SUMMARY

Excessive flightcrew fatigue as a result of trip exposure has long been cited as a factor with potentially serious safety consequences. Laboratory studies have implicated fatigue as a causal factor associated with varying levels of performance deterioration depending on the amount of fatigue and the type of measure utilized in assessing performance. From an operational standpoint, these studies have been of limited utility because of the difficulty of generalizing laboratory task performance to the demands associated with the operation of a complex aircraft.

This study examined the performance of 20 volunteer twin-jet transport crews in a full-mission simulator scenario that included most aspects of an actual line operation. The scenario included both routine flight operations and an unexpected mechanical abnormality which resulted in a high level of crew workload. Half of the crews flew the simulation within two to three hours after completing a three-day, high-density, short-haul duty cycle (Post-Duty condition). The other half of the crews flew the scenario after a minimum of three days off duty (Pre-Duty condition).

The results of this study revealed that, not surprisingly, Post-Duty crews were significantly more fatigued than Pre-Duty crews. However, a somewhat counter-intuitive pattern of results emerged on the crew performance measures. In general, the performance of Post-Duty crews was significantly better than the performance of Pre-Duty crews. Post-Duty crews were rated as performing better by an expert observer on a number of dimensions relevant to flight safety. Analyses of the flightcrew communication patterns revealed that Post-Duty crews communicated significantly more overall, suggesting, as has previous research, that communication is a good predictor of overall crew performance.

Further analyses suggested that the primary cause of this pattern of results is the fact that crewmembers usually have more operating experience together at the end of a trip, and that this recent operating experience serves to facilitate crew coordination, which can be an effective countermeasure to the fatigue present at or near the end of a duty cycle. These results have important aircrew training and aviation safety implications.

INTRODUCTION

1.1. Background

This report is the third in a series on the physiological and psychological effects of flight operations on flightcrews and the operational significance of these effects. These studies were conducted in response to a Congressional request. The original response to this request was a workshop convened by the National Aeronautics and Space Administration (NASA) and held at the Ames Research Center (ARC) in August, 1980. This workshop included representatives from the scientific community, airline pilots, and management; its expressed purpose was to make recommendations regarding the type of research necessary to understand the extent of the problem.

One of the conclusions reached at this workshop was that little was known about the effects of exposure to duty cycles on actual flight performance. Despite impressive advances in aircraft technology over the past several decades and an overall decline in the accident rate since the introduction of turbine-powered aircraft, flightcrew performance problems continue to dominate the statistics. There are, of course, many hypotheses concerning the persistence of operator performance problems in this environment, and the issue of pilot fatigue has traditionally received considerable attention. This high level of interest has stimulated a large volume of laboratory research, but much of this work is difficult to generalize to the actual operations, and the issues surrounding its applicability have prompted considerable disagreement about the extent and operational significance of fatigue-related performance decrements. As a result, NASA was asked to undertake a comprehensive program of research in order to assess whether or not fatigue-related problems are prevalent in long- and short-haul flight operations. The two major facets of this project are 1) to assess the psychophysiological effect of exposure to various types of flight and duty cycles and 2) to determine the operational significance of this exposure in terms of flight safety and efficiency. This project also sought to identify individual pilot attributes that might be associated with responses to the operational requirements of air transport flying and to identify adaptive strategies that might enable certain individuals to cope more effectively with operational requirements.

This is the second report in the short-haul series. It describes one aspect of this comprehensive program, the results of a study designed to address the operational significance question in short-haul flight operations (see Gander,

Graeber, Foushee, & Lauber, 1986; for a discussion of the psychophysiological effects). In addition, a brief overview of research aimed at fatigue-performance relationships is provided.

1.2. Related Research

Aside from research focusing on cockpit design and engineering issues, there have been relatively few pilot performance studies carried out in actual operational environments (see Graeber, Foushee, and Lauber, 1984; for a more comprehensive discussion). The reasons for this state of affairs are varied. Field research is very difficult to do because of an almost complete lack of experimental control over operational events. Moreover, flight safety considerations prohibit the examination of many variables in the in-flight environment. Thus, studies aimed at the effects of fatigue on performance have generally been confined to laboratory or part-task research environments and have focused upon individual instead of flightcrew performance.

Holding (1983) reviewed many of these studies and much of the related literature on fatigue-task performance relationships. One of the major conclusions of this review and others (e.g. Gagne, 1953) is that fatigue-performance relationships are not easy to define. The literature is populated by many studies demonstrating a deleterious effect of fatigue on some type of task performance, and about an equal number demonstrating minimal or no effects. Those tasks that do show deterioration over time (presumably as a response to some type of fatigue) tend to be relatively simple tasks characterized by repetitive stimulation. The pattern of findings for more complex tasks is less clear.

The work of Bartlett (1943) is commonly cited as one of the major contributions to the literature on complex task fatigue. In this body of work (often referred to as the Cambridge Cockpit Studies), subjects were exposed to tasks which consisted of responding on aircraft-type controls to "changes" in a variety of instruments. Fatigue was manipulated by exposure to these tasks over long periods of time. The findings of these studies seemed to indicate that as alertness declined, progressively larger "deviations" were tolerated before any corrective actions were taken by subjects. It appeared that "fatigued" subjects were more prone to distraction and seemed to suffer from a narrowing of perceptual focus such that attention was reserved for items of more central importance, such as heading and speed indicators. It was further observed that performance on these tasks became more variable (as opposed to less accurate). Moreover, subjective observations indicated that subjects became more irritable with increasing fatigue ("violent" language, etc.).

