

MEASURING THE EFFECTIVENESS OF ERROR INVESTIGATION AND HUMAN FACTORS TRAINING

*C. G. Drury, I. Richards, A. Sarac and K. Shyhalla
State University of New York at Buffalo
Department of Industrial Engineering
Buffalo, NY 14260*

K. Woodcock, Consultant

*Jean Watson
Federal Aviation Administration*

September 25, 2000

1.0 EXECUTIVE SUMMARY

This study uses a novel methodology for studying how people investigate incidents and accidents. The aim of this first year of the three year project was to develop and validate a methodology that would be suitable for measuring operationally how effective are certain human factors interventions, such as human factors training or incident investigation training. The methodology uses a set of scenarios developed from actual incidents in conjunction with our airline partners. Participants in the study are given a short paragraph to trigger an incident investigation. The experimenter provides the data from a table of facts in each scenario, as requested by the participant. Participants gradually build up a data set appropriate to the incident, stopping when they decide that they have sufficient data to produce a report to management. They produce a verbal report that the experimenter records. The whole investigation takes only about 30-45 minutes per participant, and they find it to be quite realistic.

From the set of facts requested by each participant, and the set of facts selected by them for their report, we can measure how they emphasize different types of facts, such as those concerning the people involved in the scenario, or the environmental factors which may contribute to the incident. In this way, we will be able to measure how well people use the skills and knowledge they have learned in training courses by measuring their incident investigation performance before and after the course. In the current phase of the project, we developed six scenarios and tested them on 37 participants from two partner airlines, one partner repair facility, and an incident investigation facility.

The findings were that the sample of participants was quite representative of the industry. Most were Quality Assurance investigators, but some were managers/supervisors and aviation maintenance technicians (AMTs). Many had already received training in either Maintenance Resource Management (MRM) or the Maintenance Error Decision Aid (MEDA). In their performance, these participants found about a third of the facts available in each scenario, although not all facts available were relevant to the incident. They included about a quarter of the facts they had found in their summary of the incident. There were very few differences in performance between the different job titles or between those with and without the two training programs noted above. There were almost no correlations between performance and demographic variables such as age or AMT experience. In general, participants collected most information on the sequence of steps in the tasks performed, plus the contributing causes concerning individual and social factors of the characters in the scenarios. They were less complete in collecting data on equipment and environmental factors. The types of facts collected in the investigation were closely mirrored in the synopsis reports produced, showing that while many facts were omitted from reports, the selection is not biased towards particular fact types.

On the basis of the data collected, we propose a five-stage model to describe how people in aviation maintenance investigate incidents. This model starts with the trigger, that is the data given to the investigator to begin the investigation. Next, the investigator explores the boundaries of the incident, such as when it began and when it was discovered. The main fact-finding then takes place, with data gathering and analysis alternating to guide the search for relevant facts. At some point, the investigator decides to stop data collection, often when the assembled facts constitute what they perceive as a satisfactory explanation. At this point, the final stage is the selection of facts for the report. Within these stages, there is a heavy emphasis on collecting data relevant to rule violation, for example an [AMT](#) not performing a task step or an inspector failing to check the whole of a work area for foreign objects.

The six scenarios were evaluated for their suitability for future phases of the project. All were found to be satisfactory, although two were less so than the others. The final decision on which four to retain will be made in conjunction with the airline partners in Phase 2.

2.0 INTRODUCTION

This report gives the results of the first year of a three-year study that uses an incident investigation methodology to evaluate human factors interventions. The effectiveness is measured by how well industry participants are able to put human factors knowledge to work. The first year is a baseline study of the ability of various groups of industry personnel to investigate maintenance-related incidents.

2.1 SIGNIFICANCE AND OBJECTIVES

Continuing error reduction, particularly for human errors, has been a goal of the Gore Commission report (White House Commission on Aviation Safety and Security, 1997) and National Transportation Safety Board (NTSB) directives to the Federal Aviation Administration (FAA). Maintenance errors have been assuming greater prominence over the past several years, and now constitute a major threat to the continuing reduction in accident rates. Within the aircraft maintenance industry, the most common responses to this need have been human factors programs aimed at (a) training aircraft maintenance technicians (AMTs) and others in maintenance resource management (MRM), and (b) implementing human factors-based incident investigation methodologies, such as Boeing's Maintenance Error Decision Aid (MEDA). Both of these programs attempt to change the way [AMTs](#) and others approach their jobs by promoting greater understanding of the human factors considerations underlying human work and error causation.

This research program develops and validates a common methodology that can provide operational measures of relevance to industry for these two, and other interventions. By measuring how well participants investigate incidents, a common measure (percent of causal factors identified) is obtained. Also, diagnostic information on what participants still lack in effective use of human factors knowledge can be obtained from the same methodology. These measures will be used in years 2 and 3 as before-and-after evaluations of specific interventions as well as for on-going monitoring of program effectiveness.

Specific objectives are as follows:

Year

1

- To develop an incident-investigation-based methodology for measuring the effectiveness of human factors programs.
- To validate the methodology by measuring baseline performance of different groups of participants in the aviation maintenance industry.

Year 2

- To demonstrate the methodology by testing the effects of two interventions ([MRM](#) training and a [MEDA](#)-type investigation system) in a before-and-after study.

Year 3

- To use the results of the demonstration to develop improvements in these interventions, and test these improvements.
- To make the methodology available to the industry in the form of a self-evaluation package.

2.2 RESEARCH BACKGROUND

Throughout the [FAA](#)'s involvement with aviation maintenance and inspection human factors, reduction of human error has been a prime motivation. Accidents and operational incidents have been analyzed to determine both the active failures and latent failures in their causal chains. Active failures are the immediately-precipitating events which lead directly to the incident, while latent failures are those pathogens which lie dormant within the system and pre-dispose the system to active failures (Reason, 1990). Most traditional error-removal programs concentrate on the active failures by finding the person who caused the active failure and taking steps (discipline, propaganda, and training) to prevent that person from repeating the incident. Only recently (Marx, 1998) have airlines and repair stations begun to address common latent failures, such as poor equipment and procedures, which underlie a whole set of incidents. All incidents are unique, but their active and latent failures often exhibit common patterns going well beyond the "who is to blame" finding (Wenner and Drury, 1997).

This mismatch between what accident or incident data reveals and what conclusions are drawn limits the potential effectiveness of human factors programs. Part of the issue is how decision-makers interpret causality in the incident data. Marx (1999) investigated how various participants in the aircraft maintenance system assign causality or "attribution" to the different people and factors involved in an incident. Marx used scenarios, of about one page each, describing an incident and had participants find the "causes" of the incident in rank order. In this way, he found that there was wide variation in where participants stopped their causal search. Most participants found the dominant cause to be the erring employee, but they were willing to embrace long causal chains, back to the airline management level. In particular, rule violations were often seen as a major causal factor, particularly if the scenario listed a violation of the [FARs](#).

However, the scenarios on which these investigations are based have, in practice, come from an investigation into the incident. In the investigation process there are four phases: triggering, data collection, data analysis and the data reporting. Each will be considered in turn to define factors affecting the incident investigation process.

2.2.1 Phase 1: Trigger

An incident will only be investigated if the external results of the incident trigger some action. This trigger event can be a self-report, the existence of ground damage, a letter of intent from regulators, an operational event such as in-flight engine shutdown, or even an accident. In general, the rest of the investigation process will depend upon the trigger, although some systems such as [MEDA](#) can evaluate all types of error, both self-reported and externally triggered, using the same investigation procedure independent of the outcome. Given a triggering event, the magnitude of the consequences of the error often determines the way in which the investigation is carried out. Indeed, an influential safety text (Hammer, 1989) states that the exact number and personnel involved in an accident investigation *should* depend on the severity of the injury or damage (page 267), i.e. on the triggering event. To take an aviation maintenance example, a self-report of an [AMT](#) leaving a cleaning rag inside a door may only trigger a 30-minute [MEDA](#) investigation. But if the cleaning rag had caused a door to seal improperly, causing the aircraft door to lose pressurization rapidly, this could lead to a months-long [NTSB](#) investigation.

2.2.2 Phase 2: Data Collection

The investigation process itself is critical to the final outcome, i.e. to determining effective interventions. If some of the factors important to incident causation are omitted from the investigation, then false conclusions may be drawn. Typically, investigation procedures are described in legalistic terms rather than human factors terms. Investigators are exhorted to “follow all leads.” The usual admonition is to specify who, what, why and when (Ferry, 1981, page 1). Note that “how” is absent, but “why” is present. Most texts (Hammer, 1989, Chapter 15; Ferry, 1981, Chapter 3) concentrate on the physical methods of accident investigation, for example on how to preserve evidence, how to photograph the accident scene, or how to interview accident participants and witnesses. Brown (1995) shows that if an injury occurred, there might be more effort devoted to investigating the injury than to the accident which led to the injury (p. 982).

A major determinant of the quality of the whole investigation process is the depth to which the incident is investigated. Rasmussen (1990) uses the term “stop rule” to describe the decision criterion that the investigator uses to determine when “enough” data has been collected. At the lower limit, the investigator can stop at the “sharp end” of the incident (Reason, 1990) by determining who was to blame for the event, usually the last people to interact with the system before the incident. This level will usually suffice for legal purposes, for example by finding that a written rule had been violated. But as we move to what Maurino, Reason, Johnson and Lee (1995) call the dawning of a new era in incident investigation, we are increasingly seeking the latent pathogens in our complex and well-defended technological systems. How were carefully constructed barriers breached by local conditions? Maurino *et al* (1995) discuss stop rules in this light, and make a case for investigations going beyond the individual, despite the fact that the individual may indeed have made errors or even violations. But the scientific basis for such an extension is not obvious. If we expand our investigation to cover organizational causal factors, why stop there when all events must have prior causes? “*In theory, therefore, we could go back to the Big Bang*”, stated Maurino *et al*. Typically, we move in a more empirical direction and limit our investigation to those factors over which we have some control, so that, if corrected, they would prevent future incidents.

Stop rules are the result of many factors. Rasmussen (1990) notes that investigations stop when the analyst finds a sequence of events that matches a familiar prototype. That is, we stop when we find “causes” which we recognize. This means that the analyst’s own knowledge base and biases may well influence when the search for causal factors is terminated. Rasmussen (1990) sees three reasons to stop: where information is missing, where a familiar abnormal event is recognized as a reasonable explanation, or where a cure is available. This extends the biases to include the analyst’s knowledge of potential cures. Thus, if the analyst is, for example, a personnel manager, the opportunity to redesign physical equipment may not be recognized, so that data on equipment design may never even be collected. In all, there is an intimate relationship between the stop rules chosen and the investigator’s model of the processes of causation and responsibility for change.

There is yet another issue in determining investigation stop rules. Hindsight gives clarity and prominence to events in the incident causal sequence, which would not otherwise draw comment. Because we know that the final outcome was indeed an adverse event, as we look back down the lines of causation, factors, which in the past were considered normal or even benign, take on a sinister significance (Maurino et al 1995). The more factors we investigate, the more likely we are to find one, which is beyond what we consider to be “normal” conditions. The temptation is to see that factor as causal, even though it occurs many times by chance alone and does not cause an incident. If we investigate a set of factors $[F_i]$ where $1 \leq i \leq n$, then for each factor, there is a critical level X_i which would suggest it as a contributing cause, let the probability of factor F_i exceeding the critical level X_i be p_i just by pure chance, then, for n independent factors, the probability that at least one exceeds its critical level will be $P(n)$, where:

$$P(n) = 1 - \prod_{i=1}^n (1 - p_i)$$

Note that as n increases, $P(n)$ becomes nearer to 1.0. That is, the more factors you examine, the more likely you are to find one you consider “causal”. As a numerical example, if all of the p_i values are small, say 10%, then you would be 80% sure of finding at least one “causal” factor if you examined 16 factors, all of which were only operating by chance. It is to be expected then that different analysts may find quite different causal factors depending upon how diligently they search, even if these factors were not in fact causal.

2.2.3 Phase 3: Analysis

Analysis depends upon some model of the incident generation process. It attempts to find plausible sequences that fit the known facts. It is obvious from the discussion above that the investigator’s knowledge and bias (or model) not only affects the analysis but the earlier data collection. There have always been models to guide analysis and help, by implication, specify what data should be collected. Reason’s (1990) book uses a classification of errors into slips, lapses, mistakes and violations all based ultimately on an information processing model, such as those of Norman (1980) or Wickens (1994), Maurino et al (1995) proposes a more wide-ranging model covering latent failures as well as local or active failures. Feggetter (1982) proposes a model of the information processing levels with many similarities to that of Rasmussen (1990) and uses it to develop a human factors checklist for aircraft accident investigation. Even in the realm of injury prevention, Engkvist, Hagberg and Wigaeus-Hjelm (1995) use a model based on failure modes to investigate back injury causation with the aid of a checklist. In a post-accident study, Wiegmann and Shappel (1997) used several models of human error to successfully classify about 90% of the error events in a naval aviation accident database. An obvious example is [MEDA](#), which uses an implied model of the human and the system to define both a data collection and data analysis system.

Analysis and data collection is in fact two complementary parts of an iterative process. Data on an [AMT](#) not following a procedure will fit into one analyst’s model, for example as complacency, thus prompting further questions about the [AMT](#)'s state of mind and company norms. Another analyst starting from the same procedure not followed may fit the fact into a different model, such as the quality of the written procedure, leading to further investigations of procedure content and format. Note that the analysis phase can lead to data not being collected or even being selectively omitted from further consideration. Thus, analysis is on going and not just the formal casting of the data into a fault tree or other analysis tool.

2.2.4 Phase 4: Reporting

Like data collection and analysis, reporting involves a selection of facts as well as the formation of a coherent structure for the data considered. Thus, the analyst may have collected much data on the background of each participant, but may not report that data if it contains nothing the analyst deems important or relevant. Traditional texts on accident investigation (e.g. Ferry, 1981, Chapter 16) contain direct advice on reporting. The format of the cognizant authority ([OSHA](#), [FAA](#)) should be strictly followed (page 207-209). The five essential elements of the report are defined by Ferry as facts, analysis, conclusions, recommendations and summary (page 209). “*Facts should be presented in a logical sequence, stressing those which bear on the mishap process and cause of the mishap.*” The analysis section is “*a place to order and analyze the facts*” (all from Ferry, 1981, page 209).

