

Validation of LOC-I Interventions

Judith Bürki-Cohen^{} and Andrea L. Sparko[†]*

USDOT/RITA/Volpe National Transportation Systems Center, Cambridge, MA 02142

The basic tenet of this paper is that today's national airspace systems, at least in advanced industrial countries, qualify as so-called Highly Reliable Systems (HRS). In an HRS, even the type of accident that causes the most fatalities is a rare event. This means that in an HRS, the avoidance of accidents is a frequent event. Therefore, the best way to improve an already highly reliable system would be to learn from the cases where accidents have been avoided. This is not possible, however, because you can't learn from what is unknown. Instead, safety managers resort to retrospective analyses of the most deadly accidents overall. In an unreliable system, it makes sense to correct what is wrong. In an HRS, however, any mitigation efforts that arise from rare, unpredictable, and often unique events carry great danger to upset the balance of the HRS. Such interventions must be scrupulously vetted, in a series of steps that become increasingly costly as the series progresses. This paper makes some suggestions for these steps. If the anticipated benefit from the intervention is not worth the cost of such a thorough review for unintended consequences, then it may be better to accept the existing high reliability of the system as good enough and leave the system unchanged.

Nomenclature

AA	=	American Airlines
AAMP	=	Advanced Aircraft Maneuvering Program
AF	=	Air France
AFM	=	Atmospheric Flight Mechanics
AOA	=	Angle of Attack
ARC	=	Aviation Rule making Committee
ASAP	=	Aviation Safety Action Program
ASRS	=	Aviation Safety Reporting System
ATC	=	Air Traffic Control
AURTA	=	Airplane Upset Recovery Training Aid
CFIT	=	Controlled Flight Into Terrain
CVR	=	Cockpit Voice Recorder
EASA	=	European Aviation Safety Agency
FAA	=	Federal Aviation Administration
FDR	=	Flight Data Recorder
FOQA	=	Flight Operation Quality Assurance
FFS	=	Full-Flight Simulator
GNC	=	Guidance, Navigation, and Control
HBAT	=	Handbook Bulletin for Air Transportation
HPM	=	Human Performance Model
HITLS	=	Human-In-The-Loop Simulation
HRS	=	Highly Reliable System
IATA	=	International Air Transport Association
ICATEE	=	International Committee for Aviation Training in Extended Envelopes

^{*} Principal Technical Advisor, Aviation Human Factors Division, RVT-82, 55 Broadway. Senior Member AIAA.

[†] Engineering Psychologist, Aviation Human Factors Division, RVT-82, 55 Broadway. Member AIAA.

IQTI	=	IATA Training and Qualification Initiative
LOC-I	=	Loss of Control In Flight
LOSA	=	Line Operations Safety Audit
MST	=	Modeling and Simulation Technologies
MTOW	=	Maximum Takeoff Weight
NAA	=	National Aviation Authorities
NASA	=	National Aeronautics and Space Administration
NTSB	=	National Transportation Safety Board
OEM	=	Original Equipment Manufacturer
SAFO	=	Safety Alert for Operators
SOP	=	Standard Operating Procedures
SME	=	Subject Matter Expert
TNO	=	Netherlands Organization for Applied Scientific Research
VMS	=	Vertical Motion Simulator

I. Introduction

Loss of Control In Flight (LOC-I) accidents have been identified as the number one reason for loss of life in today's world-wide air transportation system.¹ Many national and international efforts are underway to mitigate this state of affairs. To name just a few: the International Committee for Aviation Training in Extended Envelopes (ICATEE) under the auspices of the United Kingdom's Royal Aeronautical Society; the European Aviation Safety Agency's (EASA) "Gain 60 Seconds" initiative; the United States (U.S.) Federal Aviation Administration's (FAA) FAA/Industry Stall/Stick Shaker Training Working Group convened in the wake of the 2009 Colgan Air/Continental Connection Flight 3407 accident (which resulted in a Draft Advisory Circular AC 120-STALL); and an Aviation Rule making Committee (ARC) on Stick Pusher and Adverse Weather. The International Air Transport Association's (IATA) Qualification and Training Initiative (IQTI) is also concerned with LOC-I. This paper is one of over 30 papers addressing LOC-I presented in six LOC-I sessions at the combined Atmospheric Flight Mechanics (AFM), Guidance, Navigation, and Control (GNC), and Modeling and Simulation Technologies (MST) American Institute of Aeronautics and Astronautics conferences. These sessions were organized by Dennis Crider of the National Transportation Safety Board (NTSB) and Christine Belcastro of the National Aeronautics and Space Administration (NASA) and the first author. The same team organized similar sessions in 2008 and 2010.

With all these efforts underway, the industry will soon be bombarded with a plethora of proposed LOC-I awareness, prevention, and recovery strategies, ranging from procedural to training to technical solutions. The main purpose of this paper is to discuss how these methods could be validated. Before embarking on a discussion of the issues involved with validating the proposed LOC-I mitigations, however, we would like to sound a word of caution. With all this energy and expertise focused onto one problem, it is easy to lose sight of a very important fact: We are living the safest period in the history of civil aviation transportation.² In fact, there are those who say that we have reached the point of diminishing return and that efforts to make the system any safer would require such resources as to bankrupt the industry.² Figure 1 illustrates that in any system and for any tool, the cost-safety benefit function will eventually asymptote, and those last 10 percent or so to achieve perfect safety remain elusive without exorbitant investment. This investment, in fact, may drain the resources of an organization to a point where safety may be compromised.

