Development and Validation of

a human error management taxonomy in air traffic control

Bove, Thomas

Publication date:
2002

Development and Validation of

A HUMAN ERROR MANAGEMENT TAXONOMY IN AIR TRAFFIC CONTROL

By

Thomas Bove

A Ph.D. thesis

Risø National Laboratory &

University of Roskilde

2002

Supervisory Committee:

Senior scientist Henning Boje Andersen, MA,
Systems Analysis Department
Risø National Laboratory
DK-4000 Roskilde
Denmark

Professor Stig Andur Pedersen, MS,
Department of Philosophy and Science Studies
Roskilde University
DK-4000 Roskilde
Denmark

ACKNOWLEDGEMENT

The author wishes to thank all the people that have contributed to making this project possible. First, I would like to thank Preben Lauridsen and his controller colleagues from Kastrup airport for giving me the opportunity to sit and watch how they carry out their work in the different operational positions. I would also like to thank all the air traffic controllers who volunteered in participating in interviews and openly described critical episodes they had experienced. Furthermore, I would like to express my gratitude to Per Henriksen, Anne Kathrine Jensen, Ryan Sørensen and Jørgen Jørgensen who were very supportive in relation to making it possible to conduct the empirical studies. I would also like to express my appreciation to the two instructors from the Danish CAA – Pia Lendal and Karsten Balslev - who participated in reviewing all the video recordings from the simulator scenarios and patiently provided me with explanation of the errors that were observed in the scenarios. The author also gratefully acknowledges Anders Jernberg and his colleagues from the Swedish CAA for helpful assistance in interpreting incident reports.

I am greatly indebted to several people at Risø National Laboratory. In particular, a great debt is owed to Gunnar Hauland who kindly let me use the video recordings from his Ph.D.-project and patiently spent time to explain me the experimental scenarios. Furthermore, I appreciate his contribution through the many interesting office discussions we have had and for providing me feedback on a later draft of the dissertation. I would also like to thank Marlene Dyrløv Madsen who spent numerous hours analysing and classifying the data material. Finally, I am very grateful for the technical assistance that Erling Johannsen has provided me with throughout the whole project.

In the initial phase of the thesis I was involved in the development of an error taxonomy to be used in the area of Air Traffic Management. The project was carried out as a cooperative effort between Risø National Laboratory and National Air Traffic Services (NATS) for the European Organisation of Air Traffic Control (Eurocontrol) and was given the acronym HERA (Human Error Reduction in ATM). The current thesis can be seen a continuation of this work and the author wishes to thank all the people involved in the HERA project - including Steven Shorrock, Anne Isaac, Barry Kirwan, Richard Kennedy and Henning B. Andersen - for a very interesting and stimulating cooperation.

As a part of my thesis project I spent three months at the Human Factors Research Project at the University of Texas, USA and I would like especially to thank Bob Helmreich and his colleagues at the laboratory for the hospitality and inspiration I received during my stay. At the Austin lab I have obtained insight into the different classification systems that they have developed (e.g. LOSA and ASAP) which has inspired much of my work on error management. Furthermore, I would like to express my appreciation to Shanna Smith at the statistical help desk at the University of Texas for her useful assistance in relation to the statistical analysis.

I would also like to extend my thanks to Lisette Kanse who, in parallel with this thesis, has been writing a Ph.D. at Eindhoven University of Technology about error recovery. I

have benefited from the fruitful discussions we have had and been inspired by the interesting and insightful contributions that she has made within the area of error recovery.

Several human factors specialists participated in a questionnaire study where they were asked to comment on the framework developed in this thesis. I would like to express my gratitude to all those who were so kind as to take the time both to read through the material I sent to them and to provide me with constructive feedback.

Finally, I would like to express a huge thank my two supervisors on the project – Henning B. Andersen and Stig Andur Pedersen – for all their support, helpful comments and useful feedback.

Thomas Bove, June 2002

SUMMARY

The main objective of this dissertation is to develop, validate and evaluate an error management taxonomy to be used in the analysis of error events in the area of Air Traffic Management (ATM). The goal of the taxonomy is to be able to analyse the mechanisms behind human errors and their recovery. Currently an abundance of taxonomies exist to describe the mechanisms behind human errors whereas very little is known about the mechanisms underlying error detection and recovery. This is unfortunate since timely and effective interventions can often prohibit errors from having serious consequences on system safety. The goal of the present Ph.D. project is therefore to gain more knowledge about how errors are captured. One of the desired outcomes of this project is to provide a basis for reinforcing incident prevention strategies. To do so, it is important to have a structured classification scheme (a taxonomy) in which operational data about the *production*, *detection* and *recovery* of human errors can be categorised including the underlying circumstances behind these human errors and their capture.

The report is divided into four parts:

Part one – Background. The first part elaborates on the importance of human error and error management in the area of in ATM as well as some generic requirements to an error management taxonomy. For readers who are not so familiar with the domain of air traffic control a brief description of the ATM system is provided.

Part two – Literature review. In the second part a literature review is carried out to determine which categories should be included in the error management taxonomy. The focus is on taxonomies associated with human error as well as on human performance issues occurring both before and after errors. Before the occurrence of an error the focus is, in particular, on the issue of threat management which concerns how operational factors that have the potential of leading to errors and jeopardising safety are controlled. In relation to the phase after the occurrence of an error there are, in particular, four main issues that will be explored: *who* was involved in the detection and recovery of the error and/or its consequences; *when* was the error or its consequences detected; *how* was the error and/or its consequences detected and corrected; and finally *what* was the behavioural response and outcome*?* In addition, it should be possible to also give an answer to the *why*-question – namely why did the error occur and why was it successfully or unsuccessfully managed? This can be determined on the basis of so-called Performance Shaping Factors (PSFs) that can be seen as contextual factors that can have a positive or negative influence on the course of events.

Part three – Construction of the taxonomy. In the third part the error management framework will be described. It has been developed on the basis of the literature review and it has been further refined and tested on the basis of incident reports, critical incident interviews and simulator studies (the results of which will be described extensively in part four). The framework is organised around an error management model. It consists of two main components: The core of the framework is developed on the basis of the

literature review of error and error management taxonomies. The list of contextual factors is developed on the basis of the review of the Performance Shaping Factors.

Part four – Validation. Finally, the utility of the error recovery framework in relation to error management analyses will be explored. For this purpose the framework has been evaluated on the basis of different kinds of data material. First, the framework has been applied to error events found in critical incidents (both Swedish CAA incident reports and critical events elicited through the critical incident technique) and in a simulator study. On this basis it has been possible to obtain knowledge about the extent to which consistent classifications can be obtained (both across time and raters) and, furthermore, to explore the chances of discovering patterns that can yield insight into these different kinds of data material. Second, the framework has been evaluated by a series of human factors experts who have been involved in research that is highly relevant in relation to the current project. In this manner it has been possible to get a both qualitative and quantitative evaluation of the framework.

The results aimed at applying the framework indicate that it is both possible to achieve fairly robust analyses on the basis of the framework and that the framework could verify results from other studies as well as provide new insights. Furthermore, the results from the questionnaire revealed that the experts found the framework highly relevant in relation to the study of error management. In sum, the results obtained from the four studies provided support for the notion that the framework could be of use in future error management studies. In particular the framework could be useful in relation to analysing the effects of various ATM safety initiatives, be they changes in system design, operating procedures or training of personnel.

TABLE OF CONTENTS

FIGURES

TABLES

PART ONE

BACKGROUND

1 Introduction

One of the most well known accidents in the aviation history is the Everglades accident in 1972. The accident occurred at a time where the pilots were engaged in solving a problem in the cockpit in relation to a landing gear warning light (Wickens et al., 1997). Unfortunately one of the pilots had inadvertently disengaged the auto-pilot by touching the steering column and as a consequence of this the aircraft was descending. The pilots did not discover this because they were so preoccupied with the indicator in the cockpit. On the ground the controller could see on the radar that the aircraft was descending and therefore called the aircraft and asked: "How are things going out there?" The pilot thought the ATCO was referring to the problem with the indicator and therefore answered that they were doing fine. Moments later the aircraft it crashed into the Everglades swamp. Clearly, the chain of events leading to the accident was initiated by errors made by the pilots. First, by inadvertently touching the steering column and, second, by getting fixated on the single problem and not distributing their attention in an appropriate manner. In spite of this it seems in retrospect evident that the ATCO could most likely have played an active role in the recovery by making a more explicit communication with the pilots.

Air Traffic Management (ATM) has been a relatively high reliability system for some time. Even though air traffic controllers (ATCOs) every day are in many facilities required to handle a large quantity of aircraft very rarely do ATM related accidents – such as the one described above - occur. Irrespective of the impressive safety record of ATM many studies from a number of different safety critical areas - such as aviation, process control and maritime operations - have shown that a majority of incidents and accidents involve human error. The current air traffic system is in some respects stretched to its capacity limits and the challenges to safety of the ATM system may increase in the near future due to the projected traffic level increases and the introduction of computerised and automated tools. These changes will have impact on the method of operation in ATM and may affect the types of errors, the error rates and the chances of recovery. As a consequence of this it is important to be able to learn from human error events to ensure that the current high-level of the safety of the system will not be compromised.