On the other hand, McFarland (1953), in a study of aircraft incident and

accident statistics, could find no effect for the fatigue associated with long hours of flying (although accidents are a crude measure of pilot performance). Likewise, Chiles (1955) could find no fatigue effect even though subjects performed continuously in an aircraft simulator for as long as 56 hr without rest, except for intervals where they left the simulator for testing on a tracking task. It is reported that some of Chiles' test subjects had to be carried to the test apparatus, and the absence of performance effects associated with fatigue under these circumstances is indeed striking. However, aircraft flying performance was not directly measured--the primary performance criterion was associated with the tracking task. Despite what would appear to be high levels of fatigue in this study, performance on this task was well within normal limits.

Studies of automobile driving have yielded similar inconsistent results. Brown (1967) found no performance effects associated with fatigue and found that performance on a vigilance task actually improved over time (when fatigue effects should have begun to appear). In a particularly interesting study, Dureman and Boden (1972) found some evidence of performance deterioration on a simulated driving task (as measured by steering errors and braking reactions). However, when subjects were threatened with electric shock, a factor which apparently increased arousal, performance improved. These data suggest that arousal and accompanying increases in motivation may be important mediators of fatigue-performance relationships.

It is difficult to summarize the implications of the fatigue-performance literature for air transport operations because most of the studies have not been conducted in actual flight environments, or the tasks have not been relevant to the task of flying a complex transport aircraft. Even the studies that have utilized tasks with some face-validity to piloting an aircraft (e.g. Bartlett, 1943) were crude simulations by recent standards, and the primary performance criteria were relatively simple monitoring or vigilance tasks that did not comprise the entire performance spectrum. Another problem with previous studies is the manipulation of fatigue. Most studies have artificially manipulated fatigue by keeping subjects up, exposing subjects to long and continuous performance periods, or testing subjects after rigorous physical exercise. Few studies have coupled consistent manipulations of fatigue with comparable performance measures, which makes comparisons across studies hazardous at best. Even fewer studies have utilized *exposure* to normal duty cycles as a means of inducing fatigue, causing further difficulties in generalizing laboratory evaluations of fatigue effects to real-world settings.

Those studies that have examined performance as a result of actual exposure to flight operations (as opposed to more artificial manipulations of fatigue) have generally not utilized actual flight performance measures for assessment of effects (for reviews of this research see Klein and Wegman, 1980; Graeber, 1982) The usual approach has been to measure subjects' performance after exposure to

transmeridian flight (usually as passengers) on a battery of laboratory tests of psychomotor and simple cognitive performance. Only one study has measured performance in flight simulators using experienced pilots (Klein, Bruner, Holtmann, Rehme, Stolze, Steinhoff, and Wegmann, 1970). In this study, subjects were tested periodically after exposure to successive trips eastward and westward across eight time zones. Performance was measured by percent deviation from preset flight parameters, and exhibited significant deficits depending on such factors as time of testing, direction of flight, and number of days in the new time zone. While this is one of the few studies to demonstrate a potentially operationally significant effect of circadian dysrhythmia on performance, it left many questions unanswered. The measures utilized represent only an estimate of the effects of fatigue and jet-lag on raw flying skill and do not address other important predictors of performance in actual operations. The simulator tests lasted only 12 min, were limited to manual control skills, and did not tap any of the higher level cognitive skills required on the flight deck. Furthermore, the study examined a single-pilot operation. In multi-pilot operations there are a number of other factors that affect flight safety and efficiency.

Coupled with the problems of understanding fatigue effects on individual performance tasks is the increasing realization that individual pilot proficiency and the manual control skills necessary to operate a complex transport aircraft are clearly important, but not causal factors in the majority of incidents and accidents. An analysis of jet-transport accidents worldwide for the period 1968 to 1976 (Cooper, White, and Lauber, 1979) revealed more than 60 in which breakdowns of the crew performance and decision-making process played a pivotal role. Much of the aviation community seems to have accepted the notion that inadequate crew coordination, team performance, and the myriad variables that contribute to such difficulties are perhaps the most operationally significant problem areas in flight operations. Recommendations by the National Transportation Safety Board and the number of commercial and military training programs being developed to address "cockpit resource management" problems reflect this high level of acceptance. These attempts to address crew performance issues are in large part predicated on the notion that individuals will always make mistakes, and the crew should act as a system of redundancy to prevent more serious errors from occurring. There is a fair amount of research evidence (for a review see Foushee & Helmreich, 1986 in press) that the crew performance process is not working as well as it should be.

Pilots who are highly proficient in manual control skills continue to be involved in incidents and accidents with causes that fall outside the realm of the operators' psychomotor abilities. Yet, nearly all pilot performance research (and training) is centered around the measurement of dimensions that do not appear to be major problem areas in the current environment. No studies have examined the possible effects of fatigue upon higher-level decision-making skills and those

related to effective coordination of crew resources. Holding (1983) cites this failure to use measures of higher-order cognitive function, but he also mentions neglect in accounting for relatively stable effects of fatigue such as increased irritability. For example, the increased irritability associated with fatigue (e.g., Bartlett, 1943) could make it more difficult for crewmembers to work together effectively. Thus, fatigue effects may not be apparent on individual performance parameters, but significant with respect to group performance.

In the past, the major barrier to definitive research in all of these areas has been the lack of a research environment that combines experimental control with high generalizability to operations. Fortunately, the rapid advancement of simulator technology has finally provided an ideal laboratory for the study of these aircrew operating problems. It is now feasible to realistically simulate virtually every aspect of line operations to the extent that actual trips can be flown in a simulator that are almost indistinguishable from those in the airplane. Because of this high degree of realism, it is possible to study individual and group parameters with almost complete confidence that results generated in the simulator are representative of the real world. Moreover, the simulator affords a high degree of experimental control and allows the study of operational problems too dangerous to examine in flight. It is clear that arousal levels associated with realistic simulated flight more closely approximate those experienced in actual flight (at least when compared to laboratory performance tasks). In fact, Ruffell Smith (1979) reports arousal levels in high-fidelity simulation that closely approximating levels measured in flight.