These are questions for those charged with balancing the value of human life with the need for a thriving airline industry indispensable for international trade and access to remote areas. The concern of this paper is that if you already have such a highly reliable system (HRS), any intended improvement to the system must be very carefully validated to ensure that it does not upset the balance of the system and introduce harmful unintended consequences. The first rule for any interventions must be, especially for an already very safe system, DO NO HARM.

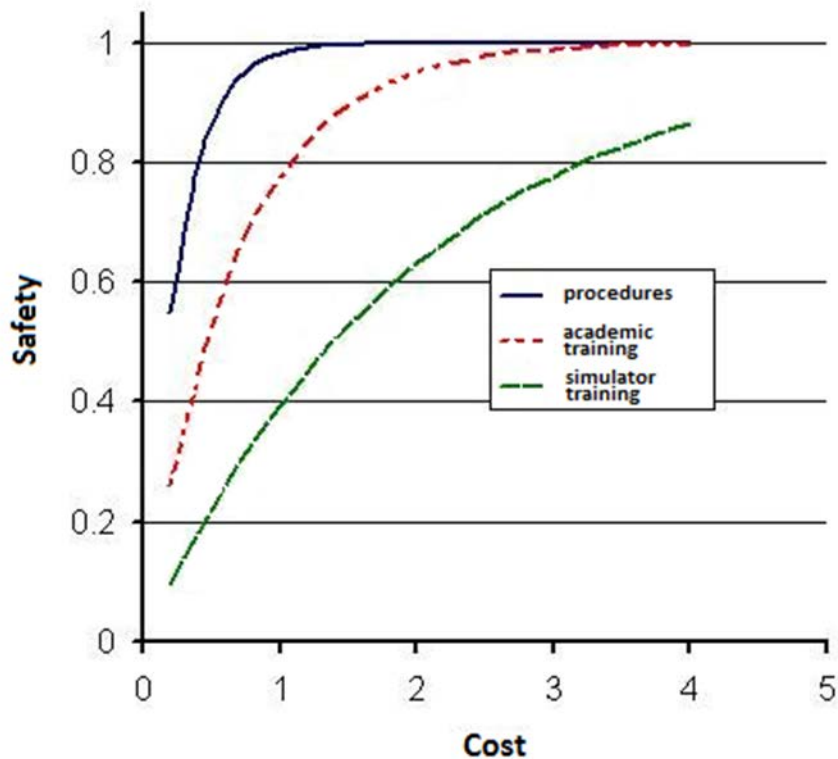


Figure 1. Hypothetical cost-safety curves for three types of mitigation.

II. A Cautionary Tale

A prime example of how a well thought-out and intended effort may have contributed to the worst accident in U.S. aviation history can be found in the events that led up to the crash of American Airlines (AA) Flight 587. In this accident, the pilot's rudder reversals in response to what is presumed to have been a mild wake vortex led to the separation of the Airbus' 300-600 tail fin. The airplane plunged into the ground, killing 260 people on-board and five on the ground.³

In response to rising concerns with LOC-I accidents and several NTSB recommendations triggered by accidents (e.g., A-94-173 by the United Express Flight 6291 stall on approach to Columbus, OH), AA took proactive measures.⁴ Even before the FAA's Flight Standards Handbook Bulletin for Air Transportation (HBAT) 95-10, "Selected Event Training," was issued on August 16, 1995, AA initiated the development of its Advanced Aircraft Maneuvering Program (AAMP). According to the NTSB AA 587 accident report,³ AA diligently involved the original equipment manufacturers (OEMs) of most of the airplanes represented in its fleet, inviting comments, traveling to the Boeing Company, and even organizing a two-day AAMP Industry Conference with participation from FAA, NTSB, Boeing, McDonnell Douglas, Airbus, and the U.S. military. AA was generally hailed for its initiative, although the issues of rudder and the fidelity of the simulator outside its validated flight envelope were raised even during this development and vetting phase. In fact, less than three months after the AAMP Industry Conference, in August 1997, FAA, Boeing, and Airbus sent a joint letter to AA stating that the AAMP was already "excellent," but elaborating on the danger of rudder reversals to the structural integrity of the tail fin. The FAA and the OEMs specifically recommended that "the hazard of inappropriate rudder use" during wake turbulence "should also be included in the discussion." AA replied that it did so, and that the booklet handed to pilots (including the AA 587 pilots) during AAMP ground school including the recommendation "High AOA [Angle of Attack] maneuvering=RUDDER" was "not a standalone document and nothing should be inferred without listening carefully to the presentation." AA also stated that the danger associated with rudder excursions at high AOA was "clearly exemplified by" showing NTSB videos on two LOC-I accidents.⁴