Studies have shown that human errors have contributed to about 90% or more of ATM incidents (Kinney et al., 1977). A fundamental question then arises, namely why do these human errors happen? By simply stating that almost all incidents are related to human errors does not advance the understanding of the incident causation and thereby the chances of mitigating the causal sequence of events. Indeed, if the investigation of critical events in the area of ATM stops at the conclusion that it was caused by a controller error, little is achieved except finding a culprit for the adverse consequences. The chances of learning from the incidents and thereby understanding why they occurred have been

omitted and, just as important, we do not obtain knowledge about how many similar errors normally are prohibited from having consequences on the system safety.

To minimise the risk of events that may compromise the safety of the air traffic it is important to develop error resistance strategies. Error resistance strategies can be divided into two main categories, namely *error prevention* and *error correction* (Lewis and Norman, 1986; Frese, 1991). Traditionally human error resistance strategies have mostly focused on error prevention. This focus is understandable since many studies of incidents and accidents in safety critical domains indicate that the underlying problem is often to be found in a combination of shortcomings of human performance in man-machine systems and the fact that most of such systems have been designed to be unforgiving to errors. An obvious solution to avoid such unwarranted consequences is to make initiatives to prevent the occurrence of human errors (e.g. through failsafe protection devices, automation and enhanced procedures).

Safety strategies narrowly based on error prevention may not be successful for several reasons. First of all, human errors will inevitably occur and it is impossible to anticipate which errors will occur in a specific task context. In particular errors that require insight into the higher underlying goals may be difficult to detect by automated detection devices (Brodbeck et al., 1993). Second, by focusing exclusively on avoiding various kinds of errors there is a risk of imposing excessive limitations on the performance which may compromise both effective and adaptive behaviour. Actually, it has been argued that the efficiency of error avoidance strategies has been exhausted in ultra-safe areas such as aviation and air traffic control, and that the end result of increased error suppression may in fact be counterproductive seen from a safety perspective (Amalberti, 2001). Thirdly, studies have shown that most errors are actually detected and recovered before leading to adverse consequences by either the perpetrator or colleagues (Amalberti & Wioland, 1997). Since human errors are inherent to real life and people have powerful capabilities to control errors it is important to a larger extent to try to manage the manageable and to support people's chances of detecting and recovering from errors. Consequently, error management should be considered an important supplementary safety goal.

In spite of a growing interest in the field of error management the understanding of how errors are detected and recovered has failed to keep pace with the understanding of the mechanisms underlying human error. A possible explanation of the scarcity of studies of the error handling process in safety science may be found in the fact that error reduction has for a long time been considered the primary and most important means to achieve high reliability and safety. In other words: the "zero accident policy", which remains the ultimate safety goal, has been interpreted as the "zero error policy" (Wioland & Amalberti, 1996). The zero-error policy is not only problematic because it ignores the fact that there is a random aspect about human errors and some types of errors may be difficult to avoid (e.g. cognitive tunnel vision and confirmation bias). The problem is also that the zero-error policy underestimates or ignores the potentially positive value of errors (Senders & Moray, 1991; Frese & Van Dyck, 1996):

- *Coping strategies.* The result of an over-emphasis on error avoidance will be a reduced ability to cope and control error-induced problems. On the other hand, the more errors we make the better we get at dealing with them. Since errors can be upsetting and frustrating, experience with error situations can be important in relation to learning to deal effectively and rationality with such situations. That is, by having experienced similar error situations before the person will know how to correct the error quickly and, as a consequence, also be less upset when such situations arise.
- *Creative solutions.* Error avoidance strategies will put limits on the range of behaviour that is possible and thereby reduce the chances of applying creative solutions to novel and unexpected problems. Such solutions might be important in relation to finding better ways of doing things.
- *Mental models.* Experience with error situations may be beneficial in the development of an understanding of the dynamics of a system. Errors constitute an important feedback concerning what the person does not know yet and can therefore be used as a means to remove previous misconceptions.

The increased scientific and practical interest in the field of error management has created an impetus for a new attitude towards human errors and towards the role of human operators in the control of complex systems. The traditional view of the human operator within the area of human reliability research has been that human operators are "intelligent but fragile" machines and, as a consequence, the role of the human actor should be minimised (e.g. through automation). A more positive attitude towards the human operator has gradually emerged as a result of recent research findings. These results have, among other things, revealed that human operators develop protections and defences against their own cognitive deficiencies (Amalberti & Wioland, 1997). In this manner the human operator plays a positive role in returning a system to a normal and safe state after the occurrence of an error. This positive role is, of course, not limited to the recovery of human errors, but also to technical failures.

If error management should be considered an important safety strategy it should be possible to build safety barriers into man-machine systems based on error management. Actually, research has shown that recovery is more than sheer luck and coincidence (Van der Schaaf & Kanse, 2000). The results indicate that recovery is something that can be planned for and that the human operator can play a powerful role in relation to preventing small failures and errors from developing into actual system breakdowns. More specifically, several researchers have demonstrated how different human factors initiatives - such as training, design and organisational culture - can support the human recovery process. Some examples are provided below:

1. *Design.* The chances of coping with errors in man-machine systems are dependent on several system factors and it is important that the system designs focus on mitigating the consequences of human error. Rasmussen (1984) has proposed that error recovery is critically dependent on the *observability* and *reversibility* of errors and their effects. Reversibility is within complex environments mainly dependent on system dynamics, but can be supported by system features such as equipment redundancy or by delaying the effect of executed actions. Observability, on the other hand, is dependent

on interface features (such as the visibility, the immediacy and the validity of feedback information) and the chances of perceiving mismatches between the expected and the actual systems state. The system is error tolerant if it is easy to observe erroneous actions and if they do not have immediate and irreversible effects (for example, in ATM mid-term and long-term conflict alerts can enhance the chances of an early recovery without any serious system consequences). By following these principles it should be possible to structure the environment to improve error detection and recovery.

2. *Training.* It has been suggested that coping with errors may be an important part of acquiring skills and expertise in a specific domain (Seifert & Hutchins, 1994). Therefore, errors should not necessarily be viewed as something that should be avoided at any prize, but instead as an opportunity to develop professional problem solving abilities and the system should support coping with them to ensure low system output error. In fact, the most proficient problem solvers may not necessarily be those who commit fewer errors, but those who have the greatest abilities to recover from their errors (Allwood, 1984). Reinforcing the error handling capabilities may be achieved through training concepts such as error management training. In this training technique trainees are encouraged to have a positive attitude towards errors (e.g. by simple heuristics such as "I have made an error. Great!") and forced to make errors (e.g. by giving them problems that exceed their level of expertise) (Dormann & Frese, 1994). Several human-computer-interaction studies (Dormann & Frese, 1994; Nordstrom et al., 1998) have demonstrated a higher after-training performance for trainees being exposed to the *error management training* compared with trainees being exposed to *error avoidance training*. The realisation of the potential benefit of error management training has also reached the aviation community. It has, for example, been suggested that to improve the error handling skills of teams and crews instructors and evaluators should change the focus from detecting and correcting errors to observing the crews' error resolution process (Tullo & Salmon, 1997). If it is successful it should be rewarded and, if not, efforts should be made to examine the error resolution process and how it can be improved.

3. *Organisational culture:* Organisations may vary in relation to their error culture. At one extreme an organisation might aim at avoiding errors at all insofar as errors may be associated with grave consequences. This approach can have some negative side effects. If, for example, an organisation has a strong error prevention philosophy it will normally imply that errors are severely sanctioned which means that people will be unwilling to report errors. The result may be that the organisation misses a vital source of information in relation to learning from errors. Furthermore, a defensive attitude might be the result and many resources will be spent on covering up errors instead of benefiting from their potential learning value. On the other extreme, organisations associated with a high-level of error tolerance - entailing factors such as openness to errors, analysis of the errors committed and long-term learning - may instead benefit from errors. Such an approach may ultimately play a significant role for the success of an organisation insofar as empirical data supports the notion that organisations characterised by an error tolerant culture have a higher tendency to also

be associated with the highest level of performance, as measured by both subjective and objective performance scales (Van Dyck et al., 1999). As also supported by the above-described studies of error management training this could indicate that the way people perceive errors and handle them influences performance. It can also indicate that these organisations become better at implementing defences and barriers to avoid the negative consequences of human errors.

As indicated by the previous paragraphs improvement in system safety in the area of ATM requires gaining systematic and detailed knowledge about the underlying mechanisms of not only error production but also error recovery. An important step in that direction is to explore whether existing error taxonomies, expanded with a classification scheme of how they were managed, can be applied to studies of human errors in ATM. Such a taxonomy can constitute a useful human factors tool in diagnosing underlying mechanisms behind human errors and their resolvement. The results can be useful in relation to analysing the effects of various ATM safety initiatives, be they changes in system design, operating procedures or training of personnel.

The goal of this project is to develop a taxonomy to study human errors and their resolvement within the area of ATM. Important benefits may be associated with attempts to develop a coherent framework to study both errors and their recoveries. This is, in particular, important because error production and error management cannot be properly understood as isolated issues but are instead closely intertwined (Amalberti & Wioland, 1997). By focusing on both error production and management in the development of a coherent error analysis framework it becomes possible to analyse these closely related issues in an integrated manner.