With these issues in mind, our objectives were to conduct an investigation of the effects of exposure to high-density, short-haul airline operations that would be highly generalizable to the operational environment. Specifically, we were interested in whether there were any behavioral or performance changes associated with typical short-haul duty cycles, but more importantly, we were interested in whether these behavioral and performance changes (if any) were operationally significant. For the purpose of this investigation, operational significance was defined as performance that has a major bearing on flight safety and efficiency. Unlike the majority of previous research efforts, this study sought to measure performance on a realistic group performance task that induced high motivation and arousal. The task was also oriented toward complex decision-making skills rather than simple, repetitive tasks involving manual control skills.

METHOD

2.1. Study Design Overview

Since the primary objective of this study was to assess the operational significance of exposure to various types of flight and duty cycles, subject crews were evaluated either before or after they had completed a three-day trip. The "target trips" in this study consisted of high-density short-haul airline operations averaging eight hr of on-duty time per day and five takeoffs and landings, with at least one day (usually the last) averaging close to eight takeoffs and landings and thirteen hr on duty. There were two experimental conditions. 1) Subjects in the "Post-Duty" condition flew the simulation as if it were the last segment of a three-day trip, whereas 2) subjects in the "Pre-Duty" condition flew the simulation after a minimum of two days off duty (usually three), as if it were the first segment of a trip. Twenty volunteer crews were run in the study (40 pilots). One Post-Duty crew was eliminated from all analyses, when it was ascertained that the captain had been informed about the operational events associated with the scenario. This left 11 crews in the Pre-Duty condition and 9 crews in the Post-Duty condition.

2.2. Subjects and Recruitment

All subjects were recruited from the ranks of active line pilots in one domicile of a major U. S. air carrier. The decision to use pilots from one carrier was necessitated by differences in standard operating procedures and aircraft configurations that are common across air carriers. Thus, for experimental control purposes and to maintain the highest possible degree of realism, the scenario was designed to simulate precisely this particular airline's operation. All of the subject pilots were currently flying the transport aircraft used for the simulation exclusively in line operations.

Both airline management and union officials were contacted and briefed about the purpose of this investigation. When approval was received from both, all pilots in this domicile were sent a brochure describing the purpose of the project and providing information about the extent and type of participation requested from volunteer subjects.

Recruitment was based on several factors. Every month, the investigators received a copy of the trip pairings (which crewmembers are flying what trips) from crew scheduling. From these pairings it was possible to determine which crewmembers were flying and when, as well as, those crewmembers with days off

during any given period. The pairings were examined to determine which trips fit the criteria for the study (see Gander et al., 1986 for a thorough discussion of trip characteristics). In general, target trip selection was based upon such factors as the number of segments flown in a day, duration of the duty day, and length of nighttime layover. The most difficult duty cycles for this aircraft type were selected each month, and were characterized by a high number of takeoffs and landings, long duty days, and the shortest overnight layovers.

When these trips were identified, the pilots scheduled to fly them were contacted by telephone (if they had not previously participated in the simulator study) and recruited for assignment to the Post-Duty condition. They were scheduled for participation within two to three hr after completion of their duty cycle.

The subjects for the Pre-Duty condition were contacted if the trip pairings indicated that they were scheduled to be off duty for a period of at least three days. Due to the nature of trip pairings in this airline (and in short-haul operations in general), this is a relatively long continuous period off duty, with the exception of vacation time which is at the discretion of the pilot. Although all Pre-Duty subjects were scheduled for the simulation after it was ascertained that they would have three continuous off-duty days, 3 out of the 22 subjects in this condition had only two off-duty days. This occurred because of unscheduled extensions of prior trips that were usually the result of aircraft or weather difficulties forcing alterations of the scheduled pattern. This could not be determined until those involved in schedule alterations had arrived for the simulation experiment.

When contacted by telephone, all potential subjects were briefed on the details of the experiment by the principal investigator. This briefing included the purpose of the simulation study and what was expected from each volunteer subject. All of the subjects were informed that the simulation would be conducted exactly as if it were an actual flight, and crews were aware that the flight was a GSO-RIC segment, but they were not informed of any other details of the simulation scenario.

After obtaining tentative agreement to participate from prospective subjects, a considerable amount of effort was spent explaining the level of experimental realism so that the subject pilots would begin to think of the simulation as "just another flight segment." All pilots were asked to bring their own headsets, flight bags, including all charts and manuals, and anything else they normally take along during line operations.

Although subjects in this study (or any other part of the project) received no compensation for their participation, approximately 60% of pilots contacted agreed to participate in the experiment. The refusal rate was higher than in the

Gander et al. (field) study (15%), and most likely reflected the fact that participation would have caused considerable personal inconvenience. The highest refusal rate was among commuters not living nearby the simulator facility and crew domicile, since participation by many of these individuals would have caused them to remain away from home for an extra night. Field study participation, although requiring more overall time, did not entail any alteration of an individual's normal schedule. Of the pilots contacted, who resided within two hr of the simulation facility, the participation rate was above 90%.

2.3. Confidentiality

Because of the sensitivity of pilot-performance data in general and the focus upon operational significance in this investigation, it was necessary to guarantee all pilots participating in this study complete confidentiality. This was done in several ways. First, all data in this study were identified by a four-digit code number, which identified individuals as captains or first officers. Thus, it was not possible for anyone, including the NASA investigators, to identify any of the participating pilots by name. Second, although the simulation was conducted utilizing the facilities of the participating airline, the company was not involved in any way with the actual data collection. NASA leased the simulator from the company and provided its own operators for the purposes of this investigation. None of these individuals was employed by the subject airline.