However, more than simply opinions might have alerted AA to the danger of unintended consequences of its AAMP rudder instructions. Just two weeks before the AAMP industry conference, on May 12, 1997, AA Flight 903 experienced an upset causing serious injury to a passenger and minor injury to a flight attendant. The pilot of AA 903 also had applied full rudder to control roll. The fin of this Airbus 300-600, however, held steady, and the flight crew recovered the airplane in time. In response to this accident, AA’s managing director of flight operations-technical stated in a memorandum to AA’s chief pilot and vice president of flight that AAMP’s roll control instructions in unusual attitude recoveries were “not only wrong, [but] exceptionally dangerous. American Airlines is at grave risk of a catastrophic upset.”⁴

The lessons of the AA 587 accident illustrate how difficult it is, even with the best of intentions, to get it right, especially when you are dealing with an already very safe system. We reiterate, first do no harm. To avoid that, listen to any cautions. Beware of single-shot narrow solutions. Ensure that you understand the full complexity of the system and of the problem before intervening. In a way, the actions of the AA 587 pilot flying illustrate this point: Misunderstanding the hazard associated with what would have been an easily absorbable wake, he took preventative actions that took his own and many others’ lives.

III. Mitigations for LOC-I

A. Extent and Complexity of Problem

Before introducing any interventions to reduce LOC-I, we first must understand the extent of the problem and its underlying causes. With regard to the extent of the problem, although LOC-I have been the leading cause of fatalities in accidents involving large Western-built jet transports since 2005, they have been relatively stable over the past ten years. It’s the reduction in Controlled Flight Into Terrain (CFIT), that has catapulted LOC-I into first place, not an actual increase in LOC-I (see Figure 2).

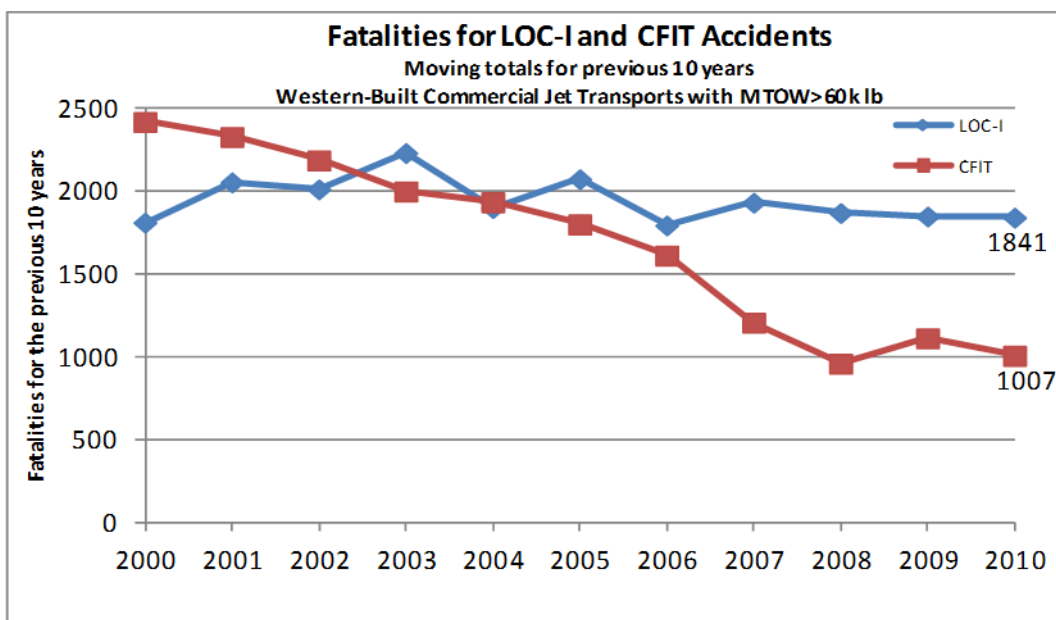


Figure 2. Fatalities for LOC-I and CFIT accidents. This figure shows that LOC-I fatalities held steady from 2000 to 2010, while CFIT fatalities have decreased. Courtesy of The Boeing Company. © MTOW=maximum takeoff weight.

The complexity of the problem stems from the many different causes for losing control of an airplane in flight. Pilots tend to say that no two upsets are alike, even in the same airplane type. The FAA/Industry Airplane Upset Recovery Training Aid (AURTA) covers the causes of upsets on 11 pages, categorizing them into environment- and pilot-induced upsets (or a combination of the two).⁵ Paul Railsback, Director of Operations at Airlines for America, differentiates between four broad categories of LOC-I, each requiring its own interventions: Stall, Incorrect Airspeed from Pitot/Static Systems, Wake Turbulence, and Missed Approaches.^{†1} Each of these has very distinct

warning signs (or lack of warning signs) and must be handled differently. And each of these, if not recognized in time and responded to specifically (or, in some cases, not responded to at all), can result in extreme attitudes, from which it would be very hard to recover. Mr. Railsback knows firsthand the potential for unintended consequences resulting from a one-size-fits-all attempt to mitigate LOC-I, having been one voice from within AA warning of the danger of training to use rudder for roll control.⁴ But he would agree that even his categorization would need further refinement to develop tailored solutions that are effective and don't entail unintended consequences.