1.1 Human error

On an intuitive level most people are able, within their own domain, to make confident judgements about whether something is an error or not. In particular, when adverse consequences are observed and human actors played a central role in the course of events it seems straightforward to attribute the cause of the situation to human error. However, when trying to analyse the concept of human error in more detail it turns out that it is a rather elusive concept and is associated with many different meanings. This confusion is an impediment to developing structured and effective countermeasures to human error. In the following we will briefly review some of the problems and ambiguities associated with the label “human error” (for elaborated discussions please refer to Rasmussen, 1983b; Reason, 1990; Woods et al., 1994) and provide the reader with a working definition of human error.

There have been some discussions about whether it is correct to speak of "errors" at all. It has been argued that the attribution of cause to the human (and not some system) components is to some extent dependent on the stop rule applied in the after-the-fact analysis of the causal chain and is therefore not dependent on any objective standards (Rasmussen, 1983b). When analysing the causes of substandard system performance the

search back through the causal chain will typically continue until a familiar and reasonable explanation has been found and a cure is available. Since the human component plays a salient role in the man-machine systems there is a large chance that the search for the causal explanation will stop when having found a "human error". The attribution of the cause to human error is also convenient because a well-known cure is available (such as "blame and train").

The label “human error” also implies that the problem is to be found is on either the human or the engineered side of the man-machine equation. Attempts to assign causal factors to either the human or the technical components may be limited by the fact that these components are closely intertwined in any man-machine system and should not be analysed in isolation. If, for instance, an Air Traffic Controller (ATCO) does not detect the presence of an aircraft on the radar due to glare or reflection, should this be characterised as an error or not? One could argue that the ATCO is not responsible for the omission and that it should not be considered an error. On the other hand, if he had moved his head slightly to the side - and thereby moved the reflection - the aircraft would probably have been detected. Consequently, it is far from always easy to decide and agree on the presence of an error. As a result of such considerations researchers have suggested that the term "human error" should be replaced with "man-machine misfits" (Rasmussen, 1983b) or "erroneous actions" (Hollnagel, 1990).

An attribution of adverse events to the human component of the man-machine system is also problematic because it from a legalistic perspective implies a neglected and thereby punishable act. In this manner the human controller becomes the scapegoat and is blamed for some undesirable consequences. This is unfortunate because the actions performed by the controller are selected on the basis of what is thought to be the most appropriate action in the given situation. Nonetheless, bad outcomes (i.e. incidents and accidents) will often be attributed to process defects (for example, a bad decision) in spite of the fact that there is a loose coupling between process and outcome. That is, good decisions may in some situations be followed by bad outcomes, but in other situations be inconsequential. Nonetheless, hindsight bias - i.e. the tendency to judge the quality of the process on the basis of the product and to over-estimate what could have been known in advance - can have strong implications for error analyses.

Finally, to label some process defect as “human error” implies that there exists a criterion or standard that the performance can be compared with and that the performance does not satisfy this criterion. The most obvious criterion to apply is the standard operating procedures. However, the standard operating procedures do not necessarily constitute an unambiguous criterion. First, not all situations can be covered by the standard operating procedures. Second, people rarely recognise rule violations made by themselves as errors, because the violations may be motivated by efficiency and/or safety concerns (for example, an ATCO may not provide a pilot with traffic information after the resolvement of a conflict due to other pending tasks in spite of this being required according to the procedures). Actually, violations and modification of formal rules might be quite rational given the actual workload and time constraints. Furthermore, such violations might be an integrated part of the established practice (Rasmussen, 1997). In addition, as pointed out

by Reason (1997) the proliferation of well-intended procedures may serve to reduce compliance with procedures and may thereby be an invitation to violations.

An alternative criterion could be violation of good working practice. That is, would other professionals with similar background have acted likewise given the constraints of the situation or was the performance below a generally accepted standard. This criterion may be of particular relevance in an area such as air traffic control insofar as it is a domain that is, in comparison with e.g. aviation, less dominated by procedures. However, this criterion is also problematic because different people may have different conceptions (dependent on e.g. background and culture) of what constitutes good working practice and, again, hindsight bias may obscure the validity of such judgements.

As the previous paragraphs have suggested it may often be difficult to agree on whether something is an error or not. It has to be accepted that human error is not something that has an objective existence on its own but is instead a social construct which meaning is dependent on consensus agreement. Even though it is probably impossible to develop an unambiguous and uncontroversial definition of the concept of human error it is useful to decide on a working approximation. In the current context the following definition lays down whether an error has been committed (adapted from Isaac et al., 2001):

> *Any action (or inaction) that potentially or actually results in negative system effects given the situation that other possibilities were available. This includes any deviation from operating procedures, good working practice or intentions.*

There are several benefits of this definition. First, the definition of *human error* is *neutral* with regard to any question of *blame.* Second, an error does not need to involve any system consequences. This is in concordance with the principle that an error should be judged on the basis of the underlying processes and not the product. Third, an action or inaction can only be labelled as an error if the actor involved could have acted differently given the constraints of the situation. That is, it does not make sense to classify something as an error if no other alternative was available. Finally, the definition accepts several different criteria or standards to which the performance can be compared, namely the standards operating procedures, good working practice or simply the actor's intentions.

1.2 Error management

In the human factors literature there have been several suggestions concerning which label to attach to the process that follows production of error - that is, the process from error detection to recovery. Some examples are *error handling* (Zapf & Reason, 1994), *error recovery* (Lenman & Robert, 1994), *failure recovery* (Kanse & van der Schaaf, 2000a) and *error management* (Frese, 1991). Both the first and the second part of these concepts are associated with some ambiguities. In the following some of these problems will briefly be described.

Both the term "management" and "recovery" are associated with some inherent problems because they can refer to different things. The concept of management is ambivalent because it can both refer to how front-line operators and high-level decision-makers within an organisation deal with errors or faults. The concept of recovery is, first of all, problematic because it can refer to both a part and the entire process that follows the production of an error or fault. Also, more specifically, it may lead to some confusion within the medical domain where recovery is used in a different context (i.e. improvement in the patient's physical state which is normally expected after an intervention). Some alternatives to these two concepts could be "handling" or "capture". Both of these concepts are more neutral in nature, but at the same time have only been used very rarely in the research literature. In short, there is no obvious choice between these different alternatives and they will therefore be used interchangeably throughout this thesis.

A more clear-cut choice of terms seems to be available when examining the first part of the above listed concepts to describe the process from error detection to recovery. In this context it is useful to examine the figure below that describes the relationship between the three central concepts, namely "error", "fault" and "failure":

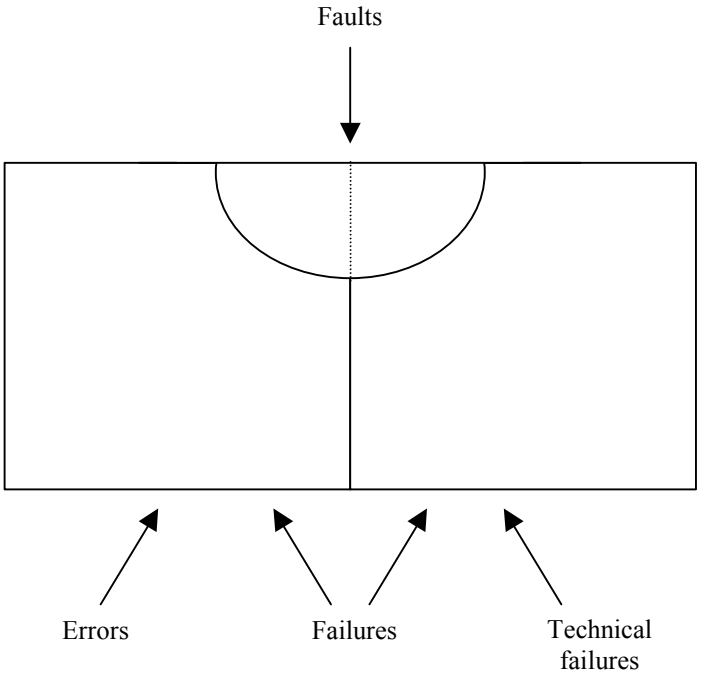


Figure 1: Graphical illustration of the relationship between errors, failures, technical failures and faults.

The error part concerns those situations where a human error has been committed. Only a subgroup of the errors leads to system breakdowns (i.e. faults) and a major part of the errors do not have any consequences at all. On the other hand, faults concern system breakdowns irrespective of whether these are human- or technology-induced. As can be seen in the figure there is an overlap between the two concepts and it is concerned with the human error induced system breakdowns. The total of the two squares can be referred

to as failures.

The ambition is in the current context that the taxonomy should be applicable to any type of system breakdown irrespective of the origin to the failure. Nonetheless, the main focus will in the current context be on human induced error situations. Therefore, we will stick to the term "error management" and not "fault management" (see e.g. Johannsen, 1988) or failure management (so, for example, a total radar blackout would in itself not be a recovery event of relevance in the current context).