2.4. Experimental Equipment

The simulator that was utilized for this study had a six-degree-of-freedom motion platform and four-window visual system. It was manufactured by CAE, Inc. of Canada, was equipped with the special effects and programmed with the aircraft performance data required to meet FAA Phase II certification, and had successfully completed the certification process.

Only minor simulator modifications were necessary for the experiment. These included the provision of input jacks so that background air traffic control (ATC) communications and the Automatic Terminal Information Service (ATIS) recordings for the various airports could be fed into the VHF radios no. 1 and no. 2. All equipment that was not functional in the simulator, such as the weather radar (the scope was present but not functional), the Automated Communications Addressing and Reporting System (ACARS), and the VHF no. 3 radio (present but not functional) was placarded as inoperative just as they would have been in the actual line operation. It was also necessary to develop software so that the simulator computer would output time-coded aircraft

performance parameters.

A portable video-data-acquisition-system (PVDAS) was designed and built by NASA for the purpose of this investigation. The advantages of this system are that it is pre-wired, and all components are mounted in a shock-proof case that can be rolled into the simulator cab and set up quickly. All of the simulator sessions were videotaped using PVDAS which included a two-channel, one-half inch VHS videotape recorder, monitor, camera, microphones, time-code generator, and three microcassette recorders. Both crewmembers and the air traffic controller were wired with lapel microphones; the captain on channel one and the first officer and controller on channel two. A single low-light, auto-iris, black-and-white, video camera was located above and behind both crewmembers at the approximate location of the cockpit door. This camera angle allowed a view of the center console, the throttle quadrant, and most of the instrument panel, with the exception of the overhead panel. About one-half of each pilot was visible, which allowed most actions to be monitored, but did not allow individual pilots to be identified. The time-code generator was used to synchronize the videotapes with the aircraft performance data and imprint a permanent record of timing information on each videotape. The microcassette recorders were also prewired and plugged into jacks mounted inside the simulator cab, which fed the VHF no. 1 and no. 2 radios. Each recorder was wired to switches so that any tape in any recorder could be fed into either radio. In this way, it was possible to provide all ATC background communications and ATIS information as appropriate depending on which radio was being utilized by the crewmembers. In this system, transmissions from the "live" controller blocked out the ATC background communications, but it was usually convenient for the controller to time instructions to the flightcrew accordingly.

Background ATC tapes were produced with the cooperation of each ATC facility that the pilots would be in contact with in the area of the simulated flight. These included GSO Ground Control, GSO Tower, GSO Departure, RIC Approach, RIC Tower, RIC Departure, ROA Approach, and ROA Tower. Each facility supervisor taped approximately 1 hr and 30 min of actual transmissions during a busy traffic period. The tapes were then edited for inconsistent information and normal pacing to produce a suitable amount of material for the simulated flight. Washington Air Route Traffic Control Center (ARTCC) tapes were produced by the investigators in the Man-Vehicle Systems Research Facility (MVSRF) air traffic control simulation at ARC using the voice-disguising equipment to simulate different aircraft in contact with Washington Center. These tapes of background communications were quite effective in producing a high level of operational realism.

Tapes for ATIS were produced by the investigators using 60-sec, continuous-loop, microcassette tapes. The tapes were mounted well in advance of the flight's approach to the appropriate facility and could be monitored at any

time and for any period because of the continuous-play feature. Tapes were produced for each of the airports used in the experiment, GSO, RIC, and ROA.

2.5. Personnel

The experiment was run with three staff members. The principal investigator dealt with the subject pilots, coordinated the data collection, and made decisions regarding the handling of flightcrew requests. He also acted as the number one flight attendant if requested by crewmembers. A second person delivered all air traffic control communications to the flightcrew and operated the simulator, including the introduction of scenario events. The third person was a retired check captain in the aircraft simulated. His role was primarily performance observation (see section 2.10), but because of his familiarity with company operations, also served as a technical adviser to facilitate operational realism. All three members of the experimental staff were present in the simulator during each run, but were out of the flightcrew's field of vision. Flightcrew members were informed that there would be no intervention by any of the experimental staff and that they should request all information through normal channels. They were advised that if simulator malfunctions caused unplanned events, they would be informed of the appropriate action. Once "airborne", intervention was necessary in only one case, when a simulator malfunction forced an early termination of the experiment approximately 10 min prior to touchdown at ROA. All data were used up until the time of the malfunction. The only exception to the non-intervention rule was that crews usually needed help during taxi operations because of poor visibility conditions and a lack of complete fidelity in the GSO ground scene (taxiway lights were low-intensity).

2.6. Experimental Procedure

When each crew arrived at the simulation facility, they were met by the principal investigator who again briefed the subjects on the purpose of the investigation. The importance of operational realism was again emphasized, and the pilots were urged to treat the simulation as if it were an actual flight. They were informed that they would have access to all resources that they would normally have in-flight, including complete ATC services, dispatch, access to maintenance, ATIS, and so on. However, no details were provided to flightcrews other than the flight's origin and destination. Subjects were asked to treat the flight as a "captain's segment", captain flying the aircraft and first officers performing support duties. This was done for experimental-control purposes and to assure consistent performance evaluation at both crew positions. After any

questions were handled, the flightcrew was escorted to a room set up as a flight dispatch facility, where they met with the flight dispatcher (played by the observer). The flight dispatcher provided the crews with the same information and paperwork that they would have normally received prior to a trip. This information included the route of flight, weather information for the vicinity of flight, fuel information, weight and balance information, number of passengers, and so on. All information was provided on standard company forms.