B. Goals and Types of Interventions

Any type of intervention must have as its primary goal a decrease in the occurrence of LOC-I. One way to accomplish this is strictly technical, by improving the stability and reliability of the airplane and its components. Other ways are more complex and can be addressed in several ways, such as increasing the predictability of LOC-I events and empowering pilots to respond effectively once they have become aware of the danger of losing control. So, you need a fail-safe airplane, with excellently human-factored flight-deck indications and annunciations that give appropriate warnings and advisories supporting effective cross checks; and you need a flight crew that is informed, alert, mindful, and flexible enough to anticipate, avoid, recognize, and respond correctly to impending or actual LOC-I events. This means that the crew should also be trained to "expect the unexpected" to avoid surprise reactions and tunneling of its attention.⁶

Aside from the stability and reliability of the airplane, most interventions, even apparently technical ones, involve the flight crew. Envelope protections and automated systems may surprise pilots by taking uncommanded actions; or pilots may start to rely on them too heavily and be unable to take over when needed. Pilots may also not understand them sufficiently and try to counteract them, such as the pilot of Colgan Air Flight 3407 who fought the stick pusher to the bitter end.⁷ Flight deck indications and annunciations must be seen, heard and understood by pilots, and discrepancies between indications must be recognized. The pilots of Air France (AF) Flight 447 did not appear to hear 54 seconds of consecutive aural "stall stall stall" warnings (that were partly masked by a concurrent aural C-chord warning horn); if they did, they may have dismissed it for lack of a complementary visual indications.⁸ The saying that "what you don't know can't hurt you" is blatantly false. What you don't know or don't understand is precisely what will hurt you, as Nassim Taleb (and David Hume before him) tried to teach us in his 2007 bestseller "The Black Swan."⁹ The black swan is a metaphor for a rare, unexpected, and tragic event, and it takes only one "black swan" to shatter the most consistent safety record. So, OEMs must be careful to design transparent systems with an easily understandable logic. Displays must contain well-organized and appropriately integrated information that is accessible to pilots when and where they need it with the appropriate saliency; and in designing them, OEMs must be mindful of pilots' sensory, perceptual, and cognitive processing limits. OEMs and avionics designers must work with human factors professionals to adapt their systems to human capabilities to create, in the end, a safe and effective human-machine system.

The next target of interventions is the flight crew itself. They must receive thorough academic instruction on – and testing of – their understanding of the aerodynamics and the systems of their airplane. There is much room for improvement in content, delivery, and verifying the success of such instruction. In many ways, this is the lowest hanging fruit, the least costly intervention, but with a potential for tremendous benefit. If the Flight 587 pilot had been fully aware of the power of the rudder of his wing-mounted-engine airplane, or of the design limits of the airplane's fin, would he have operated the pedals the way he did? Although the availability of such information in a panic situation cannot be taken for granted, there is much information that could prevent pilots from being panicked in the first place. A full understanding of the life cycle of a wake might have prevented the 587 pilots from overreacting.

Another relatively cost-effective intervention is redesign of standard operating procedures (SOP) and checklists. SOPs and checklists are designed to help pilots remember the correct sequence of actions once they have recognized a situation via their academic and simulator training. One example of such a change, is the Safety Alert for Operators (SAFO) issued by the FAA Air Carrier Training Branch 16 months after the Colgan Air Flight 3407 accident.¹⁰ The SAFO clarifies the meaning of the "minimal loss of altitude" criterion for evaluating approaches to stall contained in the FAA Practical Test Standards.¹¹ It cautions that "the reduction of Angle of Attack (AOA) during initial recovery will likely result in altitude loss," the amount of which will depend on the operational environment; and trainers should "not mandate a predetermined value for altitude loss." This change is expected to help pilots to overcome the almost instinctual urge to pull back during especially a low-altitude stall, and instead to first reduce AOA and gain speed.

Although academic training and procedures contribute to improving the stick and rudder skills necessary to prevent or recover from LOC-I, there is wide consensus that stick and rudder skills require hands-on training. Hands-on training interventions, however, put the highest demands on the training tool because of the danger of

negative training. Hands-on training for LOC-I recognition, awareness, and recovery may thus be the most costly and perhaps also the most “dangerous” intervention. Hands-on training tools cover a broad range of levels of fidelity, ranging from fixed-base training devices to Level D Full Flight Simulators (FFS) to in-flight simulators to all-attitude capable airplanes. Specialized tools ranging in fidelity from Barany chairs to centrifuges to NASA’s Vertical Motion Simulator (VMS) to the Netherlands Organization for Applied Scientific Research’s (TNO)’s DESDEMONA are also available (see Ref. 6 for an overview of training devices). With all these tools, the critical validation issue is not necessarily their physical fidelity, but transfer of training to the airplane. Negative training transfer is possible in any of them, because none fully corresponds to the real airplane for which pilots are being trained. This is due to a lack of validated flight data for upsets on the one hand and limitations in the physical capabilities of these tools on the other. For Flight 587, the NTSB specifically lists the lack of fidelity of the simulated rudder response as a factor in the accident.³

The above makes it clear that there is no one single best road to mitigate LOC-I, and that any road taken involves some danger to make things worse instead of better. The entire human-vehicle system must be considered to ensure that an already safe system does indeed become safer; and even beyond that, the foresight and the imagination of the airline company, the OEMs, and the regulator all need to join together for the common goal of improving the recognition of and the response to LOC-I.