It is in the reporting phase that management expectations and mandated formats can influence the outcome. Thus, if the person who will receive the report is perceived as expecting a short synopsis, that may well be what is produced. If the format is expected to be a report for a database, then only legal fields in the data base program will be filled in. In manufacturing industry, the forms used to report injuries are a notorious source of bias, and can even influence the data collection process itself. Typically, they place great weight on reporting the name and department of the injured person, on the nature of the injury sustained, and the outcome in terms of physician who treated the injury and expected length of disability. Data on events leading to the incident are often confined to a small text box, with a note to use additional sheets if necessary. Those who have to fill in such report forms almost never see such action as necessary. The result is a data set heavily biased towards the injured individual and against physical and situational factors. Hence, the form can literally determine the content, and lead to such traditional conclusions as the fact that almost all injuries are the fault of the worker.

The primary result of the reporting phase is a reduction in the amount of data made available. Unless the investigation is extensive and highly regulated (e.g. [NTSB](#)) then the raw notes and evidence are rarely included in the report. Thus, some active data reduction process is always taking place at the reporting phase. As an immediate example, in Marx’s study of causation, respondents were asked for the dominant contributor, the two next most influential contributors and any “root cause” of each scenario. This is a more drastic reduction of data than would normally be required, but it is often such salient features of an incident, which dominate the final report and the subsequent perceptions of future readers.

2.3 PATTERN OF FACTS DURING INVESTIGATION

From this description of incident investigation as a multi-phase activity, it is clear that the investigation itself is not free from bias and interpretation. We have many points in the process where the investigator’s own model of incident causation can have a major effect on what data is collected and what fraction of that is retained. This occurs even before we get to the point of attribution theory (Fiske and Taylor, 1984) affecting our interpretation of the results. In fact, a model of the investigation process in terms of total number of facts considered would initially increase through active investigation, then decrease through subsequent selection in later phases. A typical graph may be as shown in [Figure 1](#).

In the current study we will obtain objective measures of the first, second and fourth phases of the investigation process. The third stage of analysis is not measurable in the current study, as it is implied between data collection phase and the report phase. Our data will enable us to see how the investigation process initially expands from the few statements available in the initial scenario, and later contracts to reflect the selection of facts for the final report.

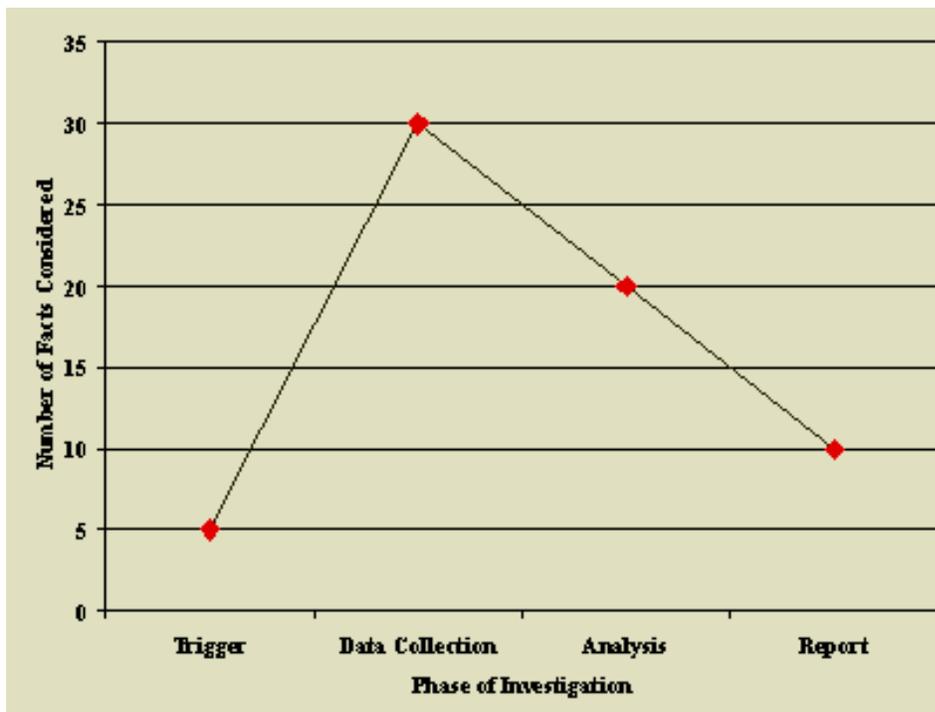


Figure 1. Expected increase and decrease in number of facts considered at each stage of investigation

3.0 METHODOLOGY

The first year's objectives are to develop the scenarios needed for the evaluation, and to undertake extensive benchmark testing at three sites. To develop a methodology of testing the effectiveness of human factors interventions the following logic is used:

1. Human factors requires understanding of human error.
2. Human factors interventions allow the participant (e.g. [AMT](#), supervisor) to learn the factors affecting human error.
3. A participant who knows more about the factors affecting human error will be better able to find the causal factors in an error investigation.
 4. We can measure a participant's ability to find causal factors in an error investigation by having them investigate a scenario and measuring what fraction of causal factors they find.
5. Our measure of fraction of causal factors found thus measures the participant's understanding of human factors.
6. We can use changes in the fraction of causal factors found to evaluate how well human factors interventions increase understanding of human factors.

The work of Woodcock and Smiley (1999) have reported a usable methodology, albeit in a different domain, and quote others who have used similar techniques in other domains. A short scenario is presented to each participant who must ask questions of the experimenter until satisfied that the incident has been satisfactorily investigated. Torell and Bremberg (1995) used a similar idea to study attribution of injury blame in home accidents. Within aviation maintenance the SUNY Buffalo team used a similar investigation methodology as part of the demonstration human factors training course for United Airlines in 1995. There, problems were summarized with a short paragraph, and trainees had to solve the problem using their human factors tools, and by asking the facilitator for more data. The fact that this worked well in a training context argues that a more complete methodology based on incident investigation should also be appropriate.

3.1 INCIDENT SCENARIO DEVELOPMENT

The objective of the project was to evaluate human factors interventions, such as [MRM](#) and [MEDA](#), by measuring how well participants investigate incidents before and after the intervention. To do this, we first had to develop suitable incidents for participants to investigate, and measure the performance of different groups of participants when asked to investigate these incidents.

The raw material for the incidents was chosen from existing incident reports at partner airlines. The criteria for choosing an incident were that it should have some human factors relevance, that it should not have a obvious single cause, and that it should be relatively easy to investigate in an interview of reasonable length. Six incidents were chosen and developed into scenarios, a term used to indicate that the description contains information on the people involved, the action, the equipment they use and their working environment. As it would be rare for a real incident to have information on all of the factors which people might want to investigate, the team had to fill in some blanks with plausible data, for example on the work background of participants or environmental conditions. This additional material was generated in conjunction with incident investigators and human factors personnel at our partner airline.

The six incidents were collected and subjected to analysis to determine what were the salient facts of the incident, whether or not these facts contributed to the incident. For example, the incident may have occurred outside so that lighting and inclement weather would be relevant facts, even though they may have been benign, e.g. a sunny spring day. At this point, airline partner personnel have again worked with us to ensure authenticity of the final scenario. For example, the original case may have included conflicting information that we have had to resolve in an appropriate manner.

Our analyses became time-lines which expanded into event trees so that all facts and events could be related to each other. We generated these event trees, but used more detailed fact tables for our interaction with participants. Each participant was given a brief synopsis of a scenario and used this as the basis for their investigation. The interviewer responded with the information requested. As each fact is requested, a code for that fact was recorded so that we could analyze the order in which facts were requested.

To ensure that each fact had a unique code associated with it, we developed a coding system both to help us to record interviews easily and later to analyze the data from these interviews. There are at least three sensible coding systems, each with points in its favor and against it. We can classify events or facts by Time Line, by Fact/Event Type or by Investigation Source. A combination of at least two schemes was needed for useful coverage. Time line has the advantage of being a natural structuring scheme for many of the people involved. Event type links directly to project objectives in that we need to know which aspects of the investigation receive most and least attention. Thus, we combined time line and event type to capture the order of information and its type. The source classification could later be mapped onto this scheme, often one-to-one but at times one-to-many, to provide an alternative interpretation of the data. In the end, the investigators themselves will structure the process; we could only anticipate these structures so that facts would be available readily during the experimenter / investigator interaction. One of our analyses will be to compare the order in which facts were requested with the order we would expect from these three classification schemes. This will reveal how investigators structure their task and perhaps how the structure changes with the

background of the investigator.

The final numbering scheme was based on the time line/event type classification.

3.2 NUMBERING SCHEME FOR INCIDENTS AND FACTS

Each fact was given a 5-digit code (e.g. 42123) so that it could be analyzed uniquely for whether it was requested by each participant, and in what order facts were requested.

- Digit 1 = Scenario Number
 1. Tagged Circuit Breakers on Passenger Entry Door
 2. Missing Cockpit Door
 3. Rudder Contact with Nose Dock
 4. Failure of Towbar
 5. Disconnected Pitot -Static Lines
 6. Rag in Horizontal Stabilizer Trim
- Digit 2 - 4 = Order of fact within scenario and main code
 - 100 = Major fact 1
 - 200 = Major fact 2
 - 300 = Major fact 3 etc.
 - 110 = Fact 1 contributing to major fact 1
 - 120 = Fact 2 contributing to major fact 1
 - 130 = Fact 3 contributing to major fact 1
 - 210 = Fact 1 contributing to major fact 2
 - 220 = Fact 2 contributing to major fact 2 etc.
 - 111 = Sub fact 1 contributing to Fact 1 contributing to major fact 1
 - 237 = Sub fact 7 contributing to Fact 3 contributing to major fact 2 etc.
- **Digit 5 = Main code for causal factor, using [TOMES](#) or [SHELL](#) models**

	TOMES	SHELL
1 =	Task	Software
2 =	Machine	Hardware
3 =	Environment	Environment
4 =	Operator	Liveware Individual
5 =	Social	Liveware Other

3.3 SCENARIO TRIGGER STATEMENTS

The following trigger statements for the six scenarios show the types of incident used. All have been de-identified with respect to airline source. In the opinion of the airline personnel who have been assisting us, they should be usable in both Part 121 and Part 145 operators.

1. Tagged Circuit Breakers on Passenger Entry Door

On Monday morning, it was discovered that the emergency power assist reservoir on the 1L door of aircraft 7026, a DC-10, was depleted. An operational check of the door revealed that it failed in the electrical mode. During troubleshooting, the circuit breakers were found to have been disconnected and tagged during maintenance at Benson's Air Service the previous month.

2. Missing Cockpit Door

During the preflight check on A/C #6833, Flight #1141, the crew found that there was no cockpit door in place. The cockpit door had been removed and not reinstalled during overnight maintenance to locate an under-floor leak.

3. Rudder Contact with Nose Dock

During heavy maintenance a B-747-200 was moved by mechanics from Bay 6 to Bay 5. During the move, the aircraft's upper rudder contacted the nose dock and sustained a 3 foot tear from the trailing edge forward.

4. Failure of Towbar

A DC-10, #5954, sustained an 8 foot gash to the fuselage when the towbar's lock opened during pushback from Gate 11 at EWR and the aircraft contacted the jetway.

5. Disconnected Pitot -Static Lines

Aircraft #1263 was about to begin Flight #371 from Tampa (TPA) to Washington DC (IAD). The pilot noted during taxi that the indications (including the A/C velocity and air temperature) from the #2 Air Data Computer were in error. On troubleshooting, a mechanic found that the pitot static lines had been disconnected from the #2 Air Data Computer.

6. Rag in Horizontal Stabilizer Trim

Upon landing flight 542 in Chicago (ORD), the flight crew reported that they had been unable to re-trim the horizontal stabilizer in either the electrical or manual modes. A mechanic in Chicago found that a rag had become jammed in the cable drum assembly, rendering the trim inoperable.

3.4 EXPERIMENTAL PROTOCOL

3.4.1 Subjects

Participants were recruited from available personnel at a variety of airlines and aircraft repair stations, and from professional aviation incident investigators. Each site had a limited number of people involved in incident investigation, so potential names were obtained from our airline contact and then each person was asked whether they agreed to participate. Four different participant groups were tested to determine whether there were differences between organizational levels and investigation experience within the organization as follows:

1. Aviation Maintenance Technicians ([AMTs](#)) with various levels of investigation experience

2. First-line supervisors and managers, both of maintenance and of Quality Assurance, again with various experience levels
3. Quality Assurance personnel, who are typically those charged with incident investigation
4. Professional aviation incident investigators not affiliated with a particular airline or repair station.

3.4.2 Interview Protocol

The interaction with each participant took place in an enclosed room, usually an office or meeting room, at their own worksite. They were provided with the written briefing and consent form shown in [Appendix 1](#). They are verbally assured of anonymity and asked for demographic data ([Appendix 1](#)). All data was stored by participant number only.

The data collection was in interview format, but was unusual in that the participant asked questions which were answered by the experimenter. The incident trigger paragraph was given to the participant (see [Section 2.3](#)). At this point, the participant was prompted to ask questions of the experimenter as if they were asking the same questions of personnel in the incident. Questions could cover the range from “What was the temperature in the repair hangar?” to “Did the mechanic have any personal problems that night?”

The experimenter answered the participant’s questions from the data sheets developed for each scenario. If the question was not anticipated, the experimenter replied that no data was available on that issue. This was typical of current practice, where nobody had thought to record, for example, the hangar temperature. Participants were given pad and paper to record facts if they desired.

When the participants declared that they would stop the investigation, they were asked to provide a verbal synopsis of the incident, as they would in writing a report. They were asked to list the contributing factors in their synopsis.