Following receipt of these materials, the crews performed flight planning duties normally reserved for this period. The dispatcher remained available to render any assistance the flightcrew might have requested. Crews were given a standard departure time based on their arrival at the simulator facility and were given adequate time for flight planning activities. After the flight planning phase, crews entered the simulator cab, were wired with microphones, and began preflight checks of the aircraft.

When it appeared that the crew was nearing completion of preflight checks and as departure time approached, the simulator operator extinguished the "aft cargo door light," prompting most flightcrews to realize the aircraft was loaded and ready. Shortly thereafter, the captain was called over the interphone by ground personnel and informed that the aircraft was ready. At this point, most crewmembers radioed for a clearance to pushback from the gate. Push-back was simulated by activating the simulator motion platform at the appropriate moment, which produced a small jolt quite similar to that associated with the tug beginning to push the aircraft backward. The recording of performance data began at push-back (see section 2.10.3). Taxi and ground operations were accomplished in accordance with standard operating procedures, as was the remainder of the simulated flight (the details of the scenario are discussed in section 2.7).

After landing, the crews were escorted back to the flight dispatch area where they filled out a number of rating forms (see sections 2.8-2.10). They were then debriefed by the principal investigator. In many cases, this was a rather extensive process since pilots were often anxious to discuss various events and actions taken during the simulated flight. Subject pilots were then thanked for their participation and received a lengthy explanation from the principal investigator concerning the importance of not discussing the details of the scenario with any other pilots (who might have later been experimental subjects). The initial briefings (when subjects first arrived at the facility) indicated that this procedure was quite successful. Only one captain (see section 2.1) was apparently aware of some of the scenario events. This was confirmed by initial questioning and later behavior during the simulated flight, and this crew was dropped from all analyses. All subjects were promised copies of the results when they became available.

2.7. Simulation Scenario

The experimental scenario was designed primarily to assess factors related to overall crew performance and decision-making. It was not intended to be a test of manual control skills, although data pertinent to these skills were available so that they could be assessed in certain critical flight phases. The scenario design was influenced by a pattern of events often seen in past incidents and accidents--relatively minor mechanical problems complicated by environmental and operational events.

Outlines of potential scenarios were developed by the NASA investigators and taken to the participating airline where specific details were developed in conjunction with subject-matter experts. For this purpose, the experts were two retired company management pilots were hired for assistance in standardizing scenario events to company operating procedures. Typical environmental conditions for the proposed area of flight were considered in great detail, so that scenario events relating to these conditions would be realistic to the experimental flightcrews. The experts also prepared all trip paperwork in standard company format and evaluated all events and procedures for accuracy and realism.

Following this process, nine "pre-test" runs were conducted with qualified flightcrews to refine procedures, train the experimental staff, evaluate the performance assessment techniques, and test scenario events. Feedback was elicited from pre-test subjects, and the videotapes were extensively evaluated. This process allowed continual refinement until the experimental scenario was finalized.

When crews arrived at the simulation facility and picked up the trip paperwork, they found the weather information for the vicinity of flight characterized by a frontal system passing through with low ceilings and visibilities (see Figure 1). This included the departure airport, which was still acceptable for takeoff but nearing the legal limit (RVR 1600, 1/4 mi). The airplane was relatively heavy, and the takeoff was runway-limited because the longer runway at the departure airport (GSO) was closed. Aircraft gross weight was 104,000 lb. Since weight was critical, it was dispatched with a minimum, but legal amount of fuel (13,200 lb). Because of this and the poor weather, which increased the probability of diversion, crews should have been concerned about the amount of fuel (possibility of extra fuel needed for a diversion), and as in the real world, crews had the option of requesting additional fuel from the flight dispatcher. This, of course, had implications for aircraft weight, and if more fuel was requested (prudent under the circumstances), baggage had to be off-loaded. The dispatcher was instructed to explain the implications of extra fuel to flightcrews and otherwise act reluctant to provide the extra fuel (as is sometimes the case). However, if the flightcrew could not be persuaded to go with the legal

minimum, the dispatcher agreed to add 5,000 lb extra fuel. Eighteen out of the twenty crews asked for and received this extra amount.