IV. Validation Methods

A. Challenges

The past decade or so has brought many so-called black-swan events: the terrorist attacks of September 11, 2001; the stock-market crash in 1987; the size of the tsunami leading to the nuclear crisis in Japan in 2011. Although each of these events, in retrospect, had at least some precursors, authorities were unable to recognize their significance. As mentioned earlier, Taleb describes a black-swan event as an event – or lack of an event – with the three attributes of rarity, extreme impact, and retrospective predictability only.⁹ Given that LOC-I has been identified as the number one cause of fatalities in commercial air transport, calling LOC-I accidents black-swan events may contradict the rarity attribute.¹ A look at the overall safety in aviation and the reason why LOC-I is in first place discussed earlier and shown in Figure 2, restores the rarity attribute, however. Nobody would dispute that the loss of 265 lives in the AA 587 accident is at least locally a catastrophic event with extreme impact. Was it predictable? As discussed above, there were authoritative voices that cautioned AA regarding the way AAMP trained rudder use. Also, there was an earlier accident in which rudder reversals played a role. Nevertheless, AAMP continued with only slight modifications to the course materials (booklet and video) based on the OEM’s concerns. Apparently, the presumed benefit of AAMP outweighed the perception of risk by AA’s ultimate decision makers. Retrospective validation of the AAMP program via the 587 accident shows that they were wrong. So, it appears that LOC-I events do indeed fulfill Taleb’s definition of black-swan events as much as any of the generally-accepted examples given earlier.

However, if LOC-I events truly are black swans, then any measures to prevent them are shots in the dark that cannot be validated. In Taleb’s definition, a LOC-I event is more akin to a grey swan, which follows at least some laws of nature. It was these “laws of nature,” derived from earlier accidents and the NTSB recommendations and the later SAFO, that the AAMP program tried to address, with unfortunate and unintended consequences. But can we conclude, even retrospectively, that AAMP was a total failure? We do know what happened to Flight 587, but how do we know how many accidents were prevented by pilots heeding other aspects of the training? One major problem with retrospective validation is that it can only be applied to what went wrong, not to what went right. There is rarely fame for preventing a black swan, only blame and shame for not preventing (or for causing) one in retrospect.

So how do we validate prevention efforts for an event as rare, multifaceted, hard to predict and surprising as LOC-I? How do we even demonstrate that these efforts follow the first rule to be observed when intervening in a HRS, to DO NO HARM? The following section will systematically go through the different options. The first option is not really an option; it is mandatory to ensure that no negative effects from the intervention. The subsequent options are to check whether the mitigation will indeed help reduce the occurrence of LOC-I.

B. First, Do No Harm

A first effort must certainly be a careful vetting of any new mitigation, be it procedures, technologies, or training, by consulting subject matter experts (SME) from all walks of the industry, including pilots, trainers, regulators, the NTSB, and OEMs. The goal of this effort is to identify – if they exist – any potentially harmful outcomes of the mitigation. As the evaluation of the AAMP has shown, however, not all experts are sufficiently

farsighted or imaginative to recognize the hidden dangers in proposed mitigation strategies, and wise naysayers may be overruled in such informal consultations.

A proven way to enrich the quality and the consistency of expert feedback is to lead the experts through a cognitive walkthrough. Typically, cognitive walkthroughs are used to evaluate user interfaces,¹² and are thus particularly useful for validation of new technologies. They may prove just as effective in discovering flaws in procedures or training, however. To conduct a cognitive walkthrough for LOC-I mitigations, the analysts conducting the walkthrough first define a particular scenario (e.g., respond to a stall warning in an airplane with a stick pusher). The pilots are then asked to “walk through” the scenario one step at a time. The goal of the walkthrough is to identify discrepancies between what the pilots are doing and what the creator of the mitigation expected them to do. The procedure can be conducted with a variety of tools – from paper mockups to FFS – making it potentially a very cost-effective strategy. Another advantage is that in such a procedure, data generation is open ended and serendipitous, unlike in experiments that aim at testing a narrow hypothesis. This advantage is shared with field observations (described later), with an additional advantage: The data are collected in an environment that is safe enough for pilots to freely explore the task. This is not true for field or line observations, where high stress and SOPs prevail. In other words, in cognitive walkthroughs pilots are allowed to act as they would do instinctually – rather than according to prescribed procedures – which provides a better indication on how they would perform during a surprising and disorienting LOC-I event.

Accidents are also a useful tool for determining potential and unanticipated problems. As discussed earlier, accidents may point to what went wrong, and they are extensively used to develop mitigations. But in addition to this, they can also be used as a “what if” model to validate mitigations hypothetically, examining whether a particular mitigation could have helped prevent a particular accident. Incidents can also be used for that purpose. Unlike accidents, the latter generally have the advantage of being reported first-hand by the pilot, and may even contain suggestions regarding mitigations. Accidents and incidents, then, may also be useful if recreated in a cognitive walkthrough with the proposed mitigation.