Several definitions of error management are available (see e.g. Frese, 1991 and Rizzo et al., 1995). Below is presented a definition of error management that – in spite of focusing on the crew level rather than the individual-cognitive level – gives a precise picture of how the concept is used in this thesis:

> *"Error management at the crew level is defined as actions taken either to reduce the probability of errors occurring (error avoidance) or to deal with errors committed either by detecting or correcting them before they have operational impact (error trapping) or to contain and reduce the severity of those that become consequential (error mitigation). It is also possible for crew actions to exacerbate the consequences of error."*
> *(Reason, 1997).*

Even though this definition is fairly long it does have the advantage of highlighting several important aspects of the concepts of error management. In particular, it puts an emphasis on not only the activities following an error – namely error trapping and mitigation – but also stresses the importance of activities preceding the error – namely error avoidance.

1.3 ATC - a system description

Air Traffic Control (ATC) is an area which, for a number of reasons, can provide a proper context for studying and analysing human error events: (1) Very little work has been done on verification and validation of measures (such as error and recovery taxonomies) for ATC and, consequently, many findings have been difficult to sustain and interpret (Hopkin, 1995); (2) ATC is experiencing many new technological innovations which may induce new types of error and alter recovery opportunities (Wickens et al., 1997) and therefore require studies of various Human Factors solutions and associated error and recovery profiles; (3) ATC is known to be a high reliability organisation where human errors are rarely allowed to develop into critical situations. That is, in spite of the inevitably occurrence of human errors only a very low rate of loss of separation occurs which to a large extent is due to the controller's ability to develop and utilise behavioural skills to control such situations (Jones, 1997). This makes ATC a very suitable context to study human error and recovery. In the following is given a short description of the ATC system.

The overall goal of air traffic control is often described as ensuring a safe, expeditious and orderly flow of traffic at all times. Basically, this means that the challenge of air traffic control is to ensure that a safe separation is maintained (i.e. both between aircraft and between aircraft and other obstacles such as mountains and ground vehicles) and at the same time to do this in such a manner that the efficiency of the air-traffic system is not compromised. The control of the traffic is accomplished by three classes of controllers located at different facilities.

- Tower (TWR) control: In this facility the aircraft are handled on the taxiways and runways in relation to landings and take-offs. The departing aircraft are handed off to approach once it is airborne and the arriving aircraft are received from approach.
- Approach (APP) control[1]: In spite of its name this facility is responsible for both climb and descent of aircraft between high levels and ground. The approach controller is therefore responsible for receiving and handing off aircraft to the tower and en route control (i.e. ACC).
- Area Control Centre (ACC) control[2]: At this facility the control of high-level traffic is handled. The flights are here guided along a series of linear routes across the sky and at different flight levels. Often the en-route control is split between low level and high level.

There are many similarities in the tasks carried out at these different controller positions, but there are also variations in the cognitive demands that they place on the controllers. For example, the approach and the en route control rely extensively on symbolic representation of the traffic information (i.e. the radar and strips) whereas the tower control to a large extent is carried out on the basis of direct perception of the traffic information (Roske-Hofstrand & Murphy, 1998). Another difference is that the tower and approach control mainly requires tactical planning skills whereas en route control requires a combination of strategic and tactical planning skills.

In the figure below is described the main actors and tools involved in the en route control.

[1] Sometimes also referred to as terminal radar control or, in short, TRACON.

[2] Sometimes also referred to as Air Route Traffic Control Centre, in short ARTCC.

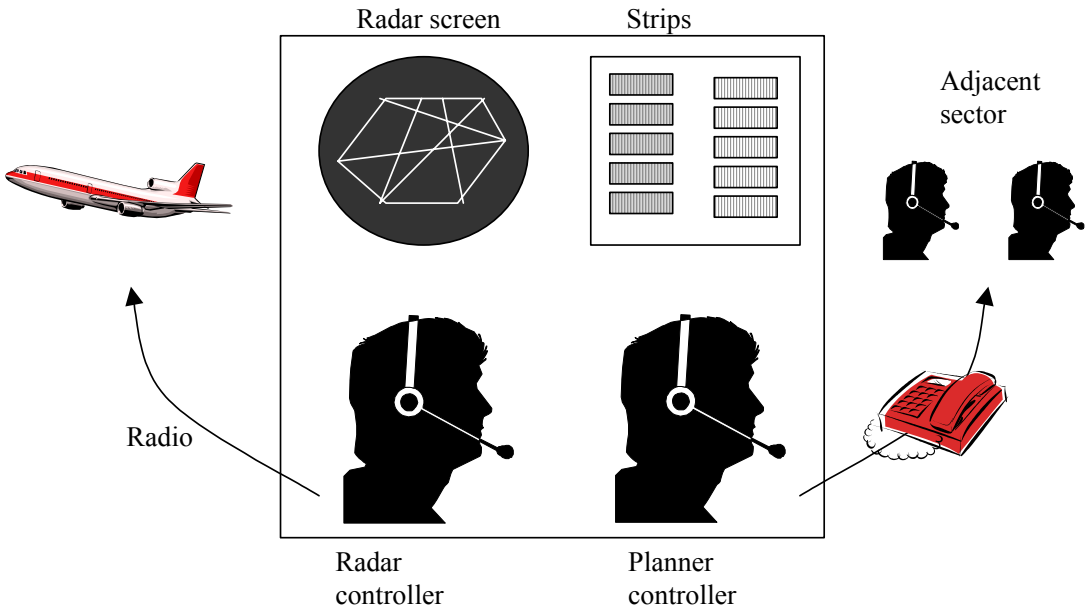


Figure 2: Schematic representation of actors and resources involved in en route ATC (Adapted from Rognin et al., 1998)

Each en route centre is divided into a series of irregular shaped sectors with both horizontal and vertical borders. For each sector two controllers - referred to as the radar controller and the planner controller[3] - can share the responsibility of providing guidance and controlling the aircraft (in the advent of low traffic several sectors may be collapsed and/or a single controller may be responsible for the whole sector or even several sectors). The controllers do not only have responsibility for the aircraft within the sector. They are also accountable for anticipating arriving traffic from other sectors and for informing neighbouring sector about aircraft leaving the sector. The exact distribution of tasks between the radar and planner controller may vary slightly from one country to another. The radar controller is responsible for monitoring the radar display to ensure safe separation and is in charge of the communication with the pilots. The planner controller organises the strips on the strip board and coordinates plans with other planner controllers from adjacent sectors.

There are several tools that are of critical importance for the controller. These are the radar display, the flight strips and the communication devices. The radar display is probably the most vital information source for the controller. Here the location of the aircraft inside or close to the sector borders can be seen and a data label for each aircraft gives information about the aircraft's callsign, altitude and ground speed. More detailed information pertaining to the individual aircraft can be found on the flight strips. This includes e.g. the call sign, flight level, departure and destination airport, flight route and aircraft type. These types of information play an important role for the strategic planning of the air traffic. The strips are also used to write instructions issued to aircraft and as a

[3] Sometimes the planner controller is also referred to as data controller.

memory augmentation in the case of unusual circumstances pertaining to an aircraft by annotating or cocking specific strips at odd angles. To accomplish the task of controlling the traffic it is necessary to communicate with other controllers and pilots which is normally done by telephone and radio (in the case where the controllers are located at the same site direct communication is normally used). The communication with the pilots is necessary to, for example, provide pilots with instructions and clearances. Coordination between controllers is an important activity in relation to handing over aircraft between different sectors. This may be between similar sectors but also between different types of sectors.

Even though it is often said that the radar is the most vital source of information this statement should be viewed with some modifications insofar as the strip is also a significant instrument for the controller. This is supported by the fact that it is possible to perform air traffic control with strips and no radar whereas it is extremely difficult to perform air traffic control with radar and no strips (Hughes et al., 1992). The importance of the strips is not only related to the detailed information they contain, but also the fact that the strips can be seen as a notepad or work site where changes in the state of the aircraft are noted from the time it enters the sector. In this manner the strips do not only provide information about the current and future state of the aircraft, but also historical information about what has been done and how the current situation has been reached.

Similar to many other safety critical areas, procedures related to both normal and abnormal circumstances play a significant role in regulating the controller's activities. There are, for example, procedures for the communication between controller and pilots and for maintaining the separation standards between the aircraft. Some of these procedures are internationally standardised by the International Civil Aviation Organisation (ICAO) whereas others are based on regional rules. In spite of the fact that many rules and procedures exist, it is in even official organisational documents specified that "nothing in these duties precludes a qualified controller from using his own discretion and initiative in any particular circumstances" (Rognin & Blanquart, 1999). That is, the procedures cannot be guaranteed to be complete or efficient in all situations and it will often be necessary to adapt these to the constraints of the situation.

1.4 Requirements to the taxonomy

A taxonomy is very similar to a categorisation system insofar as they both concern classification of phenomena into groups (or taxa). However, a taxonomy may be distinguished from a regular classification system by the requirement of being based on a sound theoretical basis (Fleishman & Quantance, 1984). Taxonomies are important as a foundation in any scientific endeavour. The reason for this is that it is necessary to agree on a common and unambiguous frame of reference to advance the understanding of the nature, origins and causes of a specific phenomenon of interest. Only by agreeing on a particular set of classifications can research results be compared and knowledge be accumulated. Therefore, taxonomic efforts make it easier to integrate results and contribute to the growth of a research field. Furthermore, it can be expected that the

development of an organising framework with a unified set of categories can be useful in relation to transferring research results to real-life operational situations.