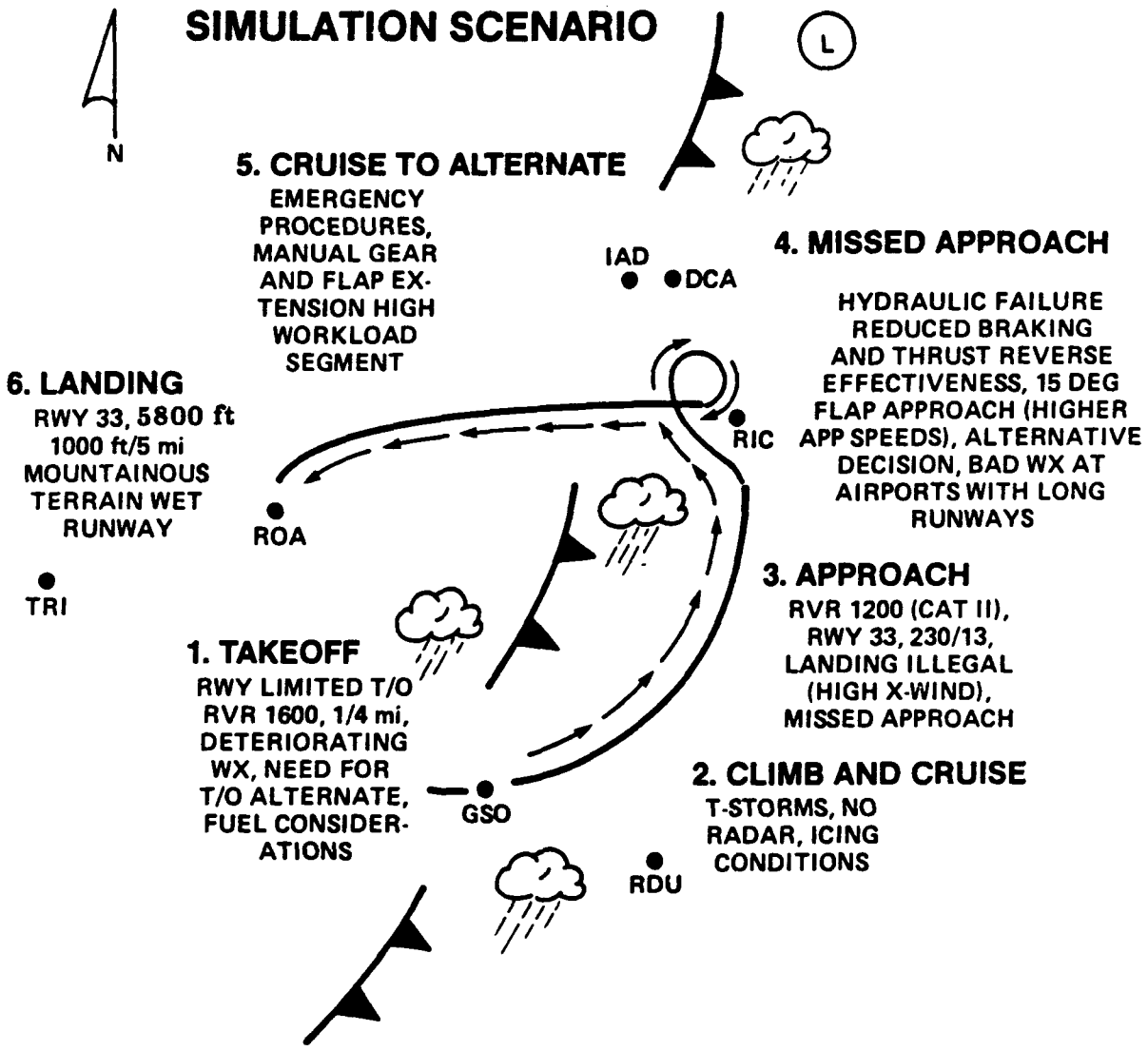

Figure 1. Overview of the simulation scenario (WX=weather, X-wind=crosswind)

After the crew had preflighted the aircraft and received clearance to push-back and taxi, ATC issued a special weather observation to all aircraft in the vicinity indicating that the weather had deteriorated (RVR 1600, 1/8 mi, rain and fog). The operational implications were that takeoff was still legal, but landing was not, and this meant crews were legally required to obtain a takeoff alternate, in case mechanical problems forced a quick return for landing. If a takeoff alternate was requested, ROA was provided by dispatch since it had the best weather within a relatively short distance from GSO. ROA was also given as the regular alternate landing site for RIC (as will be seen later, ROA was selected intentionally because of certain characteristics and because it was the site where all flightcrews were ultimately forced to land). As in all of the operational events "programmed" into the simulation, some crews realized this necessity and acted appropriately, while some did not. Once airborne, a relatively low-activity, routine segment was planned in order to allow crews to relax so that their behavior would more closely approximate actual flight behavior. This procedure has been found to be effective in both training and research applications of full-mission simulation (e.g. Lauber & Foushee, 1981). For this reason, most of the performance measures were taken after the end of this routine segment.

A few minor events were inserted in the low-workload segment as measures of crew vigilance. These included icing conditions (to make sure crews were paying attention to environmental conditions and utilized the anti-ice system when necessary) and unexpected rain and moderate turbulence along the route of flight. The aircraft was dispatched with the weather radar inoperative, and "vigilant" crews would have checked with ATC for ground-based radar advisories or pilot reports (Pireps) to assure sufficient separation from potential severe weather. Evaluation of crew awareness of such events was included in both observer-rating dimensions and in the error analyses.

The "high-workload" phase of flight was initiated when crews began their approach to the destination airport. When they received standard weather information for the arrival (via RIC ATIS), they discovered that the weather was poor and required a complicated, instrument approach (RVR 1200, Category II approaches to runway 33 in progress). Furthermore, there was a substantial crosswind on the active runway, and which was close to the legal limit (winds from 230 degrees at 9 knots). As the crews continued their approaches and contacted the tower for landing clearance, they were advised by the RIC Tower that the winds had increased to 13 knots from 230 degrees, which was 3 knots over the legal limit of a 10-knot crosswind component for a Category II approach. Some crews realized this and executed the required missed approach, and some did not. In any event, all crews, whether or not they realized that landing was illegal, were forced to execute a missed approach because those who continued the approach to RIC did not have visual contact with the ground at the decision height of 103 ft above the surface. During the missed approach, crews

experienced a "System A" hydraulic failure. The implications of this malfunction were complex: 1) landing gear and flaps had to be extended manually: 2) braking effectiveness was reduced, which meant increased stopping distances (the in-board brakes anti-skid system and thrust reversers had accumulator pressure only): 3) a 15-degree flaps approach was dictated when 30 to 40 degrees is the norm (15-degree limit is imposed in case a missed approach is necessary); 4) once the landing gear was extended manually, it could not be raised again (which has substantial implications for fuel consumption and subsequent missed approaches); and 5) nosewheel steering was inoperative (which meant that the aircraft would have to be towed off the runway.