C. Simulator Studies

Once a highly reliable organization or system is reasonably sure that a mitigation does not jeopardize its high reliability by introducing unintended consequences, mitigations can be examined for their potential beneficial effects, by provisionally implementing them and testing their effect in real operations. However, there are several obstacles to validating LOC-I in the airplane. Two are practical: Real-life LOC-I events are too dangerous and are too hard to provoke “naturally,” i.e., without a confederate pilot resorting to subterfuge. A third is methodological: It would not be possible, in the air, to keep all variables [weather, traffic, Air Traffic Control (ATC) communications] besides the mitigation constant. So it would not be possible to say with a reasonable degree of certainty that emerging effects are due to the mitigation.

The most practical and scientific method to overcome these obstacles would be to first study a mitigation’s effect in a simulator, in “with vs. without” comparisons. The operational availability of academic training could also be tested in carefully crafted simulator scenarios. The benefit of validating mitigations in the simulator is that pilots’ behavior and performance responses can be measured directly, via scenarios designed to test the behaviors targeted by the mitigation. For “with vs. without” comparisons, measured performance is compared between those who have received the mitigation and those who have not. With a tightly-controlled experiment design, any between-group performance differences will confirm that the mitigation indeed had an effect on performance. In some cases, instead of simply aggregating behavior (e.g., into a group average), it may be preferable to assess improvement (or decrement) on an individual level, especially when looking at the effects of training.¹³ Pre- and posttest designs are often used for this reason – to calculate the amount of performance gain demonstrated in response to the mitigation. This type of comparison, additionally, tends to have greater statistical power to find an effect if there is one, and a lower probability to erroneously conclude that the intervention has no effect. Even with significant statistical results, however, it is up to SMEs to determine whether the performance benefit is large enough to justify implementing the mitigation.

One concern with trusting the accuracy of performance benefits (or decrements) of mitigations that have been demonstrated in a simulator, however, may be the fidelity of the simulator. Even the most advanced FFS owned or leased by airlines are unable to accurately simulating all types of unusual attitudes, especially if they go beyond the stall break – due in part to the physical limitations of the simulator motion platform, but also to the difficulty of obtaining the flight data necessary for modeling such LOC-I events. The usefulness of FFS as a stand-in for the airplane to validate mitigations is therefore questionable. Another obstacle is that LOC-I events are so perilous and so unusual in the actual airplane that pilots often react to them with surprise, and this reaction is directly associated with profound negative physiological and cognitive responses which severely impact a pilot’s ability to respond

appropriately to the event.⁶ The difficulty then is how to reproduce, in the simulator, the same physiological and psychological response to a LOC-I event as in real life. The safety of a ground-based stand-in for the airplane, even with the most carefully designed scenarios, allows only limited validation of a mitigation intended to help pilots overcome surprise responses.

One option for a better stand-in for the airplane is in-flight simulation. In-flight simulators modify the handling qualities of a more agile airplane (such as a Learjet) to simulate the handling qualities of a larger and more sluggish transport airplane. The types of base airplanes used for in-flight simulations are able to reach beyond the normal flight envelope of the simulated airplanes, where LOC-I events occur (still, the accuracy with which they truly represent LOC-I remains ultimately unproven). A safety pilot can revert to the original configuration any time a maneuver exceeds even the normal envelope of the base airplane. Other stand-ins for the airplane that could be used for testing transfer of mitigations are advanced motion simulators such as the VMS or DESDEMONA, which would at least overcome most of the motion limitations, although not the lack of flight data. Despite the high cost of in-flight simulations and advanced simulators, their use may be justified for *validation* of interventions. For AAMP, such a validation may have helped the naysayers carry the day. For use of such advanced tools for routine initial and recurrent training of airline pilots, however, it would be impossible to absorb the cost – or risk, for in-flight simulators – for a world-wide population in the hundreds of thousands,^{§ii} not to mention the lack of availability of such tools. Thus it would be especially critical to validate the effect of new LOC-I training procedures developed for FFS or other more readily available tools by first testing quasi-transfer to the closest representation of the airplane available.

D. Field Observations

The limitations of replicating truly realistic LOC-I scenarios in the simulator may make field observations preferable for some validations. Again, however, we run into the black swan issue; the rarity and unexpected nature of LOC-I events make it impossible to anticipate when LOC-I events can be observed. The only remedy for this is to lengthen the observation time and to lower the criterion for what is considered an upset event. This would allow the inclusion of LOC-I precursors that may be found in incidence reports or Flight Operational Quality Assurance (FOQA) databases. True observational techniques also require a person to do the observing, such as in Line Operations Safety Audits (LOSA) – a serious expense for long observation periods.¹⁴ A solution, of course, is to have the pilots themselves do the observing, which in essence is already done via programs such as the FAA’s Aviation Safety Action Program (ASAP) and NASA’s Aviation Safety Reporting System (ASRS), as well as other national and international programs. This is also a cyclical strategy, considering that pilot reports often lead to the development of mitigations. Variations of field observations may also be useful, such as in-depth or unstructured interviews or case studies.¹⁵ But these too are cyclical, since interviews may have been employed in the vetting phase (do no harm), and case studies are essentially covered by accident reports.