In this project the goal is to develop an error management taxonomy to be used for studies in the area of air traffic control. Before commencing on this endeavour it is useful to decide on a set of criteria for developing and evaluating the framework. Unfortunately, there are no fixed objective criteria to evaluate the utility of a given taxonomy. However, Wiegmann and Shappell (2002) have suggested that for an error framework to be successful in this task it should be able to satisfy the following main requirements or product criteria: Reliability, Comprehensiveness, Diagnosticity, Usability and Validity. How these criteria relate to the development and evaluation of an error management framework will be elaborated in the following.

Reliability

For a taxonomy to be useful in the analysis of human errors and recoveries in research studies it is, first of all, critical that the framework is able to produce robust results. In this context there are several issues of relevance to consider:

- *Mutual exclusivity.* There should be an internal logic between the concepts within the framework to be able to achieve reliable (and useful) results. Therefore, particular attention should be paid to whether the concepts are mutually exclusive and that the taxonomy does not mix up principally independent issues[4].

- *Model-based approach.* The taxonomy may benefit from being derived from a model. This is related to the fact that the model-based approach may have certain advantages in relation to depicting the relationship between the individual components of a coherent framework (Isaac et al., 2002). Hereby it becomes easier to ensure a high level of internal consistency in the taxonomy and to achieve a high degree of mutual exclusivity between the individual categories if the taxonomy is derived from a model. Ultimately, this may increase the robustness of classifications made on the basis of the taxonomy and analyses of human-system studies may benefit from having an explicit frame of reference that the observed errors and recoveries can be related to.

- *Intra- and inter-rater reliability.* The same results should be achieved independently of where, when and by whom the classification is made. In particular, the two main types of reliability, namely inter- and intra-rater reliability (McGrath, 1994). Currently, very little research has been done to determine the reliability of error frameworks (but see e.g. Wiegmann & Shappell, 1997 and Isaac et al., 2000) and even less in relation to error management frameworks.

[4] As will be discussed later on the requirement concerning mutual exclusivity can be difficult to sustain in relation to the contextual factors that will be referred to as Performance Shaping Factors.

Comprehensiveness

For any given taxonomy it is important to consider whether the framework is able to cover all of the relevant variables that it purports to cover. In the current context this means that it should cover all the relevant categories related to the individual error management event and its surrounding context. It also important that the framework is able to analyse both normal and abnormal situations since important lessons about error management might be obtained by not only focusing on critical events, but also normal everyday events where most errors are prevented from developing into serious consequences (Helmreich et al., 2001; Maurino, 1999). Even though it is important that the framework is able to capture all relevant categories it is at the same time also important to avoid irrelevant categories (Wiegmann & Shappell, 2002). If there are many irrelevant categories it might jeopardise the framework because researchers and analysts have to spend unnecessary resources on reviewing irrelevant categories which can be a threat to reliability. The problem with irrelevant categories is also that the resulting database may contain too many "missing values". Thus, a framework can become too comprehensive.

Diagnosticity

Diagnosticity is perhaps the most crucial aspect of the framework and concerns its ability to move beyond analysing what happened to explaining why it happened. Use of the taxonomy for studies of human errors and recoveries should provide insight into their underlying causes so that potential error resistance strategies can be established. That is, the taxonomy should allow inferences about the specific causes of human error and recoveries in terms of generic psychological mechanisms as well as influencing contextual factors. The taxonomy can hereby be useful in the development and analysis of interventions to reduce the occurrence and consequences of human error.

- *Psychological basis*. Some classification systems can be used to organise directly observable features of human behaviour (such as omission, commission, repetition, etc.) whereas others concern inferences of psychological constructions or mental stages hypothesised to underlie observable human behaviour. The error management taxonomy should allow analyses of errors and recoveries in psychological and context-independent terms (and thereby move beyond the observable behaviour). The reason for this requirement is that it becomes possible to understand the underlying causes of errors and recoveries. This aspect is important because identical observable phenomena may be associated with different underlying causal mechanisms and, in similar vein, different observable phenomena may be associated with equivalent causal mechanisms. Consequently, an error management taxonomy based on underlying causal mechanisms may be most fruitful when it comes to tackling and mitigating the occurrence and consequences of human error (Reason, 1990).

- *Contextual factors.* Since errors and their capture do not happen in a vacuum but in the interaction between people and the general work environment – including

the technological, psychosocial and organisational context – it is critical that the framework is able to capture the dominant characteristics of the context that affects performance.

Diagnosticity also means that the framework should be able to generate insights that are in concordance with extant knowledge as well as being able to generate new insights. Later on we will review some hypotheses concerning expectations about how an error management framework should "behave" to be in concordance with existing research literature (see section 7.3).

Usability

Usability is concerned with the extent to which the framework can be applied to practical settings. Many conceptual frameworks are developed outside the applied setting and are never tested out in real situations. As a consequence of this it can often be difficult to apply it to complex real-life situations. Usability can be enhanced by avoiding subtle technical and psychological terminology and instead use more intuitively comprehensible concepts. This will increase the reliability of the framework and at the same time minimise the training requirements.

Validity

Validity is related to the degree to which a framework accurately reflects or assesses the specific concept that a researcher is attempting to measure. A pragmatic interpretation of the concept of validity of a conceptual framework is the extent to which it is able to satisfy the previously four described criteria. For example, content validity is directly related to the issue of comprehensiveness and relates to whether the framework adequately represents the variety and balance of the field it purports to examine. Face validity is closely related to usability and is concerned with whether the framework seems reasonable, using 'common sense', to people who might be using it. Finally, criterion validity is directly related to diagnosticity and is concerned with the framework's ability to provide insight to the underlying mechanisms of error and error management events. For this to be the case it is not only important that the framework is able to generate results that are in concordance with previous research and logical expectations, but also that it is able to uncover new and unknown insights.

There are trade-offs between the desiderata listed above. For example, the usability and the diagnosticity criteria may be in conflict (i.e. it may be difficult to achieve a high level of usability and at the same time have a high analytical power). Nonetheless, consideration to all of the above issues will be made in the development and evaluation of the error management framework.

1.5 Overview of the report

The goal of this report is to develop an error management taxonomy. Since most of the research within the area of human error has focused on error production and not error management, the main focus will be on error management. The specific content of the following chapters is:

Part Two: Literature Review

- Chapter 2: An extensive amount of research has been made in relation to understanding and systemising the mechanisms behind human errors. Some of the more prominent taxonomies will be reviewed and the most appropriate framework will be selected.

- Chapter 3: Since no off-the-shelf taxonomy exists to describe the error management process, it is necessary to develop such a taxonomy. As a starting point it is useful to review the existing research literature and on this basis make some preliminary suggestions about important distinctions. The review will focus on safety critical issues occurring both before and after errors. Before the occurrence of an error the focus is, in particular, on the issue of threat management which concerns how operational factors that have the potential of leading to errors and jeopardising safety are controlled. In relation to the phase after the occurrence of an error there are, in particular, four main issues that will be explored: *who* was involved in the detection and recovery of the error and/or its consequences; *when* was the error or its consequences detected; *how* was the error and/or its consequences detected and corrected; and finally *what* was the behavioural response and outcome*?*

- Chapter 4: Performance Shaping Factors (PSFs) can be seen as contextual factors that can have a positive or negative influence on the course of events. Some important frameworks relevant for the current context will be reviewed. The frameworks reviewed are characterised by being contextually relevant (i.e. from the aviation or the ATC domain) and/or encompass factors that can positively affect the error management process. In this manner it should be possible to also give an answer to the *why*-question – namely why did the error occur and why was it successfully or unsuccessfully managed?

- Chapter 5: Error management is not just a coincidence but is instead something that can be reinforced through different kinds of human factors initiatives. Both training and implementation of new technology concepts seem to be promising ways to strengthen the system's defences against human errors. In this chapter the focus will be on the training concept referred to as Team Resource Management and the design concept referred to as Interactive Critiquing.

Part Three: Construction of the Taxonomy

- Chapter 6: The error management framework will be described. It has been developed on the basis of the literature review and it has been further refined and tested on the basis of incident reports, critical incident interviews and simulator studies (the results of which will be described extensively in the later chapters). The framework is organised around an error management model. It consists of two main components: The core of the framework is developed on the basis of the literature review of error and error management taxonomies. The list of contextual factors is developed on the basis of the review of the Performance Shaping Factors.

Part Four: Validation

- Chapter 7: The methodological approach for evaluating the framework will be described. This includes a description of advantages and disadvantages of different kinds of approaches and some a priori defined hypotheses concerning how the framework is expected to behave (which will be used to establish the framework's criterion validity).

- Chapter 8-11: On the basis of the framework errors and their recoveries will be analysed by employing different kinds of data from the domain of ATC. The data from these analyses is used in the development, refinement and evaluation of the taxonomy. The first study is a pilot study based on Swedish incident reports (Chapter 8). The second is based on a simulator study where ATCO trainees carried out scenarios in a realistic setting (Chapter 9). The third is a questionnaire study where human factors experts could express their views concerning the relevance of both individual dimensions and the overall framework (Chapter 10). Finally, a comprehensive version of the framework is applied to the analysis of error events found cases elicited through the critical incident technique (Chapter 11).