When they completed the assignment, participants were given a de-briefing. This was to remind them of the purpose of the experiment, to reassure them of anonymity, to provide non-specific encouragement about their performance (if asked), and to remind them not to discuss the specific scenario with colleagues who may become future participants.

3.5 ANALYSIS METHODS

As each participant’s interview was completed, the audio tape, experimenter’s interview notes and participant’s written notes were reviewed. This allowed a separation of the two parts of each interview: the data collection phase and report phase. Each phase was then analyzed in a somewhat different manner.

For the data collection phase, the objective was to determine the number and types of facts requested by the participant. Each fact was noted for the order in which it was requested. The first fact was coded “1”, the next “2” and so on. If a participant asked for the same information more than once (a quite common occurrence) only the first request was coded. These data were placed into a MINITAB™ database. Later, each order data was re-coded as either a “0” for a fact not requested or a “1” for a fact requested. This allowed us to count the number and type of facts requested. The number and percentage of facts requested for each scenario were the primary measure of Phase 1 data collected.

For the report phase, a transcript of each report was made. This allowed the analysis team to make an unambiguous list of the facts incorporated into the final report produced by the participant. The order of these facts was not particularly relevant, as most reports gave the facts in time order rather than the most salient facts early in their list. From the list produced, the total number and percentage of facts were both measured.

The primary statistical analysis was of the following dependent variables:

Data collection phase: Number and Percent of total facts requested

Number and Percent of each classification of fact requested

Report Phase Number and Percent of total facts requested

Number and Percent of each classification of fact requested

The data were each analyzed by scenario number (1-6) and participant's position in the organization ([QA](#), supervision, [AMT](#), professional investigator). Subsidiary variables such as years of experience, organization, human factors training etc. were treated as covariates. The main analysis tool was a fixed effects analysis of variance using the General Linear Models (GLM) procedure from the MINITAB™ statistical analysis package.

In order to ensure that the transcription from the raw tapes and notes to data table was reliable, an inter-rater reliability analysis was conducted. Two analysts independently analyzed each of three scenarios and recorded their data on the facts requested and the order in which they were requested. The data from the two analysts was compared on two measures:

1. Overall match at the item level, i.e. which fact numbers agreed.
2. Match of fact type (TOMES) between the two analysts.

The first inter-rater reliability analysis gave rather poor results. The average overall agreement was 64.6% while the TOMES agreement was 71.7%. At this point a set of written protocols for data analysis was compiled and used by both analysts. The protocols dealt with the complexities of the data, e.g. how to handle a participant who asked for multiple facts in one question, or how to deal with the same fact being requested several times. (Separate analyses of these complex factors are given in [Section 4.3](#) after the main data analysis.)

After implementing the protocols, agreement was greatly improved. The average overall agreement was now 88.6% while the TOMES agreement was now 95.4%. The change in overall agreement was statistically significant (Paired t-test: $t(2) = 4.83$, $p = 0.040$), while that for TOMES agreement approached significance (Paired t-test: $t(2) = 3.73$, $p = 0.065$). The protocols clearly made a considerable improvement in data reliability, which was finally over 95% for the data of most interest (TOMES).

4.0 RESULTS

In this section, we present the results in a logical sequence, starting with overall analyses in [Section 4.1](#). First, we determine whether or not our sample of participants was representative. Next, we examine the question of individual differences between the participants to see whether variables such as age or experience affect the results. When it found that they generally have no effect, the analysis proceeds with overall analyses of variance on the main measures of number and percent of facts investigated. This is extended to an analysis of the types of fact investigated.

In [Section 4.2](#) the detailed investigation procedures are studied. Just what facts are investigated and in what order? What are the criteria for stopping the investigation? What is included in the synopsis as important contributing factors? [Section 4.3](#) covers the complexity factors.

4.1 OVERALL ANALYSIS

4.1.1 Was the Sample representative?

Our sample of 37 people had 5 [AMTs](#), 3 supervisors/managers, 21 [QA](#) investigators and 10 professional investigators, primarily because those were the people made available by our airline and other partners. Thus, the sample is heavily weighted towards [QA](#), who in fact perform the bulk of incident investigations in the industry. The participants were classified on Human Factors training by recording whether they had taken, or helped teach, an [MRM](#) course. Twelve of the sample (32%) had [MRM](#) training. Similarly, to measure the investigation training of interest, we recorded that 10 participants (27%) had had [MEDA](#) training.

We can also compare the age and experience distributions to the population demographics of [AMTs](#) found in a national sample compiled by the Bureau of Labor Statistics ([BLS](#), Washington, 1991). Our sample was significantly older with a median age of 43.5 year versus a [BLS](#) median age of 36.2 years (Wilcoxon test, $t = 594$, $p < 0.001$). Our sample, except for the professional investigators who were not AMT trained, were also more experienced with a median of 17.5 years as an [AMT](#) versus a BLS median of 9.4 years (Wilcoxon test, $t = 353$, $p < 0.001$). When the professional investigators are included, the difference is no longer significant (Wilcoxon test, $t = 426$, $p = 0.146$). The age and experience distributions of our sample of participants are shown in [Figures 2](#) and [3](#).

[ANOVAs](#) were performed to check whether the characteristics of participants differed between the three job types. There were significant differences on two of the measures. Years in Current Job showed that [AMTs](#) had spent much longer in their current job (19.7 years) than managers (1.5 years) or [QA](#) investigators (4.3 years) or professional investigators (3.4 years), ($F(3,32) = 10.6$, $p < 0.001$). For years since obtaining an A&P License there was a significant difference ($F(3,32) = 13.9$, $p < 0.001$), where [AMTs](#) averaged 29.3 years, while managers averaged 13.7 years, [QA](#) investigators 16.7 years and professional investigators 0.7 years. Other summary statistics for the overall sample included a median of 2.8 years investigating incidents, a median of 12.5 investigations in the past year and a median of 1 day since the previous investigation.

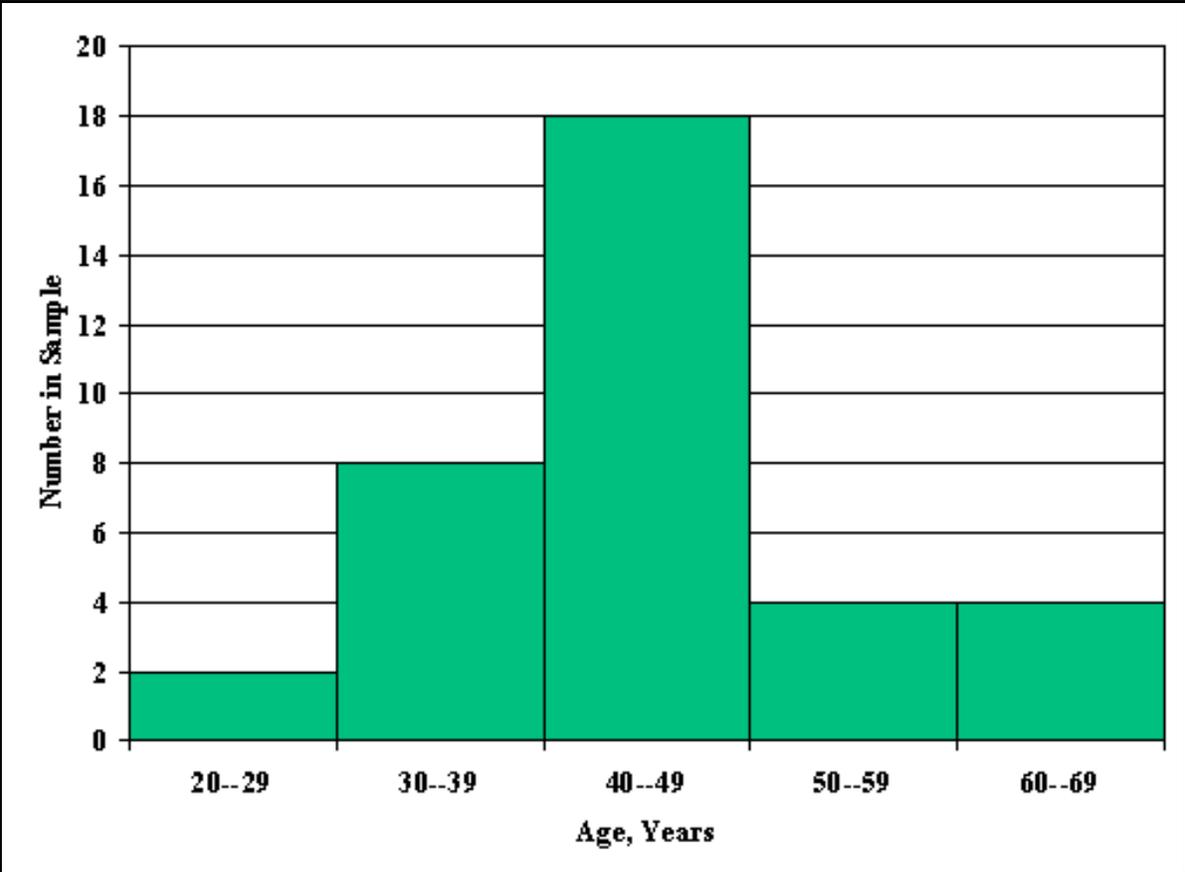


Figure 2. Age distribution of sample

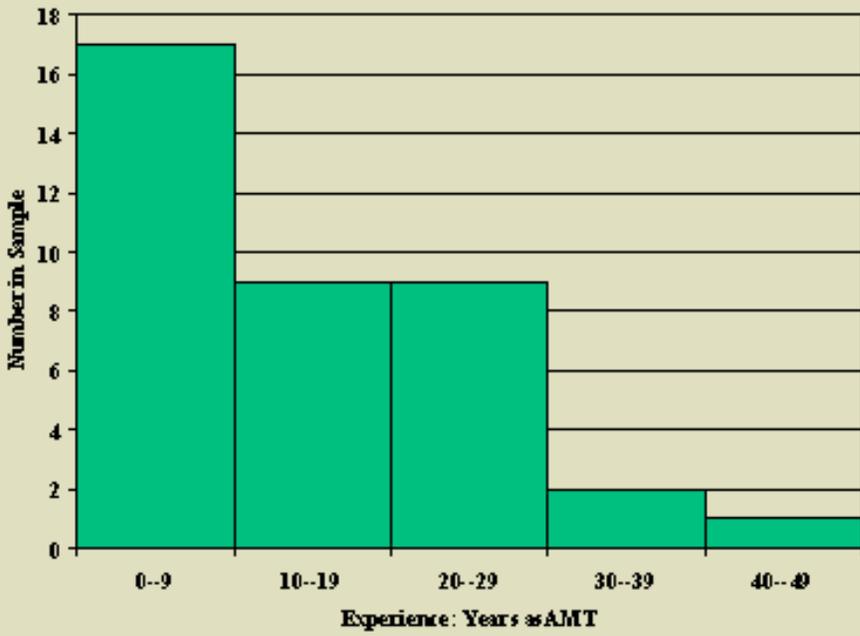


Figure 3. Experience distribution of sample

The conclusion is that our sample was older and more experienced than a sample of [AMTs](#), primarily because most had left the hanger floor for staff positions in [QA](#). The sample perhaps had more training in [MRM](#) and [MEDA](#) than most [AMT](#) samples, but again these levels would not be unusual for [QA](#) investigators. Managers and [QA](#) investigators change their job titles more frequently than [AMTs](#), but all were well experienced in the industry and in incident investigation. The professional investigators had typically not had either AMT training, MEDA training or MRM training as such, although they had had proprietary incident investigation training and some human factors exposure.

4.1.2 Individual Differences

An initial inter-correlation between the individual difference measures (Age, Years as [AMT](#), Years in Current Job, Years Investigating Incidents, Number of Incidents Investigated in the Past Year, Time Since Last Investigation) showed that these measures were only rarely related to performance of the investigation. Factor Analyses of the individual difference measures gave four factors, while three factors were derived for the overall performance measures. These were Investigation Depth, comprising number and percent of facts asked for; Synopsis Depth, comprising number and percent of facts in the synopsis, and Time Spent, comprising time to investigate and time to give the synopsis. When the three performance measures were correlated with the four individual difference factors, no significant correlation was found among the twelve possible correlations. On the basis of this analysis, it was concluded that demographic factors played little if any role, so that none of these individual difference measures were included in further analyses.

4.1.3 Total Facts Investigated and Reported, and Times Taken

With 37 participants unequally distributed across job types, no full factorial analysis of the facts investigated and time taken using both Scenario and Job Type as factors was possible. Thus, the GLM procedure was used to give analyses of variance (ANOVAs) with Scenario as the main variable and Job Type as a covariate. The difference between the six scenarios in the number of facts investigated just failed to reach significance ($F(5,29) = 2.42, p = 0.060$). [Figure 4](#) shows the mean number of facts investigated in each scenario. There were no significant differences between the six scenarios by percent of facts investigated or time to investigate, which averaged 26.7 minutes. In terms of synopses, there were significant differences between the six scenarios for both number and percent of facts included in the synopses ($F(5,29) = 4.89, p = 0.002$, and $F(5,29) = 8.88, p < 0.001$ respectively). Again, there were no differences between scenarios in times taken for this part of the task. It took an average of 4.8 minutes for participants to summarize their findings in the synopsis. The mean numbers of facts in each synopsis are shown for the six scenarios in [Figure 5](#).