“Found” experiments offer another alternative. Like field observations, “found” experiments compare data that have occurred naturally in the field. Usually, these studies arise from an unplanned event that has already occurred, such as LOC-I events. The analysis involves the examination of data from before and after the event, to seek out the most probable cause. The challenge, however, is to obtain data from after the event – especially considering the devastating outcomes of many LOC-I events. It took two years to find the Flight Data Recorder (FDR) and the Cockpit Voice Recorder (CVR) of Air France Flight 447. Even if an FDR is recovered, the sampling rate may be as low as one Hertz for some parameters. CVR data from today’s accidents may be as short as 30 minutes and may be overwritten if the accident is such that the CVR continues to receive power.^{16,17}

Retrospective data can also be used to *look for* a difference (even if one has not been noticed) between two groups that differ on a single element. For example, it may be of interest to compare fleets with and without an AOA indicator to see whether pilot performance differs. The challenge with any retrospective data, however, is availability; FOQA data may be the only useful means of performing retrospective analyses, but such data are heavily protected from unauthorized access and thus very hard to obtain.

E. Pilot Performance Models

Given that all aforementioned methods have major issues affecting the accuracy or the statistical power of the results, a last method discussed is the use of pilot performance models to investigate the effect of different LOC-I mitigation methods. Human performance models (HPMs) are the holy grail of human factors research – wouldn’t it be great if we could do away with the costs and limitations of human-in-the-loop simulations (HITLS) and model the human along with the system? However, human, or in the context of LOC-I, pilot-performance models are only

as good as the data used to create them, and here we have the same fidelity problem as with the simulator: just as we don't have validated flight data beyond the normal flight envelope for the simulator, we don't have all the data necessary to model a system of such infinite complexity as a human being.

By far the best-modeled element of human performance is manual control, starting with the pioneering work of Duane McRuer in the 1950s and 1960s.¹⁸ However, on today's flight decks, manual control has been relegated to a relatively minor role during normal operations, having been largely replaced by supervisory control and other cognitive activities that are much less transparent and thus much harder to model. Nevertheless, because of their obvious practical advantages over full-fledged HITLS, HPMs may soon be able to provide useful information for system development and in system evaluation.

An excellent introduction to the capabilities and limitations of HPMs is given in the 2008 report on NASA's 6-year Human Performance Modeling project, edited by David Foyle and Becky Hooley.¹⁹ This project was conducted as part of NASA's Aviation Safety and Security Program. Five modeling teams were tasked to apply a different HPM to two pilot performance problems: 1) pilots' taxi navigation errors in current operations; and 2) pilot behavior and performance during approach and landing using a future synthetic vision system displaying the airport environment. The five HPMs applied to this problem were 1) Active Control of Thought-Rational (ACT-R); 2) Improved Performance Integration Tool/ACT-R hybrid (IMPRINT/ACT-R); 3) Air Man-machine Integration Design and Analysis System (Air MIDAS); 4) Distributed Operator Model Architecture (D-OMAR); 5) Attention-Situation Awareness (A-SA) model. Each of these well-known models was generated for a different purpose and has different strengths and weaknesses. The five teams were provided with a rich set of HITLS data relevant to the two pilot performance problems for model development and validation. Armed with these data, the teams were able to modify the models to explain observed behaviors related to the two problems, suggest design or procedural changes, and explore those in what-if simulations. This led Richard Pew, the Principal Scientist at BBN Technologies, to state, in the foreword, that "modeling has advanced to the point that it is able to represent relatively complex aircraft - air traffic control interactions with sufficient realism to assess alternative systems and procedure designs." It appears, however, that the availability of HITLS data was critical to the success of the models, and that at least to date a combination of HITLS and HPM may be the most promising tool.

V. Conclusion

The basic tenet of this paper is that today's national airspace systems, at least in advanced industrial countries, qualify as so-called Highly Reliable Systems (HRS). In an HRS, even the type of accident that causes the most fatalities is a rare event. This means that in an HRS, the avoidance of accidents is a frequent event. Therefore, the best way to improve an already highly-reliable system would be to learn from the cases where accidents have been avoided. This is not possible, however, because you can't learn from what is unknown. Instead, safety managers resort to retrospective analyses of the most deadly accidents overall. In an unreliable system, it makes sense to correct what is wrong. In an HRS, however, any mitigation efforts that arise from rare, unpredictable, and often unique events carry great danger of upsetting the balance of the HRS. Such interventions must be scrupulously vetted, in a series of steps that become increasingly costly as the series progresses. The paper makes some suggestions for these steps. If the anticipated benefit from the intervention is not worth the cost of such a thorough review for unintended consequences, then it may be better to accept the existing high reliability of the system as good enough and leave the system unchanged.