- Chapter 12: On the basis of the empirical data and the predetermined evaluation criteria (reliability, comprehensiveness, diagnosticity and usability) conclusions about the framework will be made.

- Chapter 13: Finally, the results of the project are discussed and suggestions concerning future research in the area of error management are made.

PART TWO

LITERATURE REVIEW

2 Human error

The controller's task is highly cognitive in its nature and is dependent on mental processes. This means that most of the tasks are covert and cannot be directly observed on the basis of resulting behaviour. As a consequence, the taxonomies that will be reviewed in this chapter are cognitive in nature. As previously mentioned a taxonomy may be distinguished from an ordinary classification system by the requirement of being based on a sound theoretical basis. In compliance with this demand a framework will be selected that is suitable for the task of analysing human errors related to accidents and incidents. Since a vast amount of error taxonomies have been suggested, the focus will be on finding a taxonomy that maximises the chances of conducting robust analyses that ultimately provides a sound basis for reducing and mitigating the effects of human error.

A large part of the existing error taxonomies are grounded in an information-processing model (Wiegmann & Shappell, 1997). In general these models draw on the metaphor of the human as a computer and describe a number of mental processes or stages occurring from registration of stimuli from the environment by the use of sensory organs (eyes, ears, etc.) to the execution of a response (verbal or motor). Each of these stages is hypothesised to transform the stimuli, or information, and at each of these stages the transformation process may be in err or information may be lost. The exact amount of stages and their specific function may vary to some degree from one taxonomy to another. However, most of the models contain roughly equivalent information processing sequences and the main variation is the number of steps between the on-set of a stimulus event to the execution of a response. Some of the more prominent and influential frameworks are in this context a traditional Information Processing Model (see e.g. Wickens 1987, 1992), the Skills-Rules-Knowledge model (see e.g. Rasmussen 1982, 1983a) and The Model of Unsafe Acts (see e.g. Reason, 1990).

Since no extensive review and evaluation has been made of the individual frameworks that will be presented in the following it is not possible to determine the extent to which they are able to comply with the previous enlisted requirements on an objective basis. However, the frameworks will be evaluated on the basis of the extent to which they have some positive or negative attributes related to the previously identified criteria.

2.1 Skills-Rules-Knowledge Model

A natural point of departure in relation to understanding the mechanisms behind human error is the work done by Rasmussen insofar as his pioneering work, especially up through the eighties, has had a significant impact on how errors are conceptualised in the human factors literature. The contribution of Rasmussen lies both in the development of a human error taxonomy and in broadening the understanding of what is meant by the term human error. In the current context the focus will be to examine the skill-rule-knowledge (SRK) framework which has for sometime been considered a market standard when it comes to human error taxonomies. Originally this framework was developed on the basis

of verbal protocols of technicians engaged in electronic trouble shooting tasks (Rasmussen & Jensen, 1974).

An important notion behind the SRK framework is that human behaviour can be controlled at different levels of conscious control dependent on the degree of familiarity with the task and the environment (Rasmussen, 1983b). More specifically there are - as depicted in the figure below - three different levels of control, namely skill-, rule- and knowledge-based behaviour.

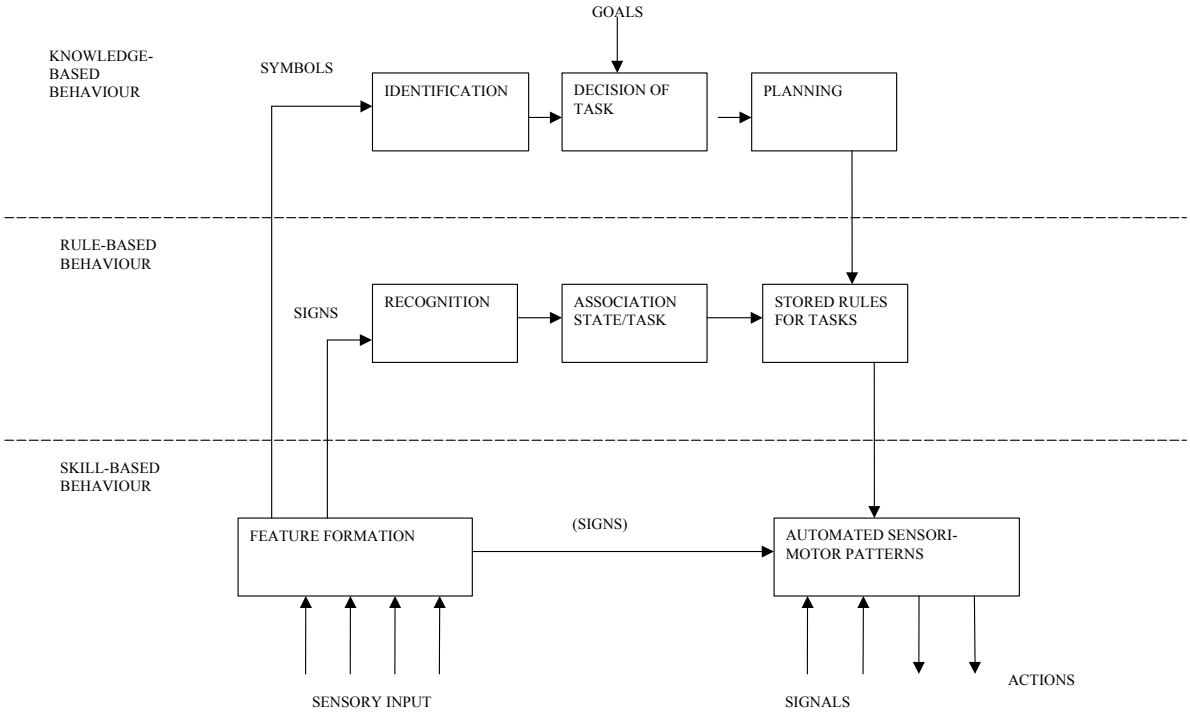


Figure 3: The Skill-Rules-Knowledge model.

At the skill-based level the behaviour is regulated by the lowest level of conscious involvement and is characteristic of highly routinised and automated activities. Such activities are mainly regulated by perceptual-motor systems of the human cognitive apparatus and by spatial-temporal information from the environment. The advantage of requiring few or no conscious resources is that the limited conscious resources can be used for other purposes such as planning ahead and monitoring past plans. Errors on this level are based on variability of force, space, or time coordination.

Rule-based behaviour is also a kind of behaviour that becomes activated in familiar work situations, but is distinguished from skill-based behaviour by requiring some degree of conscious involvement. The behaviour is controlled by stored rules either derived from experience or from other's know-how. Such pre-packaged rules can be activated on the basis of perceptually available information in the environment. It is important to note that the stored rules can be activated without necessarily having any understanding of the

functional properties of the environment. Errors at this level may be associated with erroneous classification of situation - and thereby application of the wrong rule - or incorrect recall of procedures.

When faced with an unfamiliar problem where no pre-packaged solutions are available it is necessary to move to the knowledge-based level of behaviour, which is the highest conceptual level. At this level a new plan has to be generated. This can be accomplished either through physically carrying out trial-and-error experiment on the environment, or through conceptual reasoning based on an understanding of the functional properties of the system being controlled (often referred to as the mental model of system) or a combination of both. Errors at the knowledge-based level of behaviour are associated with problem solving and goal selection tasks.

Conclusion	
Advantages	*Reliability.* An adaptation of the SRK framework has been used in relation to real-time studies of anaesthetist errors (Jensen, 1997). In this study the SRK framework was thoroughly evaluated and on the basis of several kinds of validity and the framework achieved ratings from satisfactory to high. The framework demonstrated inter-rater reliability Kappa values of 0.49 and 0.71 in naturalistic and simulated environments, respectively. A very similar version of Rasmussen's taxonomy has been applied to the analysis of human errors in an aviation accident database (Wiegmann & Shappell, 1997). The taxonomy could in this case accommodate well over 3/4 of the pilot causal factors contained in the database. The obtained inter-rater reliability Kappa values were as high as 0.935 which reflects an excellent level of agreement. *Diagnosticity.* The SRK-framework has been widely adopted in many contexts insofar as it provides the opportunity to gain insight into how human behaviour can be controlled at different levels of conscious control. Errors at these different qualitative levels might be controlled through different means and might also be detected by different kinds of mechanisms (Reason, 1990).
Disadvantages/ limitations	*Usability.* A problem is that it is not always easy to distinguish between the different levels of cognitive control – especially in environments that are less dominated by procedures such as air traffic control. *Comprehensiveness.* The framework does not distinguish between two important qualitatively different error types, namely slips and lapses (Sarter & Alexander, 2000).

2.2 The Model of Unsafe Acts

Reason is another prominent figure in relation to the scientific endeavour of studying the mechanisms behind human error. In particular, his book "Human Error" (1990) has had a tremendous influence on how the human factors community understands and analyses human errors. The significant contribution of his work lies in the identification of both basic human error types and the causal factors present in the system (so-called latent failures) before an accident sequence involving human errors actually begins - in short, the crucial role of organisations in industrial disasters. In the current context the focus will be on the Model of Unsafe Acts that can be used to examine the mechanisms behind individual unsafe acts.