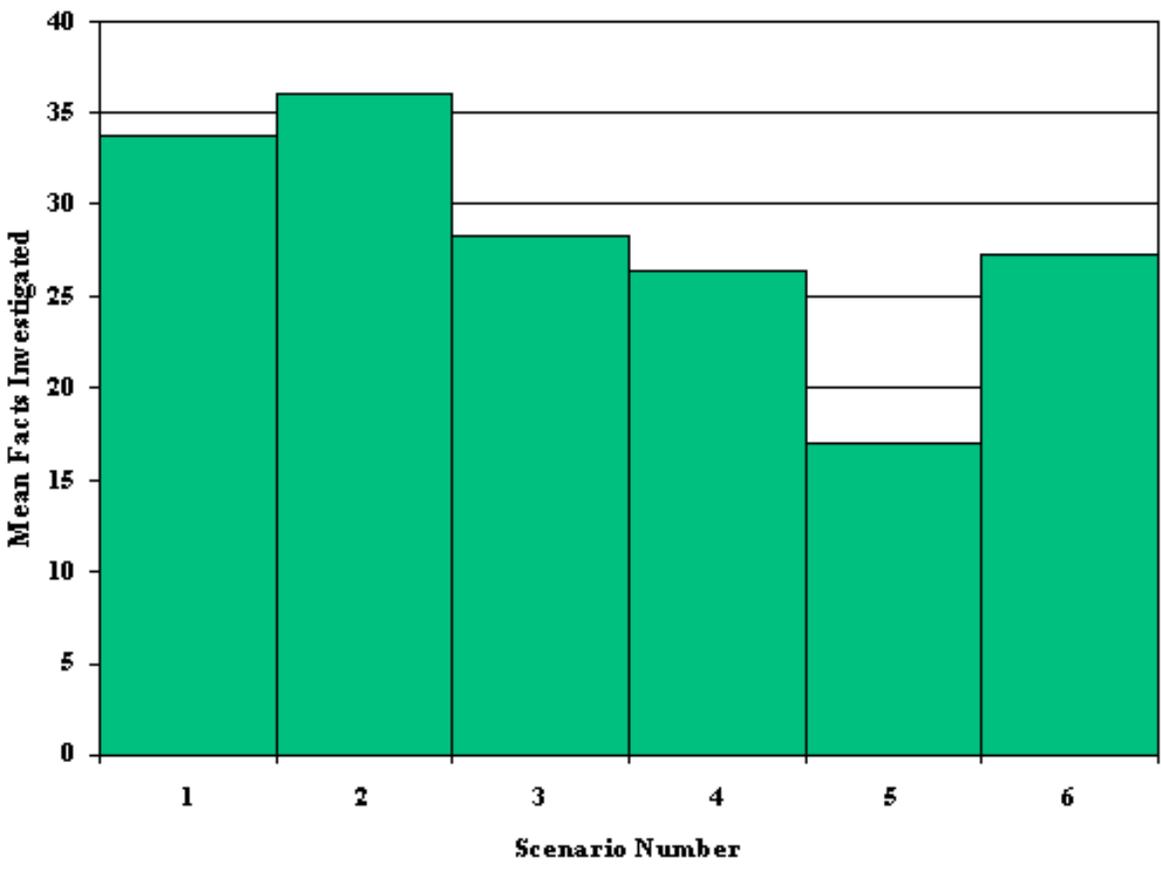


Figure 4. Number of facts investigated by scenario

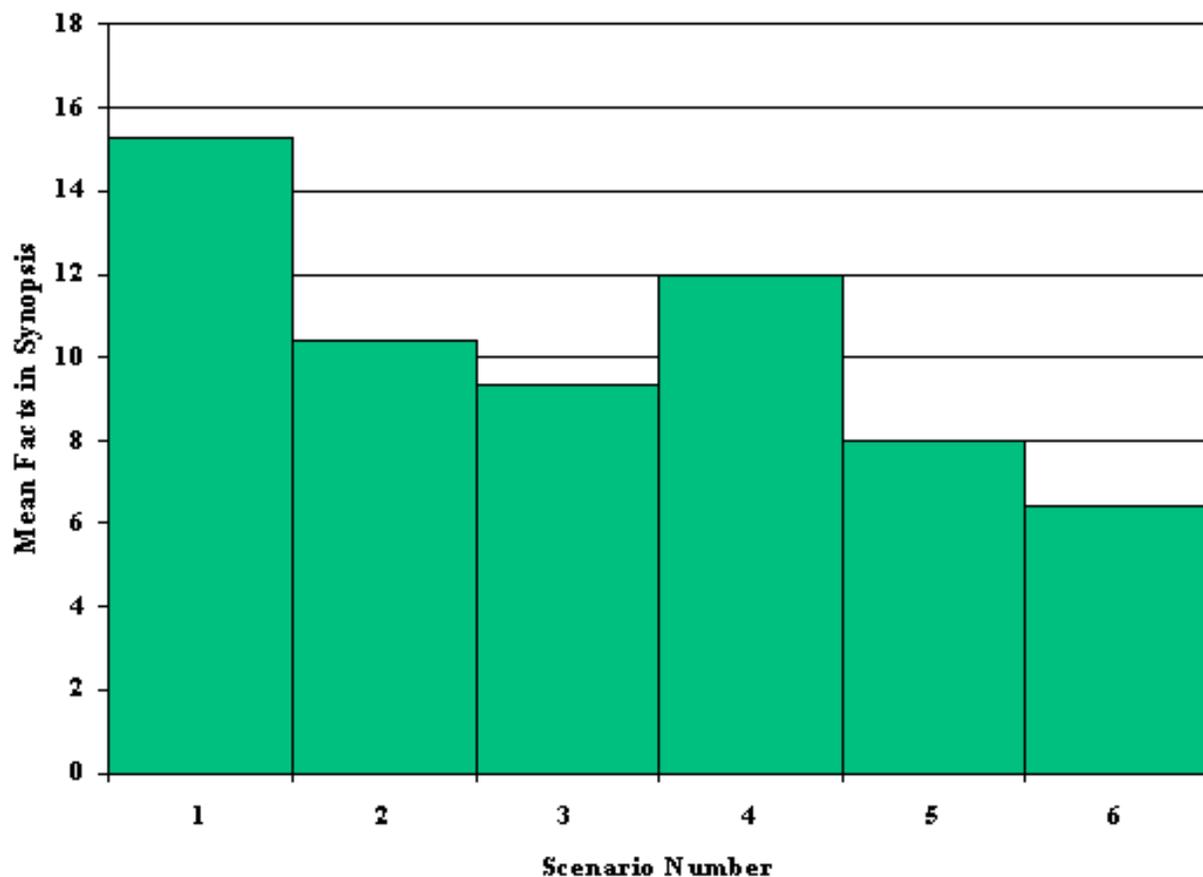


Figure 5. Number of facts included in each synopsis

From these analyses, we can see that it is more typically the number of facts rather than the percentage of facts which gives differences between the scenarios. [Figure 6](#) shows the correlation of 0.82 between total facts in each scenario and number investigated. This correlation was statistically significant, $p = 0.047$). If the percentage is relatively constant across scenarios, this implies that people investigate at a depth proportional to the total size of the scenario. More complex scenarios elicit more fact finding and more conclusions in the synopsis, although not more time spent investigating. Thus, measures such as the percent of facts investigated and reported can to some extent equalize results across scenarios. In the remaining analyses, where the objective is to characterize the scenarios for future use, only numbers of facts will be analyzed, not percent of total facts.

The next logical analysis is to break down the facts investigated by the type of fact, i.e. whether the facts refer to Task, Operator, Machine, Environment or Social factors. To do this, we make use of the classification system given in [Section 3.2](#). Because not all fact types were represented in all scenarios, there can be a lack of balance in this analysis, but overall it can show what were the facts receiving most and least investigative effort from participants.

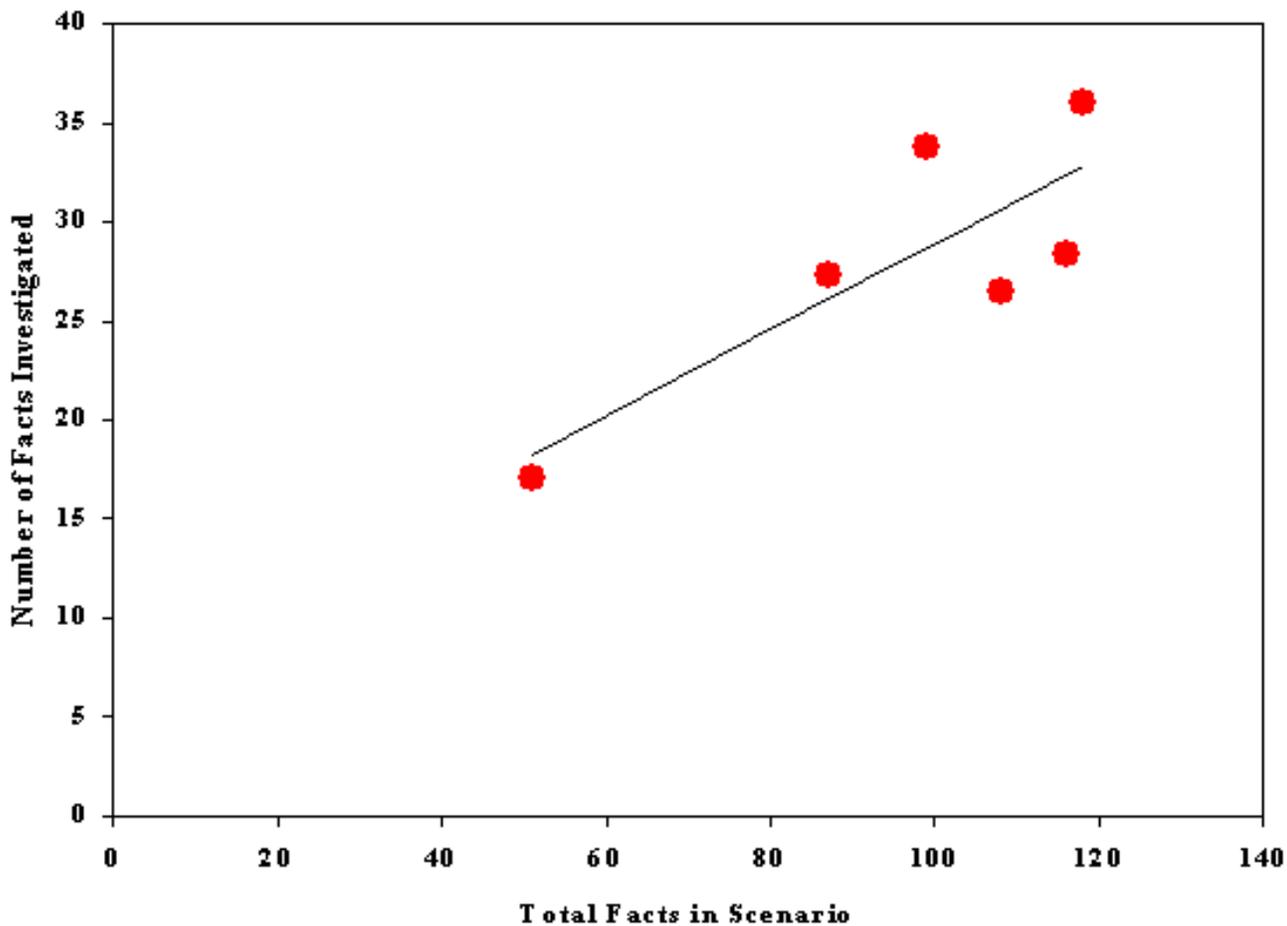


Figure 6. Correlation between total facts and facts investigated

The number and percent of each fact type (T/O/M/E/S) were counted and analysed using ANOVAs with scenario type, [TOMES](#) type and their interaction as factors, and Job Type as a covariate. As noted above, only the Number of Facts analysis is reported. Both main effects and the interaction were significant ($F(4, 149) = 85.8, p < 0.001$ for [TOMES](#), $F(5, 149) = 4.5, p = 0.001$ for Scenario and $F(20, 149) = 8.1, p < 0.001$ for the interaction), as was the covariate of Job Type ($F(1, 149) = 4.49, p = 0.036$). The pattern of facts by fact type is given for each scenario in [Figure 7](#). Note that most of the facts sought by participants were Task facts as these established what activities were performed in what order in the scenario. Machine facts were not investigated in great numbers, except for Scenario 4 where a piece of equipment (the towbar) failed, a major cause of the incident. Social factors were investigated strongly in Scenarios 2 and 3 where multi-person cooperation (or lack of it) was a key issue. Environment factors were of little interest to participants except in Scenario 4 where the tow bar failed in darkness at a gate with a tight space for maneuvering. Operator facts were investigated consistently across scenarios. The background of the scenario characters was examined to help determine why they might have acted as they did.

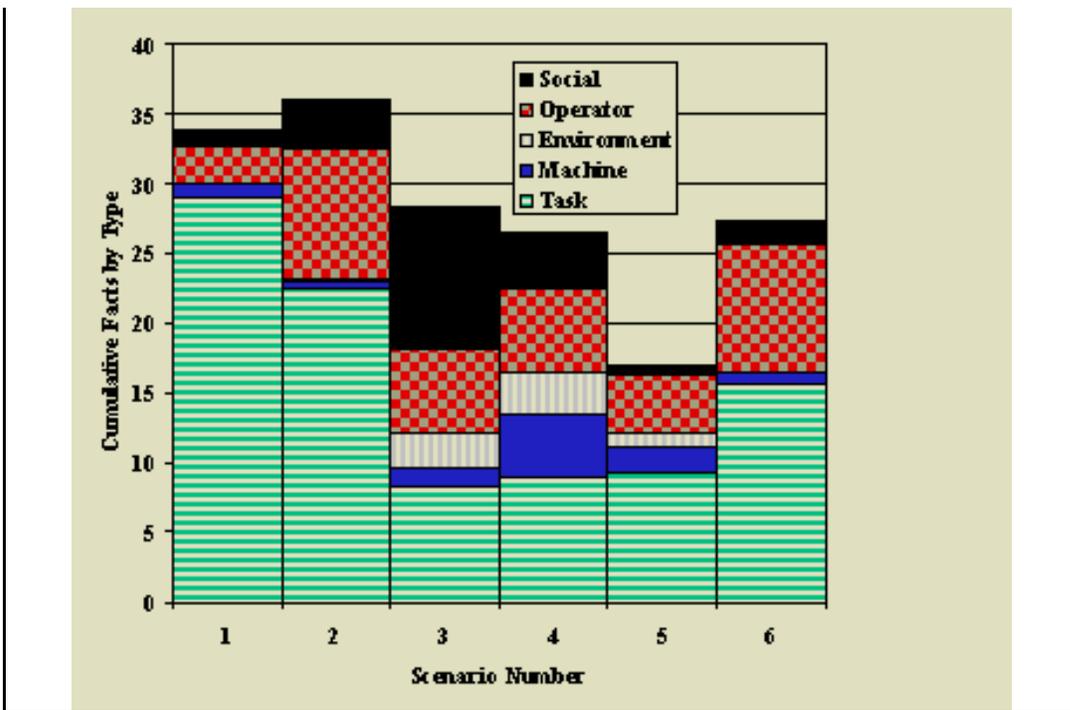


Figure 7. Types of facts investigated for each scenario

In this analysis the covariate of Job Type was significant but the size of the effect was small except for professional investigators. [AMTs](#) investigated an average of 5.2 facts of each type, while managers investigated 5.3, [QA](#) investigators 5.4 and professional investigators 6.4. Further analysis of the data with a separate ANOVA for each fact type (TOMES) showed no significant differential patterns by Job Type.