VI. Acknowledgments

This is an invited paper reflecting the current thinking of the authors, not of the Department of Transportation or any of its administrations and institutions. The authors thank the Federal Aviation Administration (FAA), NextGen Human Factors Division (ANG-C1) for enabling their participation in AIAA LOC-I prevention activities, in particular our Program Managers Tom McCloy and Michelle Yeh. We thank Tom McCloy for his thoughtful review of an earlier draft. At the Volpe Center, we thank our Chief, Maura Lohrenz, for comments on an earlier draft; and Amanda Mattson for support with the references and figures. Special thanks go to Robert A. Curnutt, Technical Fellow at The Boeing Company, for providing Figure 2, and Paul Railsback, Director of Operations at Airlines for America, for lending his expertise.

VII. References

- ¹"Statistical Summary of Commercial Jet Airplane Accidents," Boeing Commercial Airplanes, Seattle, 2011.
- ²Zajac, A., "Airline Crash Deaths Too Few to Make New Safety Rules Pay," *Bloomberg Business Week* [online journal], URL: <http://www.businessweek.com/printer/articles/268182?type=bloomberg> [cited 23 July 2012].

³National Transportation Safety Board, “In-Flight Separation of Vertical Stabilizer American Airlines Flight 587 Airbus Industrie A300-605R, N14053, Belle Harbor, New York, November 12, 2001,” NTSB/AAR-04/04, 2004.

⁴Ladkin, P. B., “The Crash of AA587: A Guide,” RVS Group, Rept. RVS-RR-04-03, University of Bielefeld, 18 November 2004.

⁵Carbaugh, D., Rockliff, L., and Vandal, B., “Airplane Upset Recovery Training Aid Revision 2,” *Flight Safety Foundation*, URL: http://flightsafety.org/files/AP_UpsetRecovery_Book.pdf [cited 23 July 2012].

⁶Bürki-Cohen, J., “Technical Challenges of Upset Recovery Training: Simulating the Element of Surprise,” *Proceedings of the AIAA Modeling and Simulation Technologies Conference*, Toronto, Ontario, 2010.

⁴National Transportation Safety Board, “Loss of Control on Approach Colgan Air, Inc. Operating as Continental Connection Flight 3407 Bombardier DHC-8-400, N200WQ Clarence Center, New York February 12, 2009,” NTSB/AAR-10/01, 2010.

⁸Bureau d’Enquêtes et d’Analyses, “Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro – Paris,” July 2012.

⁹Taleb, N. N., *The Black Swan: The Impact of the Highly Improbable*, 2 ed., Random House, New York, 2007.

¹⁰Federal Aviation Administration, “SAFO: Possible Misinterpretation of the Practical Text Standards (PTS) Language ‘Minimal Loss of Altitude,’” SAFO 10012, July 2010.

¹¹Federal Aviation Administration, “Airline Transport Pilot and Aircraft Type Rating Practical Test Standards for Airplane,” FAA-8081-5F, 2008.

¹¹Wharton, C., Rieman, J., Lewis, C., and Polson, P., “The Cognitive Walkthrough Method: A Practitioner’s Guide,” CU-ICS-93-07, 1993.

¹³Kirkpatrick, D. L., “Evaluation,” *Training and Development Handbook: A Guide to Human Resource Development*, edited by R. L. Craig, 3rd ed., McGraw-Hill, New York, 1987, pp. 301-319.

¹⁴EUROCONTROL, “Line Operations Safety Audit (LOSA),” *Skybrary* [wiki], URL: [http://www.skybrary.aero/index.php/Line_Operations_Safety_Audit_\(LOSA\)](http://www.skybrary.aero/index.php/Line_Operations_Safety_Audit_(LOSA)) [cited 23 July 2012].

¹⁵Burgess, R. G., *In the Field: An Introduction to Field Research*, Unwin Hyman Ltd, London, 1984.

¹⁶EUROCONTROL, “Cockpit Voice Recorder (CVR),” *Skybrary* [wiki], URL: [http://www.skybrary.aero/index.php/Line_Operations_Safety_Audit_\(LOSA\)](http://www.skybrary.aero/index.php/Line_Operations_Safety_Audit_(LOSA)) [cited 23 July 2012].

¹⁸McRuer, D. T., and Weir, D. H., “Theory of Manual Vehicular Control,” *Ergonomics*, Vol, 12, No. 4, 1969, pp. 599-633.

¹⁷Kaminski-Morrow, D., “Cockpit Recorder Overwritten in Qantas A380 Engine Incident,” *Flight International* [online magazine], URL: <http://www.flightglobal.com/news/articles/cockpit-recorder-overwritten-in-qantas-a380-engine-incident-349690/> [cited 23 July 2012].

¹⁹Foyle, D. C., and Hooley, B. L. (eds.), *Human Performance Modeling in Aviation*, CRC Press/Taylor & Francis, Boca Raton, 2008.

ⁱ Information from Paul Railsback on different types of LOC-I categories comes from a personal communication (e-mail from July 17, 2012)

ⁱⁱ In 2010, there were 103,000 airline and commercial pilots in the U.S. alone (<http://www.bls.gov/ooh/transportation-and-material-moving/airline-and-commercial-pilots.htm>)