The structure of the model is largely derived from Rasmussen's SRK framework. The model is shown below.

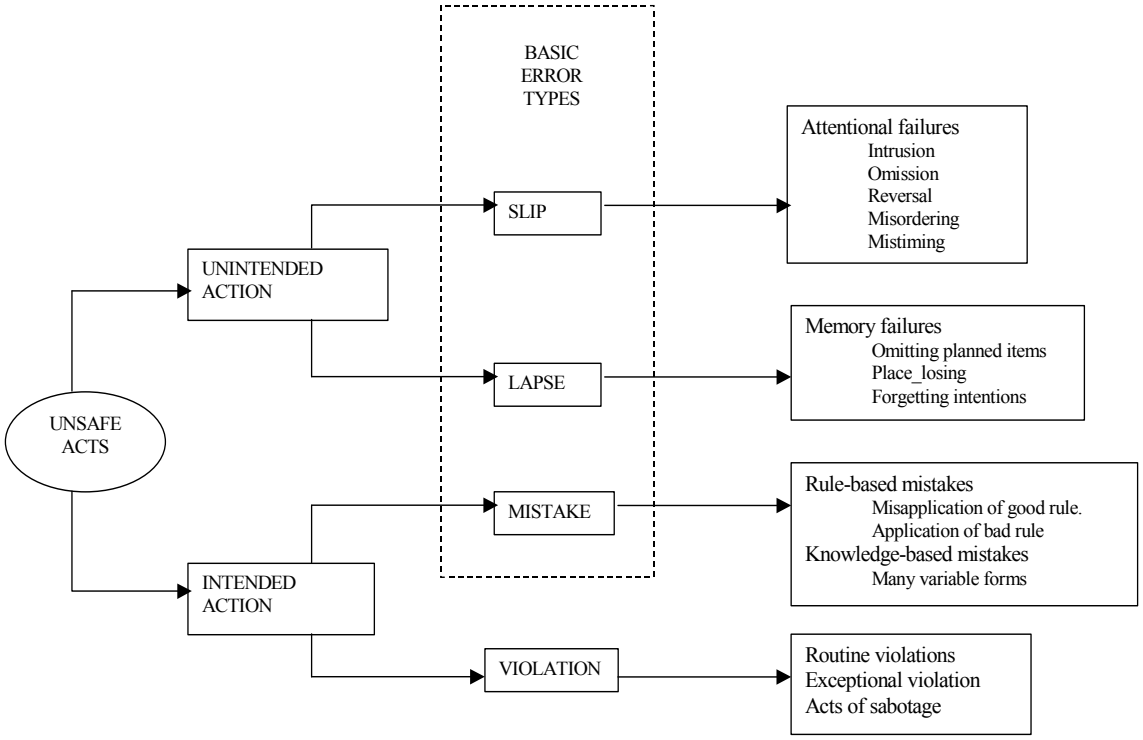


Figure 4: The model of unsafe acts.

According to the model there are two major groups of unsafe acts, namely intentional and not intentional. This distinction does not refer to the outcome of human activity (which in the case of human errors are by definition unintended) but is instead related to whether the actions are carried out as planned. The unintended actions can be manifested either as slips or lapses. Slips are monitoring or attentional errors where an action is planned but another one is carried out. These errors occur at the skill-based level of behaviour. Lapses are the other type of unintended actions and involve memory failures. Such

memory failures can manifest themselves through, for example, forgetting planned items or forgetting intentions.

The intentional part of the unsafe acts includes two main groups, namely mistakes and violations. Mistakes concern activities that run according to the plan, but where the plan is inadequate to achieve the desired goal. Such mistakes can be divided into two groups, namely rule-based and knowledge-based mistakes. Rule-based mistakes are associated with familiar situations where either a bad rule is applied to the situation or a rule that is perfectly adequate for certain circumstances is applied to situations that require a different set of actions. Knowledge-based mistakes, on the other hand, can occur in situations where no off-the-shelf solutions are available and a new plan has to be generated. In such situations slow, limited and effortful cognitive resources have to be applied to an often complex problem-solving situation.

Violation is a group of unsafe acts that concern deliberate deviations from rules, procedures or regulations. In the model three kinds of violations are listed. Routine violations are violations that happen on a regular basis and are perhaps reinforced by norms and values within a specific social context. Exceptional violations are those that only occur in rare and exceptional circumstances. Such violations may in particular occur in unusual situations where the existing procedures are not applicable to solving the current situation and where knowledge based reasoning therefore is required. The final group of violations is acts of sabotage and is separated from the other violations by the fact that the negative consequences are intended (and therefore normally not an issue for Human Factors research).

Conclusion	
Advantages	*Comprehensiveness.* In comparison with Rasmussen's framework Reason's model has the advantage of including violations as a group of unsafe acts. Furthermore, it distinguishes between slips and lapses at the skill-based level. *Diagnosticity.* Similar to the SRK-framework the model of unsafe acts has been applied in a lot of human factors research because it provides insight into the underlying mechanisms of intended and non-intended actions.
Disadvantages/ limitations	*Usability.* Even though the structure of the framework is somehow more intuitively comprehensible compared with the SRK-framework the distinction between cognitive levels – rule- and knowledge based level – is far from easily determined in less proceduralised environments such as ATC. *Reliability.* In Reason's model it is implied that an unsafe act is either a violation or an error, but in practice it can be both. That is, many violations are used as short-cuts to procedures that are considered unnecessary and inefficient. Such violations can be labelled "intentional non-compliance errors" to underline their dual property of being both an error and a violation (Helmreich et al., 2001). *Reliability.* Similar to Rasmussen's classification system it can be

	difficult to distinguish between the different levels of cognitive control.

2.3 Information Processing Model

Many models of human error are grounded in information processing theory. The information processing models are generated on the basis of synthesis of a number of results from an abundance of experimental studies aimed at studying specific aspects of the human "information processing machinery". One of the most well known models is the one proposed by Wickens (1992). The model describes critical stages of information processing in relation to a decision-making situation. This descriptive model may be useful to describe mental operations or stages that occur between the onset of a critical stimulus event in the environment and the response of the decision-maker. It is assumed that the different information processing stages in the model are characterised by transforming the input and that they demand some time for their operation.

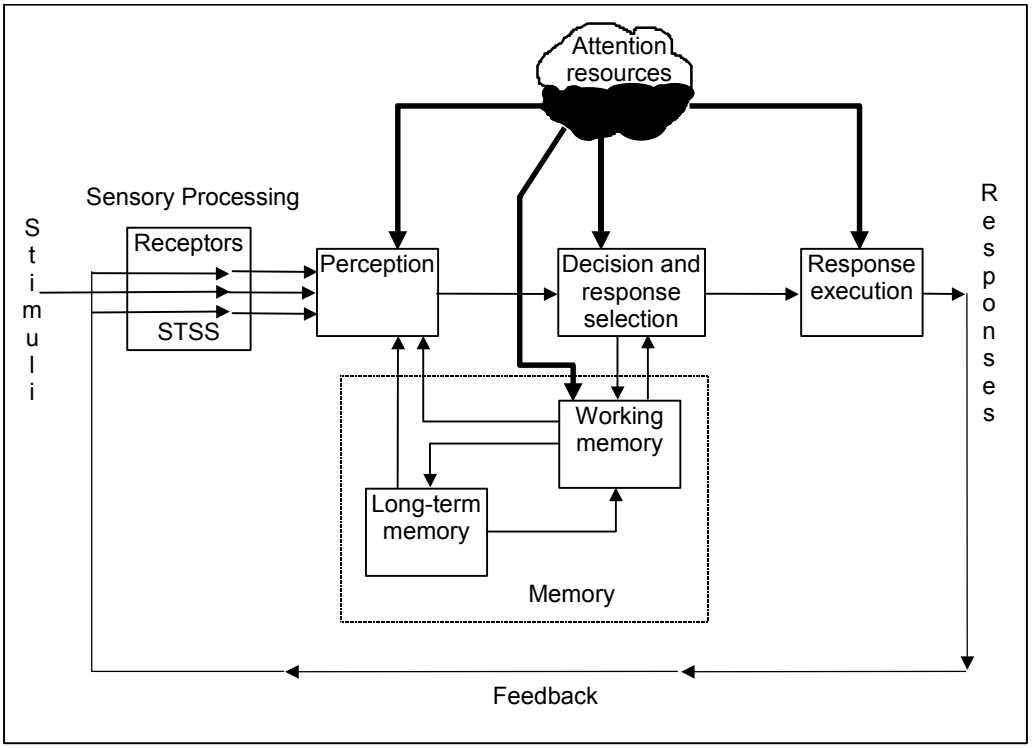


Figure 5: Wickens' model of human information processing

To get an overview of the model we will in the remainder of this section "fly over" the landscape of the model and explore some of its main characteristics. As a starting point we have a situation that can be characterised through a number of cues from different sources in the environment. The cues are initially processed through *short term sensory stores (STSS)* where the representation of physical cues are prolonged for a short period after the stimulus has physically terminated without requiring any conscious attention

(Wickens, 1992). Different kinds of sensory stores have been suggested for different sensory modalities such as echoic memory for auditory stimuli and iconic memory for visual stimuli.