The same analyses were performed on the facts included in the synopsis to determine which aspects of the investigation were retained from Phases 2 and 3 through to Phase 4. This showed again that the scenarios were significantly different in the number of facts of each type included. The effects of Scenario ($F(5, 149) = 5.9, p < 0.001$), [TOMES](#) fact type ($F(4, 147) = 50.0, p < 0.001$) and their interaction ($F(20, 149) = 14.7, p < 0.001$) were all significant while the covariate did not reach significance. The pattern of synopsis facts is shown in [Figure 8](#).



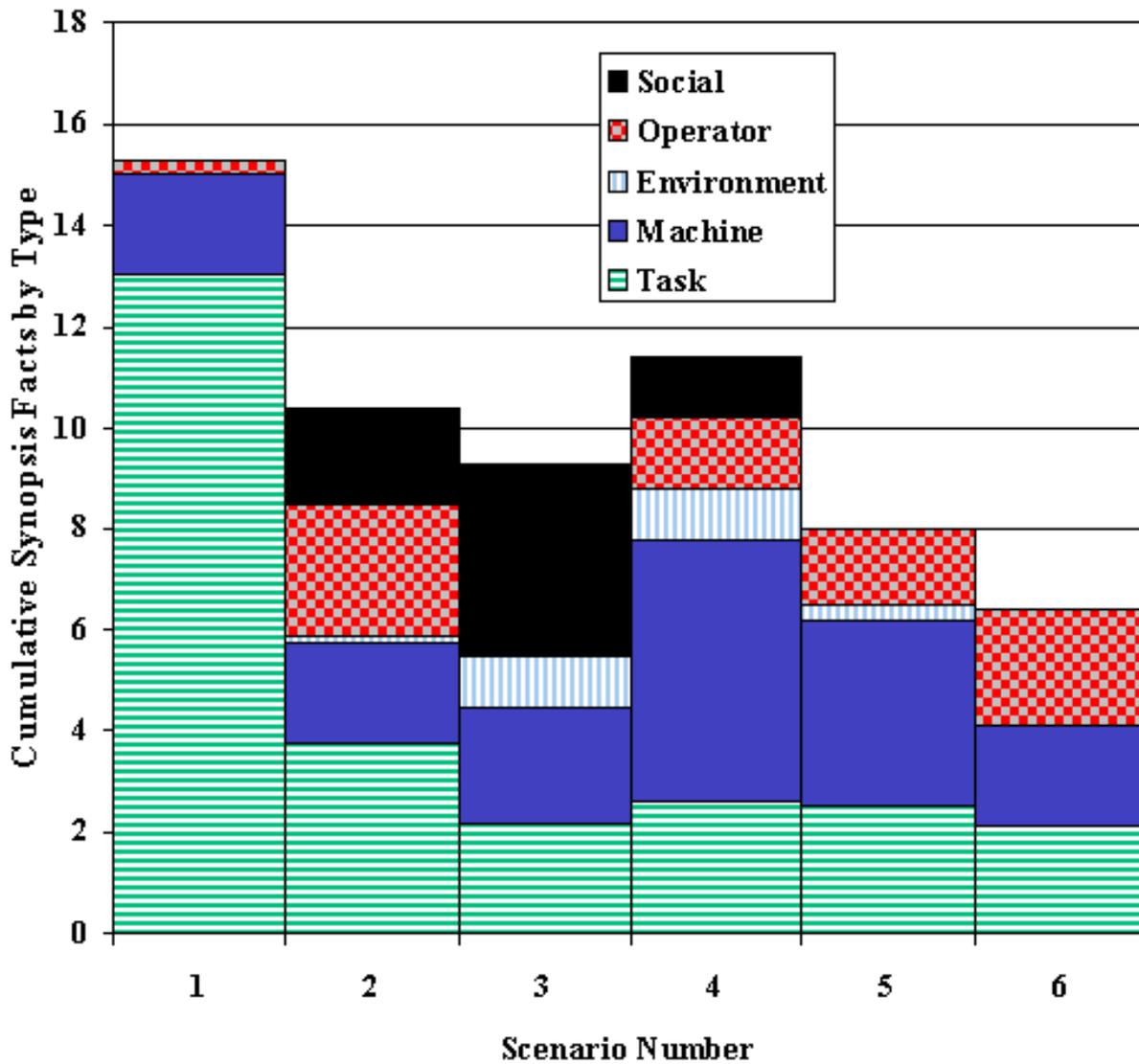


Figure 8. Types of facts included in the synopses for each scenario

Perhaps the most striking feature of [Figure 8](#) is its similarity of pattern to [Figure 7](#). Indeed, the number of facts investigated by fact type correlates quite well with the number of facts of each type in the synopses ($r = 0.72$, $p = 0.16$). [Figure 9](#) shows this correlation with the type of fact labelled on each point. Generally, if a type of fact is favored in the investigation, it is also favored in the synopsis. Participants may select facts to include in their summary of an incident, but they do not appear to selectively omit one type of fact rather than another. The odd point is for Machine facts, which are included in the synopsis at a higher rate than in the data collection. This was because most participants used machine facts from the trigger to provide context for the synopsis, for example by stating aircraft type, tail number and location of defect.

There were, however, some differences between the pattern of facts across scenarios between Phases 2 and 3 and Phase 4 as can be seen in [Figures 7](#) and [8](#). Comparing these two figures, the increased representation of Machine facts is apparent throughout. Beyond this, Task facts are under-represented in the synopses of Scenarios 2 and 6. The large contribution of Social facts in scenarios 2, 3 and 4 occurs in both the investigation and the synopsis.

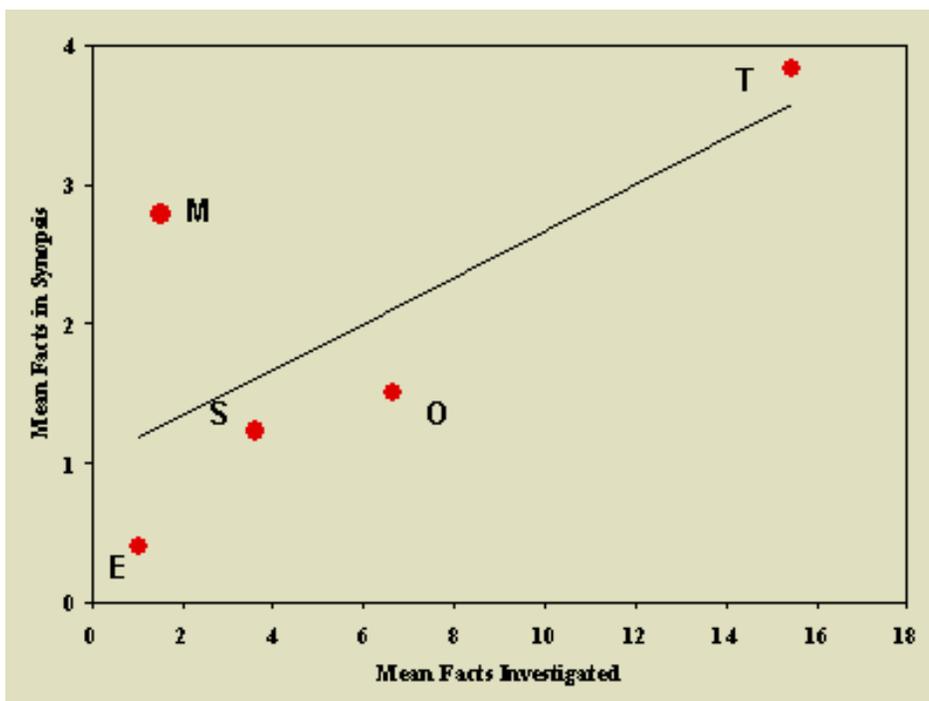


Figure 9. Correlation between types of facts included in the synopsis and investigated

4.2 ANALYSIS OF ORDER OF FACTS INVESTIGATED

For this analysis, the original files listing all facts were used. These files were organized in approximately chronological order. Thus any positive correlation between the order of listing and the order of asking for each fact was a direct test of whether the participant investigated that scenario in chronological order. This correlation was performed for all 37 participants, and the correlation coefficients used in an [ANOVA](#) to test for differences between the scenarios. There were no significant differences in that analysis, so all 37 correlations were taken as a single group. [Figure 10](#) shows the distribution of these correlations, which can range from -1 for the opposite of chronological order, through 0 for random order, to $+1$ for perfect chronological order. As can be seen, most were positive and many were quite large. Indeed, of the 37 participants, 17 were significantly positive and only 2 significantly negative. A t-test of the distribution in [Figure 10](#) shows that the mean (0.31) of the distribution was positive ($t = 4.7, p = 0.001$).

The next analysis was to determine which facts were asked for most often in each scenario, and when participants sought them. To do this, we reordered the facts in each scenario by the number of times each was asked for, and by the investigation order for each participant. Note that in this report, we are not presenting extensive examples of the facts themselves as if these become widely known in the industry, the scenarios will be compromised for future years of the project. Hence, we will classify the facts and use this classification to present the results of the analysis.

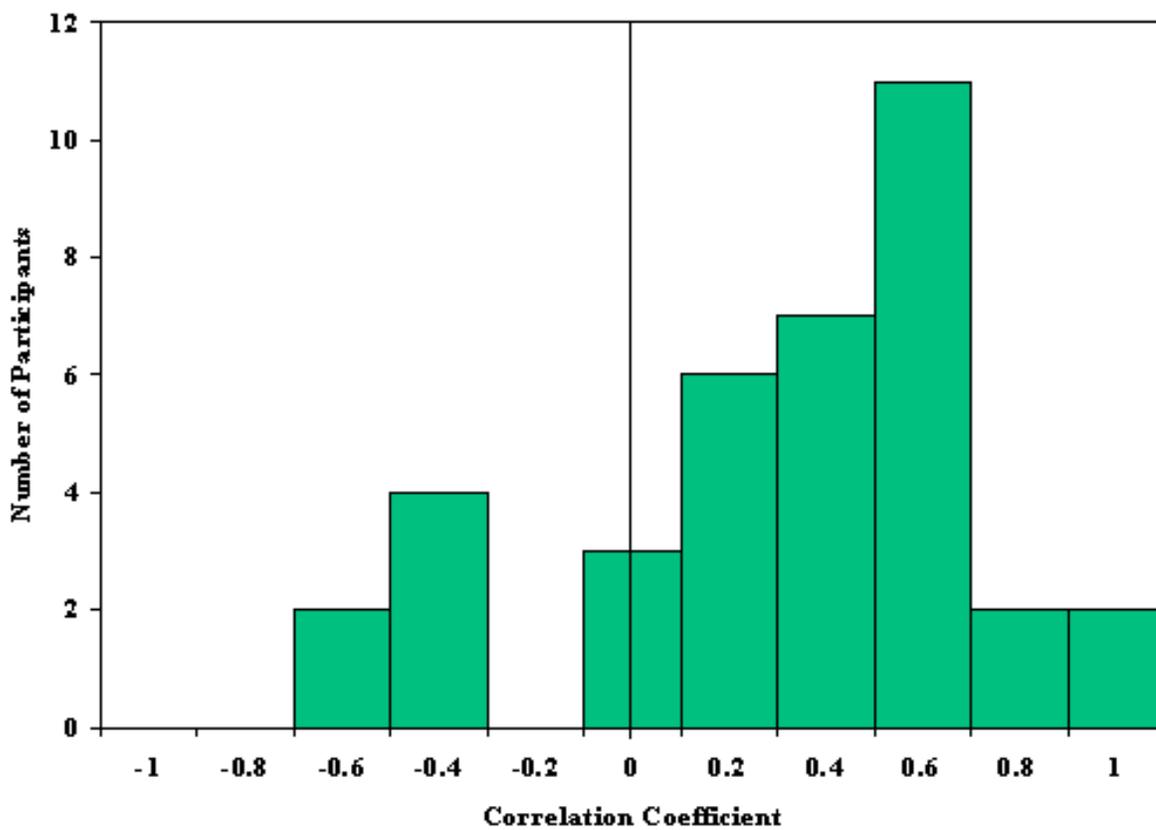


Figure 10. Distribution of correlations between investigation order and chronological order

What we are doing is analyzing verbal data (the facts) to look for patterns that repeat. The sequence of each investigation is typically quite messy in detail, but when enough investigations were ordered and compared, patterns did emerge. It is possible to perform strict pattern detection, for example by creating a to/from matrix that can be used to discover repeating instances of one class of fact being followed by another. Such an analysis has not been performed at this stage as the research is still exploratory to some extent. What follows is an initial classification based on ordering of the scenarios.