Crews were then faced with a number of complicated decisions to make and procedures to execute. They had to decide where they were going to land, since the original destination did not have legal landing conditions, and in some cases there was only a limited amount of fuel (recall the original dispatch with minimum fuel). They had to diagnose the failure, realize the implications, and secure the failed system. Since higher approach speeds and reduced braking effectiveness were primary problems, it was clear that the most desirable alternate landing site was one with a relatively long runway.

The operational problems induced by the hydraulic failure were more severe than they normally would have been because of poor weather in the general vicinity (this was consistent with the initial weather briefing that crews received in dispatch prior to the flight) and limited fuel. The airports that should have been considered as alternates included TRI, IAD, DCA, RDU, CLT, and GSO, but all had visibilities of 1/4 mi or less, and some were closed due to weather conditions. The only reasonable alternate was an airport (ROA) with acceptable weather, but a with a relatively short runway (runway 33 at 5800 ft) that was wet, sloping downhill, and adjacent to mountainous terrain. ROA had a 1000-ft ceiling, 5 mi visibility, and winds from 360 degrees at 10 knots. The relatively good ROA weather was not surprising to most flight crews, who were familiar with the fact that it is an area frequently characterized by favorable conditions when other airports are marginal. Although the relatively short runway and mountainous terrain posed a dilemma for many flightcrews, it was the best choice compared to the other possible alternates.

Another feature of the scenario was that the manual-gear and flap-extension procedures were time-consuming and required a fair amount of pre-planning. Moreover, the flight time from RIC to ROA was relatively short (approximately 30 min), so time had to be apportioned carefully. In short, this simulation required a high level of crew coordination for effective performance--a good test of high-level decision-making and crew performance. It should again be stressed that some of the features of this scenario are similar to those seen in past incidents and accidents.

2.8. Demographic Data

Each subject pilot was asked to complete an extensive background questionnaire compiled to obtain information on demographic and lifestyle variables such as sleep, nutritional habits, and personality profiles related to pilot performance. (see Gander et al., 1986; for a more complete discussion). While many of these variables were intended for use in the physiological investigation, it was necessary to obtain this information in this study in order to assure that subjects in each condition were matched on variables such as age, general health, and experience. However, the relatively small sample size did not allow extensive analyses of performance as a function of these variables.

2.9. Fatigue Measures

Subject perceptions of fatigue were assessed utilizing the same techniques that were developed for use in the field study (see Gander et al., 1986; for a detailed discussion of these measures). Immediately after the simulation, subjects reported their sleep-wake schedules for each of the four previous nights. They were also asked to rate the quality of their previous night's sleep on four dimensions: difficulty falling asleep, deepness of sleep, difficulty arising, and how restful the sleep was. In addition, subjects completed a 26-item mood adjective checklist (e.g., Moses, Lubin, Naitoh, & Johnson, 1974) and estimated their level of fatigue by placing a mark on a 10 cm. line representing a continuum from most alert to most drowsy (e.g., Wever, 1979). All of these measures were taken from the "Daily Log Book" (see Gander, et al., 1986) used in the field investigation, and are shown in Appendix D.

2.10. Crew Performance Measures

A variety of measures were utilized in an effort to assess the performance of flight crews in this simulation study. These measures were aimed at the assessment of both individual and crew performance parameters. Since the primary objective was to determine whether any observed performance changes were operationally significant, a heavy emphasis was placed upon methods designed to quantify the crew performance process. As previously discussed (section 1.2), the reason for this emphasis was the finding that the vast majority of incidents and accidents in air transport operations appear to be due to breakdowns in crew performance rather than a lack of individual knowledge and skill.

The crew-performance measures included expert observer ratings, subjective assessments of workload, aircraft handling data, error analyses (both real-time and videotape), and crew communication patterns. Each will be discussed in detail.

2.10.1. Observer Ratings. Two types of ratings were obtained. First, a rating form was developed that was organized into categories relevant to the performance dimensions of interest in the test scenario. This instrument can be seen in Appendix A. It was partitioned by phase of flight (i.e., preflight, taxi/takeoff, climb, cruise, approach/missed approach, emergency procedure/hydraulic failure, cruise to alternate, and approach/landing). Within each section were the individual dimensions to be rated. The expert observer was asked to rate both the captain and the first officer on each dimension as it was observed and if it was applicable. Each dimension was scored on a five-point Likert scale. Scale anchors were defined as follows: 1=below-average performance; 2=slightly below-average; 3=average; 4=slightly above-average; and 5=above-average performance. Many of the dimensions were common to each phase of flight (i.e. crew coordination/communications, a/c handling, planning/situation awareness, procedures, overall performance, etc.). However, some were relevant only to specific flight phases or situations (e.g. thunderstorm awareness, stress management, takeoff alternate, etc.). If for some reason the category was not observed (not appropriate or the observer was uncertain), the observer was instructed to circle the "n/a" response provided with each scale. These ratings were made "real-time" as the simulation progressed. This type of rating form was designed as a systematic approach to the types of performance judgements made routinely by supervisory check pilots in training and evaluation. Thus, the observer (a retired check-captain) was highly experienced in making such ratings.

The second type of rating (Appendix B) was more general and completed by the expert observer immediately after the simulation. It consisted of nine categories (overall knowledge of a/c and procedures, technical proficiency, "smoothness," crew coordination and internal communication, external communication, motivation, command ability (for captains), vigilance, and overall performance). The observer completed this overall form for both captains and first officers These ratings were also made on five-point Likert scales (anchors defined same as above) and were intended to assess the expert observer's overall impression of performance throughout the simulation on each dimension.

2.10.2. Subjective Workload. Pilot perceptions of workload were assessed via a technique developed by Hart and coworkers (e.g. Hart, Battiste, & Lester, 1984). This rating form consists of 10 workload-related dimensions: task difficulty, time pressure, performance, mental effort, amount of attention required, complexity, busy-ness, motivation, fatigue, and overall workload (see Appendix C). Each dimension was scaled on 7-point Likert scales with high

numbers indicating larger amounts of the dimension in question. These measures have proven useful in assessing subjective levels of pilot workload in a number of experimental settings.

