were no significant effects of being a flight instructor, per se, on RRS scores, there is still a need for a more representative population to be able to better generalize the results of the study.
- Extraneous variance in the experimental setting and measurement techniques resulting mostly from the nature of aviation (weather, mechanical malfunctions, etc.) were present.
- The testing effect of a repeated measures design without a control or pretest condition.
- Mono-operation bias whereas the focus was on only one aspect of a multi-part training program. The influence of the academics, Bonanza flight, and ground simulator were not controlled for or measured in the current format of the study.
- The use of only one performance measurement technique created a mono-method bias which only measures one aspect of the construct of training effectiveness.
- The possible unreliability of the RRS where measurement error may have occurred as it has not been cross-validated or scrutinized as an accurate representation of pilot performance.

Conclusions and Future Direction

Future directions of the study will focus on addressing the identified limitations discussed above. New protocols, forms, and data collection efforts have been established and implemented as countermeasures to the threats to validity which can be controlled or measured. For example, instructor calibration and collection of more detailed flight time and pilot experience will aid in controlling potential confounding effects in the study. More exacting pilot demographics and flight times allow for control of experience levels. Measurement techniques have been established to specifically test for the training effects at each stage of the training. These enhancements to the research protocol will provide richer data from which the training program will ultimately benefit.

Acknowledgements

This research was supported, in part, by the U.S. Department of Transportation, Federal Aviation Administration Human Factors Research and Engineering Division, Contract Number DTFA01-02-C-00088, and the New Mexico Department of Transportation-Research Bureau. The authors wish to thank all of the pilots who have participated in the Flight Safety Research Center Upset Recovery Training Program.

References

Gawron, V. J. (2000). *Human performance measures handbook.* Mahwah, NJ: Lawrence Erlbaum Associates.

Federal Aviation Administration (2002). *Flight Instructor Practical Test Standards for Airplane with Change 1.* FAA-S-8081-6B. Washington, DC: Flight Standards Service.

Jensen, R. S. (1995). *Pilot judgment and crew resource management.* Aldershot: Ashgate Publishing Limited.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.