In addition to facts having types (e.g. [TOMES](#) or [SHELL](#)), we can use other attributes of each fact appropriate to the study of the order of investigation. Phases 2 and 3 of an incident investigation cannot be separated clearly in many investigations, and certainly not in the current study. Participants were performing analysis (Phase 3) throughout the data collection (Phase 2), and using these analyses to guide further information gathering. But for the process we recorded, it was clear that the first task in the investigation was to determine the temporal boundaries and initial characters. This stage comprised four types of information.

4.2.1 Discovery

How was the error or incident brought to public scrutiny? This was often given in the trigger paragraph (e.g. “During the preflight check on A/C #6833, Flight #1141, the crew found that there was no cockpit door in place.” from Scenario 2) and so was sometimes absent from the investigation. The discovery events determine the end point of the period of time being investigated.

4.2.2 Operational Trigger

What was the maintenance or operational events that started the sequence of events leading to the incident (e.g. “The cockpit door had been removed and not reinstalled during overnight maintenance to locate an under-floor leak” from Scenario 2)? This determined the early boundary to the time being investigated.

4.2.3 Initial Actors

Who were the characters in the scenario performing at least the tasks subsequent to the operational trigger? For Scenario 2, the following facts were classified as Initial Actors:

22104	10/26 0740	Adam Westcliff assigned to NRR
22154	10/26 0740	Martin Highland assigned to NRR

(Note that in these excerpts from the scenarios, the first column gives the fact number (see [Section 3.2](#)) while the second gives the time of the event. Some facts, such as those giving the demographic background of the characters or typical policies in the hangar, do not usually have associated times but are given times in this report to avoid the appearance of empty cells in a table.)

In some instances another class of fact was included in the initial stage of the investigation as it helped determine the temporal sequencing of events:

4.2.4 Return to Service (RTS) Decision

Who signed the RTS and when? This is an important part of the checking procedure for any aircraft as it represents the point at which the aircraft is officially deemed airworthy.

With these facts in place the rest of the investigation could proceed in somewhat of a temporal order. There were three classes of facts in this main stage of the investigation:

4.2.5 Work Sequence

These are primarily task events and define the steps accomplished by the characters, both those defined by the established work procedures (e.g. workcards) and those taken in response to problems, such as [AMTs](#) being assigned to other work for lack of parts to complete a workcard. Within the work sequence are both physical acts on and around the aircraft (e.g. removing a forward galley) and paperwork / reporting actions (e.g. signing off for the removal of the forward galley).

4.2.6 Inspection Sequence

Although these are items of work, they are broken out as a class because they are performed by different characters in the scenario and because they typically occur after the main work sequence. A typical inspection sequence fact would be:

27421	10/28 0400	Inspector Harry Bird checks briefly inside Aircraft #6833 and sees important systems OK
-------	------------	-----------------------------------------------------------------------------------------

4.2.7 Contributing Factor

These are facts may be in the sequence of tasks (e.g. [AMT](#) does not complete shift change log). They also include parallel details from our classes of Machine, Operator, Environment and Social, such as the following:

22775	10/26 1430	Millennium Air does not enforce shift turnover log policy, which is often ignored by AMTs and Inspectors but followed by leads and supervisors
-------	------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------

A further possible class of facts that were requested by some participants were:

4.2.8 Post Discovery

Facts that establish the actions after discovery. These were not the focus of our study, but a few were included, such as:

28311	10/28 0730	Cockpit crew notifies hangar manager that cockpit door is missing
-------	------------	-------------------------------------------------------------------

With these classes of fact, it is possible to see how the investigation proceeds. [Figure 11](#) shows a schematic that summarizes our findings on this study, with the stopping rule explicitly added to show that all investigations have to cease at some point.

In the Boundaries Stage, the investigator determines the starting and ending points of the investigation, and who were the major characters involved early in the maintenance process. These boundaries may be exceeded in later stages, for example when the investigators asks about prior training of [AMTs](#), or prior problems with a particular system on a particular aircraft fleet. The same consideration can apply to post-discovery facts. However, the main purpose of this stage is served by gathering facts on the discovery, the operational trigger and the initial actors. Within this stage, the ordering of these three classes of fact is not specified. Also, other facts are interspersed within these classes, for example to follow up on the background of actors or the operational trigger. Thus, the Boundaries Stage merges into the Sequence Stage with no sharp division.



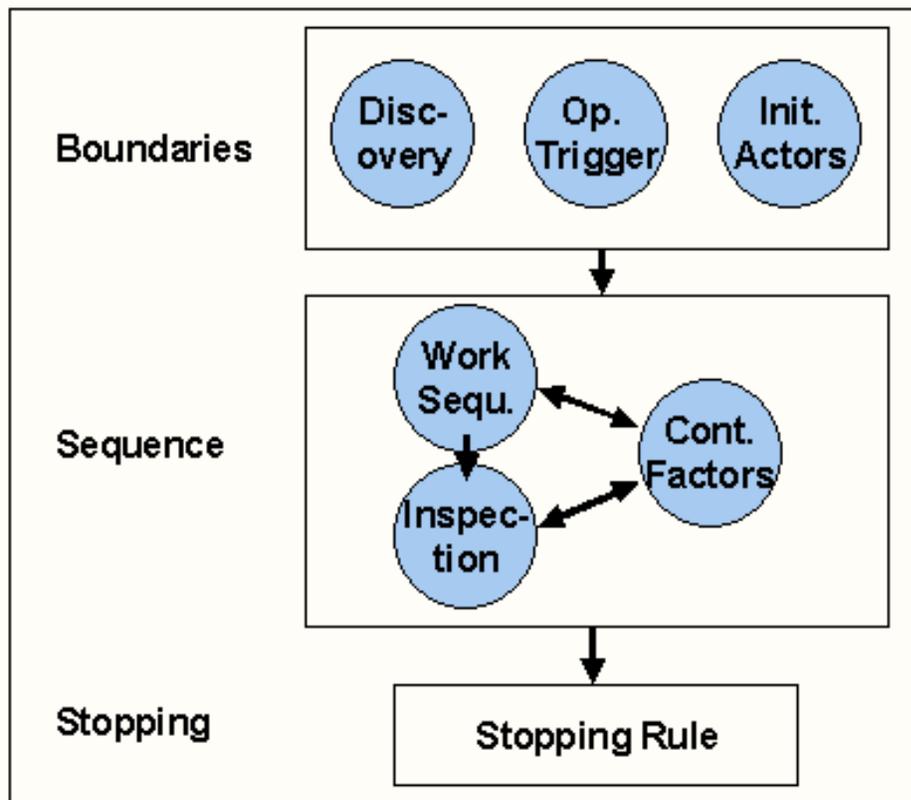


Figure 11. Three stages of investigation sequence in Phases 2 and 3

Within the sequence stage, the investigation usually proceeds from the work sequence, which is the actions primarily performed by [AMTs](#), utilities, managers and supervisors, to the inspection facts that are performed by inspectors before the work package can be completed and the aircraft returned to service. Not all investigations followed such time line, as was demonstrated in [Figure 10](#). Participants left out logical events, returned to events out of sequence, backtracked when a new fact did not fit their expectations or when an inspection item was followed physically by another work item. At any time during the sequence, participants would ask for contributing factors before continuing with the sequence. For example, in Scenario 2, one participant used the following sequence of a work sequence facts followed by a social fact and then by an inspection sequence fact:

22331	10/26 0800	Adam Westcliff and Martin Highland disassemble coat closet
22425	10/26 0800	Norm in hanger is to remove it for easier floor access
27421	10/28 0400	Harry Bird checks briefly inside #6833 and sees important systems OK

Hence, in [Figure 11](#), the main order is shown as **Work Sequence** to **Inspection Sequence**, with side arrows to **Contributing Factors**. The impression from the investigations was one of a combination of logical analysis, working forward and backward from the limits found during the Boundary Stage, with contributing factors requested either as a natural occurrence in this process, or in a more random fashion. One point to note is that many of the questions concerned the paperwork system. Participants tried to establish the paperwork trail often concentrating on who signed off on particular workcards, whether the work instructions were from the Maintenance Manual, a workcard or a non-routine repair card ([NRR](#)), and whether specific actions taken by characters were those defined by the written procedures.

The final stage depicted on [Figure 11](#) is the Stopping Rule. We did not explicitly ask participants why they stopped, but some conclusions can be inferred from their behavior in the investigation and in the synopsis. A few participants did mention why they stopped. One stated that he would stop at a particular point if this was a normal investigation, for example requested by the manager or supervisor to determine what caused a delay. However, he would go further (and did in the interview) if he wanted to “get to the bottom” of the incident. Another stated that “I could continue but I'd just be beating a dead horse because I think I know what happened.” Yet another claimed that he had determined the causes from the synopsis, and only asked a few questions to confirm these conclusions. For a few participants, statements in their synopses closely resembled facts that they had not originally asked for. This suggests that they were inferring facts in their reporting rather than strictly selecting facts from an explicitly-gathered store.

However, most of the participants stopped after finding what they considered to be a final key fact which “explained” the incident to their satisfaction. Interestingly, nobody stopped at a single “cause”, perhaps because the scenarios were selected to encompass multiple causes, or because investigators know that incidents are typically multi-causal. The final key fact was of course different between participants as they discovered their facts in a different order, and nobody exhausted all of the facts. Indeed, across the six scenarios the mean percent of facts requested was 33%, with a range from 20% to 42%. Across all individual participants, the lowest was 9% of facts and the highest 59%.

When participants stopped investigating, they summarized their findings, including a much smaller set of facts in their synopses. Across the whole data set, participants used only an average of 8% of facts, with a range across scenarios from 3% to 15%. Thus, they retained about a quarter of what they had discovered into Phase 4 of the investigation process. Many of the facts retained were again concerned with the paperwork trail: signoffs, use of manuals, etc. There were other contributing factors from all of the fact types, but mainly concentrating on Task Operator and Social types as shown in [Figure 8](#). The participants seemed to be trying to determine whether any violation of written procedures had occurred. They used the other contributing factors to help “explain” why some of these deviations from correct procedures occurred.

4.3 ANALYSIS OF COMPLEXITY FACTORS

As noted in [Section 3.5](#) not all responses from the participants were in the form of a unique question giving a unique fact. The following were measured to better define the complexity of the interview process:

1. Errors made by the experimenters in responding to the participants. As only 7 such errors were made in the whole study, these are not considered further.
2. Participants requesting multiple facts. We measured both how many times this occurred and how many facts were released in response to such questions.
3. Repetitions of questions which elicited the same fact. The number of such repetitions was counted. Only the first request for a fact was recorded in the main data.
4. Extra facts requested that were not in the scenarios. We counted the number of each type of extra facts (TOMES) as well as the total number requested.

Measures 2-4 were analyzed using fixed effects General Linear Models ANOVAs with Scenario as the factor and job type as the covariate.

The number of questions eliciting multiple facts and the total facts released in this way both gave significant differences between scenarios (Questions: $F(5,29) = 4.4$, $p = 0.004$; Total Facts: $F(5,29) = 3.84$, $p = 0.009$) and between job types (Questions: $F(29) = 4.37$, $p = 0.045$; Total Facts: $F(1,29) = 5.83$, $p = 0.022$). [Figure 12](#) shows the relationship between these values for each scenario. The correlation of 0.977 was significant at $p < 0.001$. Note that each multiple question elicited 2.67 facts on average. The Job Types effect showed that the professional investigators asked the most multiple questions (4.7) and received the most multiple facts (13.7).

Repetitions were analyzed in the same way, but only yielded a significant scenario effect ($F(5,29) = 3.84$, $p = 0.009$). Scenario 1 had the most repetitions (13.5 per respondent) against an overall average of 5.1. Participants sought considerable reassurance from the experimenters.

Extra facts were analyzed by scenario and fact type with job type again as a covariate. Only fact type was significant ($F(4,149) = 28.4$, $p < 0.001$). The main extra fact type requested was Task (4.4 times) followed by Machine (1.6 times), Operator (1.2 times), Environment (0.4 times), and Social (0.03 times). Participants asked more questions about Task factors, and requested more unavailable data about the same factors. Indeed there was a significant correlation of 0.90 between facts requested and extra facts requested ($p = 0.038$).

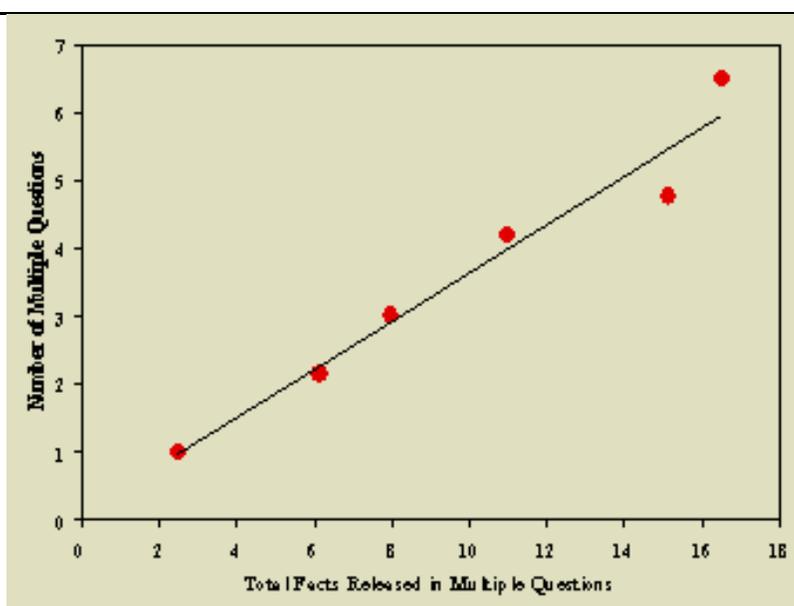


Figure 12. Multiple questions: correlation between the number of multiple questions asked and the number of facts released.

5.0 DISCUSSION

Our overall aims of the first year of the project were to develop a methodology for studying incident investigation and to validate the methodology using the baseline performance of different groups of participants. Both of these aims have been met, and in addition we have learned a great deal about how people in the aircraft maintenance industry go about investigating incidents.