2.10.3. Aircraft Handling Data. These data were recorded directly from the simulator computer. Twelve parameters related to the aircraft configuration and handling were sampled every 15 sec. These included: airspeed, altitude, vertical speed, magnetic heading, engine #1 (engine pressure ratio) EPR setting, glideslope deviation, localizer deviation, gear position, flap position, #1 navigational frequency, and #1 DME reading, and elapsed time (see Appendix F). All measures were time synchronized with the videotaped records so that other performance parameters could be examined along with aircraft configuration. These measures were utilized primarily for analysis of final approach data at ROA and for crosschecking during the error analyses.

2.10.4. Error Analyses. Error analyses were undertaken using two independent sources of data. First, the expert observer kept a record of all errors observed during the course of the simulation. The second source of error data came from the videotape records. Using these records, two independent, "blind," observers reviewed the tapes for operational errors. When an error was recognized by one or both observers, the tape was stopped and the segment containing an alleged error was reviewed at least twice by the observers. After this process, both observers had to agree that the error had occurred or it was not counted in the analysis. Both sources of error data (real-time data collected during the simulation and data from the videotape analyses) were compared for reliability. The video-tape error-coding process captured all errors recorded real-time by the expert observer, however, it also revealed others that were missed during the course of the experimental runs. This is not surprising considering the fact that the videotapes could be endlessly reviewed. Reliability analyses revealed 81% agreement between both sources of error information, yielding a high level of confidence in these data.

Since assessing the operational significance of performance differences was a primary objective, an attempt was made to categorize errors according to their level of severity. This process was accomplished by both of the observers who had undertaken the videotape error-analysis. A three-level classification was utilized--Type I errors were defined as minor, with a low probability of serious flight safety consequences; Type II errors were defined as moderate severity, with a stronger potential for flight safety consequences; and Type III errors were classified as operationally significant errors, those having a direct negative impact upon flight safety.

Examples of Type I errors include missed clearances that were quickly corrected, checklists not run according to standard operating procedures (missing an item, or accomplished from memory), and altitude deviations of less than 200

ft. Type II errors included the failure to notify the company or ATC of an abnormal situation, altitude deviations between 200 and 400 ft., failure to use ice-protection, over or under speed for current configuration, delayed recognition or handling of the hydraulic failure, and failure to run a required checklist. Type III errors included failure even to recognize that the hydraulic failure had taken place, deployment of speed brakes and thrust reversers prior to touchdown, failure to consider alternates other than the company recommended ROA, failure to notice the crosswind or takeoff and landing restrictions, improper handling of the abnormal procedures, and altitude deviations of more than 400 ft.

Each videotape observer rated each error for severity and both observers discussed the rating for each recorded error. In cases where there was disagreement between observers about the correct classification, each observer presented his case in an attempt to reconcile the discrepancy. In most cases, agreement was reached. However, in a few cases when the observers could not agree, the error was assigned the lesser severity rating (there were no cases in which there was substantial disagreement between observers on a single error such as a Type I versus a Type III rating). All data that was used for analytical purposes represented the observers' agreed upon ratings.

2.11. Flightcrew Communication Patterns

Since past research has shown the interaction process of flightcrew members to be a significant predictor of flightcrew performance (e.g., Foushee & Manos, 1981), extensive analyses of communications data were undertaken in this investigation. The procedure used to analyze within-cockpit communication patterns was adapted from the Foushee and Manos (1981) procedure that was in turn derived from the work of Bales (1950). Using this approach, each statement or phrase was coded into one of eighteen categories of communication: command, observation, suggestion, statement of intent, inquiry, agreement, disagreement, acknowledgement, answer supplying information, response uncertainty, tension release, frustration/anger/derisive remark, embarrassment, repeats, checklist, non-task related, non-codable, or ATC communications. Coders worked independently. If there was any doubt about how a given speech act should have been classified, the coders were instructed to place them in the non-codable category. Thus, all speech acts were included in the total communications analyses, even those classified as non-codable. A complete list of communication categories and operational definitions is contained in Appendix E.

Two coders were trained extensively in the coding procedures by the principal investigator. The pre-test videotapes were used for training purposes so that prior to actual coding, neither coder had seen the experimental videotapes. The training consisted of the selection of particularly demanding 10-min segments

of the pre-test tapes on which coders practiced transcribing and coding. A point-by-point agreement method was used to calculate interrater reliability (e.g., Kazdin, 1982). This method is substantially more conservative than commonly used methods which compute interrater agreement based upon total frequencies. The point-by-point method is more conservative because it takes into account both the number of agreements and disagreements (or instances where one observer recorded one category and the other either recorded a different category or nothing at all). Total-frequency-based methods do not account for disagreements or non-events in this manner. Using this method, 71% agreement was established between the two coders.

After the coders' training had established a satisfactory level of reliability, each was randomly assigned half of the tapes to code. Each coder was assigned an equal number of tapes in the Pre- and Post-Duty conditions to rule out potential bias. Moreover, coders were blind to the condition when they were performing the communications analyses.

Although it would have been desirable to have each coder review all of the tapes, this was impractical due to the extraordinary amount of time necessary to code each tape (up to 40 hrs. per experimental run). As a compromise procedure to ensure adequate interrater reliability, coder agreement was checked halfway through the coding process and at the end using selected segments of the pre-test videotapes. On both reliability checks, an adequate level of reliability was evident, 74% on both occasions, which was slightly higher than the training criterion.