5.1 WERE THE PROJECT OBJECTIVES ACCOMPLISHED?

As noted in the Introduction, the research reported here supports that of Marx (1999) who studied attribution of causal factors in incidents. The difference was that while he gave participants the total set of relevant facts about each incident, we only gave an initial trigger paragraph of facts, leaving participants to discover additional facts for themselves. Thus we are able to study more of the processes related to incident data: the data gathering rather than just the data interpretation. In fact, we did not ask the participants to list causes or contributing factors in any order, just to provide a synopsis of their findings including the contributing factors.

We used a methodology based on that of Woodcock and Smiley (1999), transferring it from the industrial accident domain to the aviation maintenance incident domain. This transfer was successful. Participants generally treated their experience with us seriously, although one or two clearly had other things on their minds and were quite brief in their interviews. The six scenarios we developed were all based on real incidents and were extensively reviewed and tidied by both the [SUNY](#) team and by our industry partners. These partners were from the Human Factors function of their organizations and so provided another link between the human factors knowledge and the industry knowledge bases necessary to provide realistic and useable scenarios. Material had to be added to the raw data from the incident reports provided by our airline partners as many facts of interest in investigation, particularly human factors issues, were not originally collected. At times, facts well-known within an organization had been omitted from the original reports, for example standard procedures for tag-out or shift change. While some further modification of the scenarios may be helpful before they are used in later phases of the project, the set of six scenarios used here worked successfully. They covered a wide range of tasks in heavy and line maintenance, and also on from line operations (Scenario 4). They all elicited many requests for information, and were substantial enough to require reasonable synopses from the participants. Later we shall discuss choice of a subset of the six scenarios for use in later project phases.

The sample of participants tested for the baseline data was adequate to validate the scenarios and methodology, but somewhat small to provide definitive measures of differences between different groups in the industry. With 37 participants (all that could be provided by our partners in the time available) we had a high representation from Quality Assurance investigators, as would be expected in an industry where such people are the main investigative arm. We were able to test fewer [AMTs](#) and managers / supervisors, primarily because they were more difficult for our airline partners to release for the 30-60 minutes required for the interview process. Because of this we were rarely able to detect differences in performance between the three groups of participants, but these differences were generally small anyway in absolute magnitude. The only reliable difference found between the groups was that [QA](#) investigators asked for out 15% more facts than [AMTs](#). Subsequent analysis showed that this was due to [QA](#) investigators having more concern for the Machine (equipment) fact type than either of the other groups. The conclusion reached is that treating all of our participants as a single group will not lead to erroneous conclusions. For the group we tested, their age and experience distributions were representative of an established workforce typical of major airlines and repair facilities. The sample may be older and more experienced than the industry as a whole which includes regional airlines, general aviation and smaller repair stations where many [AMTs](#) start their working lives.

On the matter of individual differences, there was a considerable range of performance exhibited, from 8 to 58 facts being requested, with an average of 26.6. However, we could find no reliable predictors of this or other performance measures using the demographic information we collected such as age, [AMT](#) experience or investigation experience. This is typical of many previous studies (e.g. Chervak, Drury and Ouelette, 1996). We have concluded before that all of our sample were experienced enough to perform the task, and that extra experience did not translate into improved performance (or alternatively to more complacency). This finding of few significant individual difference effects was reinforced by the lack of effects of [MRM](#) or [MEDA](#) training. We obtained a good sample with almost half having had [MRM](#) training and over one third having attended [MEDA](#) training. Among these participants at least, we found that these two training programs did not have much measurable effect. That is not to say that our future phases, designed to measure the effects of such training courses directly, will not find significant results. We will be using a more direct before-and-after methodology in the future where the same participants will be tested before training and after training. This will be a much more powerful experimental design than just sampling randomly from the existing population.

5.2 A MODEL OF INCIDENT INVESTIGATION

Using the literature from aviation and other fields, we developed a four phase model of incident investigation in the Introduction ([Section 2.2](#)). On the basis of our data collection and analysis on this project we can modify this somewhat idealized model to give a more accurate representation of what actual investigators do. Two of our four Phases (Trigger, Data Collection, Analysis and Reporting) were well supported by the study. Investigators were constrained to use a Trigger, here the trigger paragraph of the scenario, so that it is hardly likely that our data would not support this phase. The Data Collection and Analysis phases could not be separated in our study, and indeed it is doubtful whether they ever can in practice. Even in large public investigations, for example in air crashes or chemical releases, data and analysis guide each other throughout the investigation. Initial hypotheses are formed, data is collected to test these hypotheses and new analyses performed based on the outcome. This is an iterative process, as indeed it is for all social processes. For example, Fiske and Taylor (1984) discuss information seeking as a social strategy to reduce uncertainty of attribution (pages 33, 107). Thus if we must describe the process of incident investigation rather than merely prescribe it, we need to remove the temporal distinction between Data Collection and Analysis and treat them as two steps in an iterative loop rather than as distinct phases. In fact, our model of these two phases as shown in [Figure 11](#) could well be expanded to show the analysis steps explicitly. We did not collect direct evidence on this matter (e.g. by having participants “think aloud” during their investigation) so that we restrict our model to one based on the observed order of asking for facts of various classes.

In [Figure 11](#) we show that an initial stage is the exploration of the boundaries of the system under study. This is primarily a temporal exploration, as the spatial boundaries are largely implicit, e.g. the hangar or the departure gate. Participants needed to go beyond the temporal and spatial boundaries at times as their knowledge of the incident expanded. For example, the shift work history of the characters or the working norms in the organization will require accessing information from before the operational trigger of the incident. In this Boundary Stage the investigator extends the information from Phase 1 to help structure the rest of the data collection and analysis, so that in one sense this stage provides a logical bridge to the Sequence Stage.

The Sequence Stage is where the investigator collects data in a more-or-less chronological order, starting with Work Sequence and continuing to Inspection Sequence. At times this process must also proceed in reverse as the investigator attempts to reconstruct the incident from both temporal boundaries. For example, questions about why a cockpit door was removed can alternate with questions on why this removal was not picked up on inspection. In a heavily regulated and proceduralized industry such as aviation maintenance, there are procedural barriers (c.f. Reason, 1997) to error propagation in the form of independent inspection procedures. These are well known to all participants, so that they can deduce that this is a barrier that must logically have failed for the incident to occur. Thus many investigators approached their data collection from both ends, although the temporal ordering predominated ([Figure 10](#)).

The Stopping Rule Stage was not requested explicitly and perhaps should be in future uses of this technique. Participants stopped when they had either reached a point that would satisfy their superiors or would be inherently satisfying to themselves. They stopped collecting data when they found an explanation which matched in some way their expectations and experience. As Rasmussen (1990) put it, they had found a sequence of events that “matches a familiar prototype”. We found no evidence that would contradict such a statement, but then we did not ask for it. There must logically be a stopping point for each investigation, so we are on safe ground including it in our model, but more information on what exactly caused the data collection to cease would be of great practical interest in future work.

Our final Phase of Reporting remains from the model proposed in the Introduction, perhaps because we asked for a synopsis explicitly. In fairness, hardly any investigation takes place without some written findings separate from the data collection notes and other evidence so that participants were quite prepared to provide a synopsis. We asked for a verbal synopsis, such as would be given at a morning managers' meeting, rather than a written one to avoid participants having to take too long in the overall task. A written synopsis may have been more formally and comprehensively structured, but we suspect it would have been hurried so that participants could resume their regular duties without delay. In the synopsis, participants listed the (facts and inferences) they found most relevant or explanatory. This comprised a much reduced set of facts from that considered in the data collection stages, showing that not all investigation material was considered relevant beyond the investigation. We found no evidence of selection bias, in that some types of fact were suppressed or forgotten. The pattern of fact types retained in the synopsis was almost an exact reflection of that collected in the previous stages. The synopsis of an investigation, based on our findings, is likely to be an accurate reflection of the data collected, although not necessarily of the relevant data available to be collected. Any selection of the data appears to be at the collection stage rather than the reporting stage.

Based on our findings, we propose a model of the incident investigation process comprising what we shall now term Stages to distinguish it from the original proposal's Phases. It is given here as [Figure 13](#) in block diagram form as a summary of the discussion of this section of the report.

Within this framework, some observations can be made about the types and numbers of facts collected. First, not all fact types were collected equally. There were many more Task facts than the other types as the whole sequence of events had to be described, whether the facts were contributory to the incident or not. Thus, the fact that Harry Bird was assigned to inspect the [NRR](#) which started the missing cockpit door scenario is just a fact. Unless there were anything special about Harry Bird, it would not be even a pointer to a contributing factor, just a fact that would be needed in the investigation. Many facts in the reporting phase are of this nature: they give names, dates, aircraft tail numbers, maintenance locations etc. Thus, we would expect that these, which are strictly classified as "Task" would be over-represented in the data collection and reporting phases. It may be interesting in subsequent years of this project to distinguish between such non-contributory facts and those, such as an omitted step in a procedure, which are task facts and also contributory.

Of the other fact types, Operator and Social types were collected regularly and just as regularly made it through to the reporting stage. They seemed to provide the explanatory model sought by the participants, perhaps reflecting social norms (an [AMT](#) was tired after overtime or the inspector did not discuss his findings with the supervisor) or they may have come from an industry awareness of human factors. The aviation industry is probably unique in its emphasis on social and operator factors as comprising Human Factors, largely as a result of the awareness created by [MRM](#) programs and earlier [CRM](#) programs. In other industries, Human Factors implies at least as much emphasis on Machine and Environment factors as modifiers of human behavior and performance. In our experiment, there was a noticeable lack of data collection on Machine or Environment factors, unless these were direct contributors, such as the failure of the towbar (Machine) or the dark area in which the towbar was stored, leading to its defects not being seen (Environment). Our sample of [AMTs](#), managers and investigators behaved alike in this, except for the [QA](#) investigators who found relatively more Machine factors. If we are to have well-rounded investigations in practice, emphasis needs to be more evenly distributed in training and awareness.

One other factor that only becomes visible with detailed analysis of the actual facts requested is the emphasis on the paper trail in investigations. This is natural in a highly-proceduralized industry, where the belief is that strict adherence to written procedures and defined regulations will lead to safe operations. Additionally, when an investigative body such as a regulatory authority (e.g. [FAA](#), [CAA](#)) performs an investigation, they also emphasize procedural adherence. This may be due to the same belief in the efficacy of procedures, or it may be due to the ease of proving a violation rather than an error arising from other contributing factors. If the rules say you must do a particular act, and you do not, then you are de facto to blame for at least contributing to the incident. This is a valid but rather short-sighted view of the reality of a complex industry. No set of rules can ever be complete or universally-valid, although we try to approach this in many industries. But such a view leaves the equipment, training and environmental shortcomings of the real world out of the picture. As we have found in any really detailed investigation (e.g. those of aircraft accidents by [NTSB](#)), such factors can be major contributors to the propagation of error through the system towards a final undesirable outcome. No participant in our sample ignored such factors, and indeed many listed rule inadequacies (e.g. poor workcards) along with other contributors. But there was an emphasis on finding rule violations. Some participants even used this as their stopping rule.

Finally, we can return to our pattern of facts in an investigation, given as [Figure 1](#) in [Section 2.2](#). We now have actual measures of the numbers of facts at each stage from counting those in the trigger paragraphs (usually 7 or 8) and from counts in the study. [Figure 14](#) shows the findings in this form, with the new model presented above providing the new structure. As shown in this figure, the stages cannot be separated explicitly, reflecting the real process of investigation. To go deeper, we would have to create a more detailed version of the model of [Figure 13](#). It is interesting that the pattern proposed in [Figure 1](#) was largely validated by the data collected in this study. In general, about 20% to 25% of the facts collected appear in the reporting stage. It is assumed that more would be retained in a written report, but it is striking that only a fraction of the data collected is selected by the participants as relevant. This finding suggests that investigators should maintain files of raw data and investigation notes, rather than just their final reports on incidents. Most do this. However, it is the final report that forms the raw material for analysis later for trends and correlations in computer databases. We need to be aware of the potential biases in using such reported data rather than the original investigation. Useful contributory causes can be omitted from the data collection, but they can also fail to survive the final “cut” for the written report.

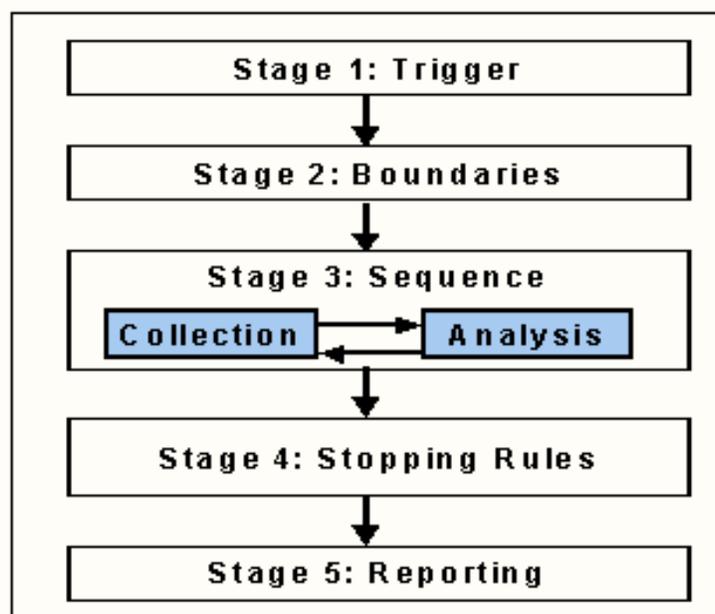


Figure 13. Model of incident investigation process

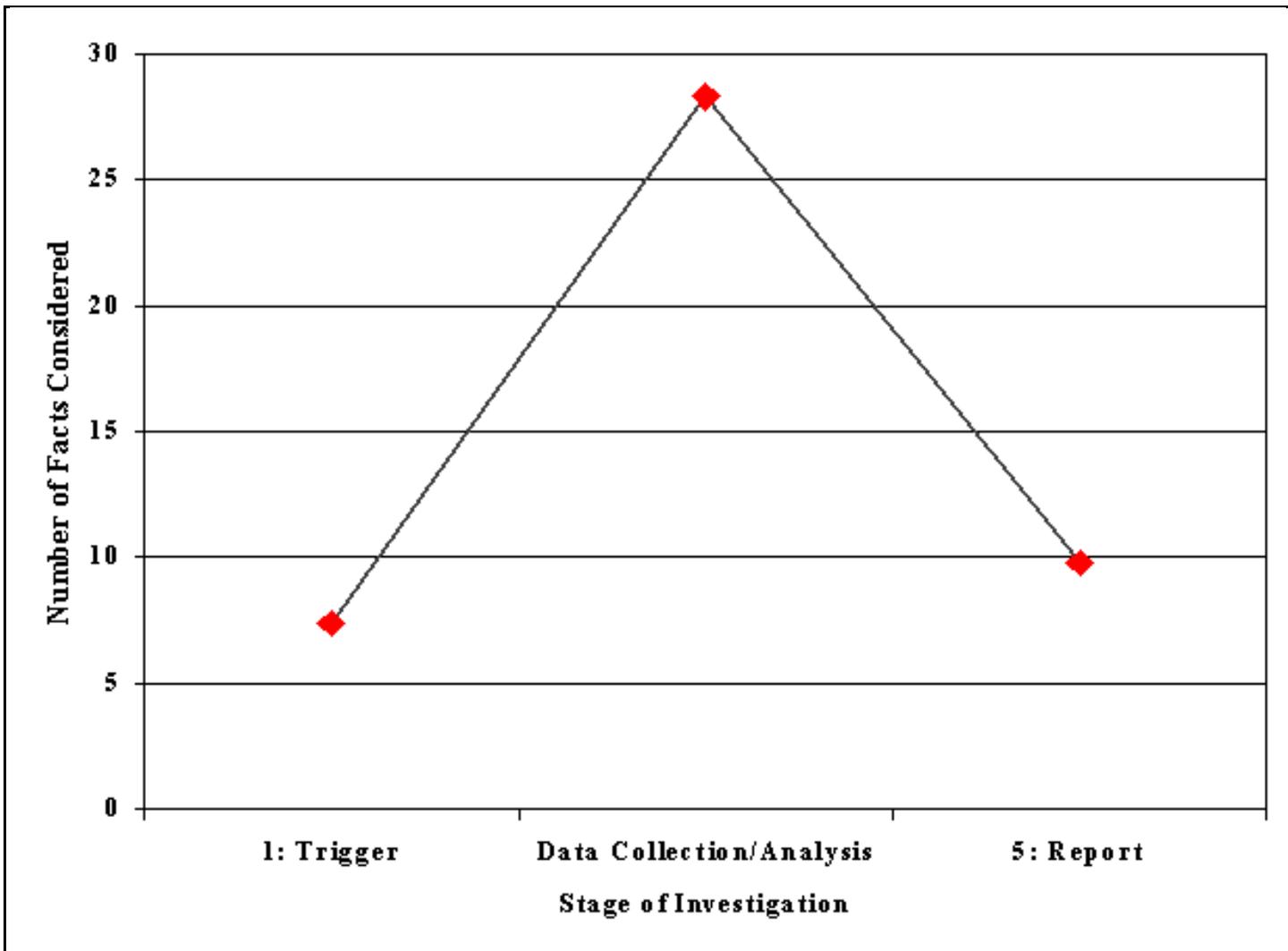


Figure 14. Findings for pattern of investigation

5.3. SCENARIOS FOR FUTURE PROJECT PHASES

In this study we have deliberately developed and tested more scenarios than we will need for future phases. This allowed us to use actual data from our sample of participants to help us select the four we shall need later. In fact, we could perform the rest of the phases with only two scenarios, using one before and one after each intervention ([MRM](#) training or [MEDA](#) training). However, four scenarios will make it more difficult for participants to share information among themselves in casual conversation

scenarios were collected from different partner airlines, and we found early in pre-testing that some of these incidents were quite famous within an airline. We would not want to use a scenario where the participants knew the original incident on which it was based.

On the basis of our study, we did not find any scenarios that were obviously unsatisfactory. Scenario 1 was quite difficult, and many participants did not fully investigate even the sequence of events. It also had almost no Environment facts so it is suggested that this scenario be eliminated. The other candidates for elimination are Scenario 5, which was quite short and Scenario 6 that also had few Environment facts.

A final decision on which to eliminate will be made with an expanded set of airline partners who have volunteered to take part in Phase 3 of the project. It may be that certain scenarios would be foreign to their operations, so that it is wiser to delay the final choice until Phase 2 begins.

6.0 CONCLUSIONS

The main conclusions from this study are simple: the methodology developed for studying incident investigation was successful in producing results that fit with the literature on incident investigations. On the basis of this phase of the study we can proceed to the future before-and-after studies of Human Factors interventions with some confidence that the methodology is valid and produces internally consistent results. In addition, we now have considerable data on how representative personnel within aviation maintenance go about investigating incidents, and a model to guide future data collection and analysis.

7.0 BIBLIOGRAPHY / REFERENCES

1. White House Commission on Aviation Safety and Security (1997). Gore Commission Report.
2. Reason, J. (1990). Human Error. Cambridge, U.K.: Cambridge University Press.
3. Marx, D. (1998). [Discipline and the blame-free culture](#). Proceedings of the 12th Symposium on Human Factors in Aviation Maintenance. London, England: CAA, 31-36.
4. Wenner, C. L. and Drury, C. G. (1997). Deriving targeted interventions for ground damage, SAE Technical Paper Series 972591. In Proceedings of the 1997 SAE Airframe/Engine Maintenance & Repair Conference (AEMR '97), August 1997.
5. Marx, D. (1999). [Maintenance error causation](#). Chapter 2. In Human Factors in Aviation Maintenance - Phase XI, Progress Report. Washington, D.C: FAA, Office of Aviation Medicine.
6. Hammer, W. (1989). Occupational Safety Management and Engineering. 4th Edition, NJ: Prentice Hall.
7. Ferry, T. S. (1981). Modern Accident Investigation and Analysis. NY: John Wiley and Sons.
8. Brown, I. D. (1995). Accident Reporting and Analysis. In J. R. Wilson and E. N. Corlett (Eds.) Evaluation of Human Work. London: Taylor & Francis, 969-992.
9. Rasmussen, J. (1990). The role of error in organizing behaviour. Ergonomics. 33.10, 1185-1199.
10. Maurino, D. E., Reason, J., Johnston, N., and Lee, R. B. (1995). Beyond Aviation Human Factors. Aldershot. England: Avebury Aviation.
11. Norman, D. A. (1980). Errors in Human Performance. Report #8004, Center for Human Information Processing. San Diego: University of California.
12. Wickens, C. D. (1994). Engineering Psychology and Human Performance. New York: Harper Collins.
13. Feggetter, A. J. (1982). A method for investigating human factor aspects of aircraft accidents and incidents. Ergonomics, 25.11, 1065-1077.
14. Engkvist, I-L, Hagberg, M., and Wigaeus-Hjelm, E. (1995). Interview protocols and ergonomics checklist for analyzing overexertion back accidents among nursing personnel. Applied Ergonomics, 26.3, 213-220.
15. Wiegmann, D. A. and Shappell, S. A. (1997). Human factors analysis of post-accident data: Applying theoretical taxonomies of human error. The International Journal of Aviation Psychology, 7.1, 67-81.
16. Fiske, S. T. and Taylor, S. E. (1984). Social Cognition. Reading, MA: Addison-Wesley.

17. Woodcock, K. and Smiley, A. (1999). Developing simulated investigations for occupational accident investigation studies, prepublication draft.
18. Torell, U. and Bremberg, S. (1995). Unintentional Injuries: Attribution, Perceived Preventability, and Social Norms. *Journal of Safety Research*, 26.2 63-73.
19. Bureau of Labor Statistics (1991). Washington, D.C.
20. Chervak, S., Drury, C. G. and Ouelette, J. L. (1996). [Field evaluation of Simplified English for aircraft workcards](#). Human Factors in Aviation Maintenance - Phase Six, Progress Report, DOT/FAA/AM-96/xx, National Technical Information Service, Springfield, VA.
21. Reason, J. (1997). [Approaches to controlling maintenance error](#). In Proceedings of the FAA/AAM 11th Meeting on Human Factors Issues in Aviation Maintenance and Inspection, San Diego, CA.

8.0 ACRONYMS

AAM Medicine	FAA's Office of Aviation
ANOVA	Analysis of Variance
AMT	Aviation Maintenance Technician
BLS	Bureau of Labor Statistics
FAA	Federal Aviation Administration
FAR	Federal Aviation Regulation
GMM	General Maintenance Manual
MEDA	Maintenance Error Decision Aid
MRM	Maintenance Resource Management
NRR	Non-routine Repair
NTSB	National Transportation Safety Board
OSHA	Occupational Safety and Health Administration
QA	Quality Assurance
QC	Quality Control
RTS Service	Return to
SHELL	Software / Hardware / Environment / Liveware / Liveware system
SUNY	State University of New York
TOMES	Task / Operator / Machine / Environment / Social system

9.0 APPENDIX 1

9.1 BRIEFING/DEBRIEFING, CONSENT AND PARTICIPANT INFORMATION FORMS

Incident Investigation in Aviation Maintenance Briefing

This experiment will take less than an hour of your time.

The purpose of this experiment is to see how you investigate aviation maintenance incidents. We are performing a project for the FAA's Office of Aviation Medicine which will help us to provide better incident investigation methods. (Phase 2 and Phase 3 only: We are also using the data to measure what is learned during a training course.) Although the FAA is funding this study, they will never see the raw data from our session, and cannot trace any responses back to you as an individual or to your company.

After collecting some data about your demographics, I will give you a short description of an incident. This is the sort of description you may get via email or by telephone from somebody asking you to investigate the incident.

I have access to all of the data about this incident so I want you to ask me for specific data just as you would if you were investigating this incident at your own worksite. If a particular fact you request was never collected, I will let you now. You may take any notes you like on the pad provided.

Please continue to ask for information as you normally would, and stop when you normally would. When you get to the stopping point, I would like you to give me a synopsis of the incident, as you would in a report. Please list the contributing factors in your synopsis.

I will be analyzing the questions you ask, and so would like to make an auditory tape of the investigation. You may refuse to be tape recorded, in which case I will be taking more extensive notes. The audio tape will be erased after we collect the data from it on what questions you asked. If you agree to be audio taped, please initial the box on the consent form. I would also like to have your notes to help me reconstruct your investigation.

Let me emphasize again that all data are confidential. Your data will be recorded under a subject number only. I do not even need to know your name. No raw data can get back to your company or to the FAA. You may withdraw from the experiment at any time without penalty.

If you agree to continue, please read and sign the consent form, then we will begin.

Any questions before we start?

Colin G. Drury, Department of Industrial Engineering
University at Buffalo, 342 Bell Hall, Buffalo, NY 14260
716-645-2357 x2117, or email at: drury@buffalo.edu

Incident Investigation in Aviation Maintenance Consent Form

I, _____, have read and understand the briefing information, and my questions have been answered. I understand that participating in this study is voluntary. I understand that I need not provide any information that I do not wish to disclose, and that I do not have to answer any questions that I do not wish to answer. I also understand that I can choose to withdraw from this experiment at any time without penalty, and that I may choose to withdraw my data at the end of the experiment. There are no known risks, hazards or inconveniences to participating in this research.

I agree to participate in this research.

Signature: _____

Experimenter's Signature _____

Date: _____

I agree to auditory tape recording of this session. (Please initial if you agree.)

For further information you can contact the Principal Investigator on this project:

Colin G. Drury, Department of Industrial Engineering
University at Buffalo, 342 Bell Hall, Buffalo, NY 14260
716-645-2357 x2117, or email at: drury@buffalo.edu

Questions about your rights as a research subject can be directed to the Human Subjects Review Committee, 810 Clemens Hall, University at Buffalo, 716-645-2711.

You may keep a copy of this form.

**SUNY Buffalo
Participant Data Sheet**

Date/Time	
Participant Number	
Scenario Number	
Airline / Organization Number	
Site Number	

Job Title	
Years as AMT	
Years as Investigator	
HF Training Details	
Investigation Training Details	
Number Investigated in Previous Years	

Time Since Last Investigation	
Age, gender	

Incident Investigation in Aviation Maintenance Debriefing

This was part of an experiment to see how people investigate incidents in aviation maintenance. (Phases 2 and 3 only: We are also using the data to measure what is learned during a training course.)

Do you have any questions about this research or your participation?

I want to reassure you that the data are confidential, and that no raw data will be released to your company or the FAA. The audio tapes will be erased as soon as we have analyzed your data.

Please do not discuss the specific incident you have just seen with any colleagues. They may be future participants in the experiment and we do not want to bias their data.

If you have any further questions, please contact

Colin G. Drury, Department of Industrial Engineering
University at Buffalo, 342 Bell Hall, Buffalo, NY 14260
716-645-2357 x2117, or email at: drury@buffalo.edu

9.2 LIST OF FIGURES

- Figure 1 Expected increase and decrease in number of facts considered at each stage
- Figure 2 Age distribution of sample
- Figure 3 Experience distribution of sample
- Figure 4 Number of facts investigated by scenario
- Figure 5 Number of facts included in each synopsis
- Figure 6 Correlation between total facts and facts investigated
- Figure 7 Types of facts investigated for each scenario
- Figure 8 Types of facts included in the synopses for each scenario
- Figure 9 Correlation between types of facts included in the synopsis and investigated
- Figure 10 Distribution of correlations between investigation order and chronological order
- Figure 11 Three stages of investigation sequence in Phases 2 and 3
- Figure 12 Multiple questions: correlation between the number of multiple questions asked and the number of facts released
- Figure 13 Model of incident investigation process
- Figure 14 Findings for pattern of investigation

FINAL REPORT