Perception concerns integrating and assigning meaning to the physical cues. The most basic form of perception is detection which concerns determining whether a signal or target (such as an aircraft on a radar monitor) is present. More complicated processing is required if it is necessary to determine which class the target belongs to - a process referred to as recognition or identification. Classification of perceptual events includes making absolute and relative judgements. Absolute judgement concerns identifying a stimulus on the basis of its position along one or several stimulus dimensions (e.g. speed of an aircraft) whereas relative judgement concerns determining the relative difference between two or more stimuli (e.g. which aircraft has the highest speed).

After having assigned meaning to the physical cues a choice of action should follow. This occurs in the stage referred to as *decision-making and response selection*. Problems may occur if the cues are absent, vague, ambivalent or conflicting. Furthermore, problems may occur if conflicting goals are present. If the decision situation is new and unanticipated situation most of the active processing related to understanding the situation and making decisions have to be carried out in the *working memory*. This may be associated with some problems because working memory is characterised by being temporary, fragile and limited. The limitations of working memory may be circumvented if relevant *long-term memory* structures exist. This may lead to recognition-primed decision-making which refers to a relatively rapid and automatic process whereby experts make decisions based on recognition of similar situations in the past (Klein, 1989).

The decision to initiate the chosen response is separated from its execution in the model. That is, when the decision is made, it has to be carried out. This phase is in the figure denoted *response execution*. In this phase errors are typically associated with problems of automaticity which refers to the fact that people are able to carry out highly practised action sequences with few or no attentional resources and such activities are associated with a specific type of error, namely slips. The outcome of the decision can function as a basis for further pick-up of cues and decision-making. In addition, the outcome can also function as the basis for decision-making in the future by being stored in the long-term memory. In this manner there is a direct link between decisions made in the past and the decisions made in the present.

The stages of perception, decision-making and response selection and response execution are, as illustrated in the figure, largely dependent on the available attention resources. It is hypothesised that there exist a limited amount of attentional resources that can be distributed among the different cognitive activities. Four different kinds of problems may be associated with the attention (Sanders & McCormick, 1992). (1) *Selective attention* is a characteristic of those situations where several sources of information should be monitored for the occurrence of specific events. Problems may be associated with *selective attention* when people have to choose which aspects of the environment to direct their attention at (i.e. a top-down process). (2) *Focused attention* concerns those

activities where a source of information should be attended to and other sources excluded. Therefore, if some processes require a lot of resources only limited resources will be available for the remaining processes. Attentional problems may be associated with focused attention because people's attention is often drawn to the most salient stimuli in the environment that may not necessarily be the most important ones (i.e. a bottom-up process). (3) *Divided attention* is required where two or more tasks should be carried out simultaneously and attention must be paid to both. Due to the limited pool of cognitive resources there is a risk that one or several tasks receive insufficient resources for successful performance. (4) *Sustained attention* concerns detecting a signal over prolonged periods of monitoring time. Vigilance decrements have been observed to occur as a result of sustained attention over prolonged periods of time with the results that speed and accuracy in signal detection is reduced.

Conclusion	
Advantages	*Reliability.* In a study by Wiegmann & Shappell (1997) where the information-processing model was used in the analysis of an aviation accident database a Kappa index of 0.660 was achieved which is considered "good" by conventional standards. *Usability.* The model has the advantage of corresponding conceptually very well with the controller's task. It might therefore be easily applied to the domain of Air Traffic Control. *Diagnosticity.* Breakdowns in the cognitive processing at the different stages in the model might be associated with different kinds of remedies (actually several human factors books such as Wickens (1992) are dedicated to this issue). *Comprehensiveness.* All the main cognitive error types seem to be covered by the information-processing model.
Disadvantages/ limitations	*Usability.* Some parts of the framework might appear too theoretical and not applicable to practical settings. This, in particular, is the case for the sensory processing part of the framework. That this category is less relevant in practical settings is, for example, supported by a study of Wiegmann & Shappell (1997) of pilot errors where less than 3 % of the errors fell within this category. Furthermore, in another study by McCoy & Funk (1991) of ATC operator errors based on a modified version of Wickens' model the sensory processing part was not included in the analysis.

2.4 HERA

HERA is acronym of Human Error Reduction in ATM. It is a comprehensive technique that has been developed to analyse the mechanisms and circumstances behind human errors in the area of air traffic management. The HERA technique contains one of the most elaborated and detailed taxonomies for error analysis and has at the same time been specifically adapted to the analysis of human errors in ATM. The framework has been developed on the basis of a review of both academic and industrial research of the past five decades (Isaac et al., 2002; Andersen & Bove, 2001).

At the core of the HERA technique is a slightly modified version of Wickens' (1992) model of human information processing. Below is shown a depiction of the underlying model:

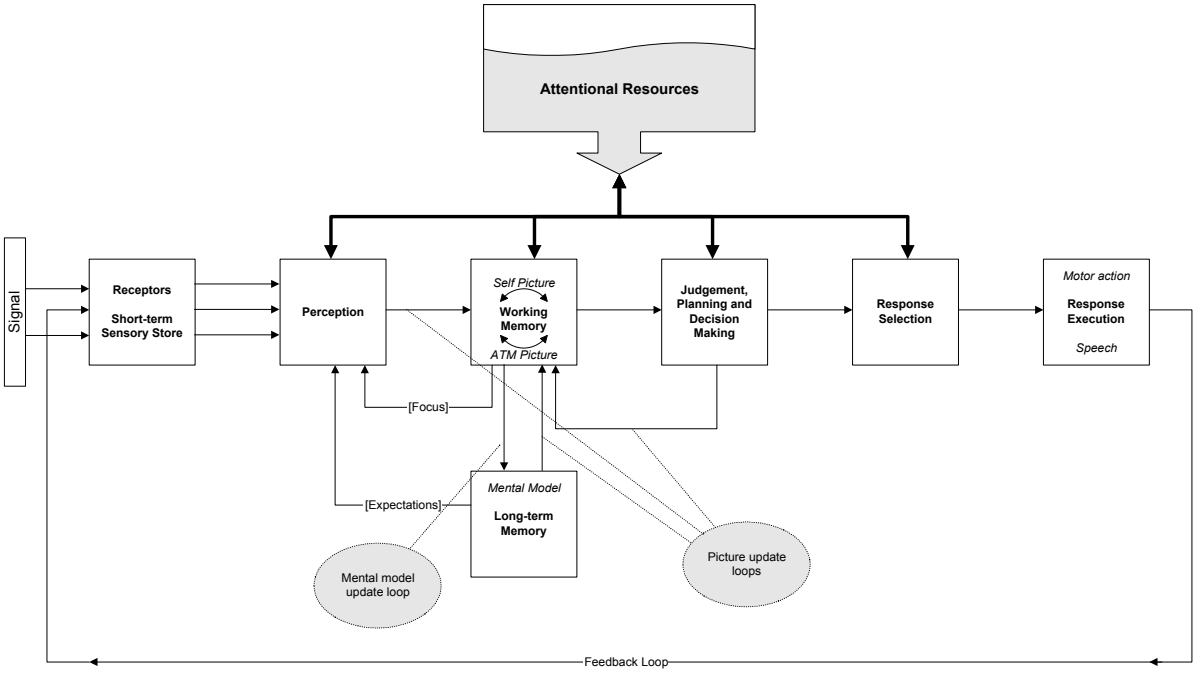


Figure 6: The HERA model

The model contains five different cognitive domains - each which may be associated with qualitatively unique errors. To get a better understanding of how errors can be associated with each of these cognitive domains it is useful to review some practical examples:

- Perception and vigilance: This cognitive domain concerns issues related to receiving and understanding information. A typical kind of error associated with this domain is hearback error. That is, a controller fails to pay attention to the content of a pilot's read back and hears what he/she expects to hear.

- Working memory: This domain concerns the short-term storage of information. For example, a controller may forget to carry out tasks necessary to ensure continued safe separation between aircraft in spite of having intended to do so.
- Long-term memory: The long-term memory contains more permanent information based on the person's training and experience. A typical error associated with this domain is if the controller recalls a procedure incorrectly because, for example, the procedure is rarely used or has not been used recently.
- Judgement, planning and decision-making: Controllers are constantly required to make projections of trajectories, plan future actions and to make decisions. These activities may all be associated with errors. For example, the controller may mis-project the future position of two aircraft and consequently not consider any need to monitor them further.
- Response execution: Sometimes people carry out actions that they have not intended. A well-known example is when a controller gives a clearance to one flight level but had intended to give clearance to another flight level. This is often referred to as slip of the tongue.

Analysis of human errors is within the HERA technique structured around these different cognitive domains. In the following is briefly described how each individual error can be analysed on the basis of factors associated with the individual act and outside factors. In essence, the way of analysing human errors within the HERA technique is to a large extent inspired by Rasmussen's (1982) multi-facet taxonomy where the analysis is not stopped at some inappropriate or undesirable behaviour but is continued to an investigation of what caused the human to act as he or she did.

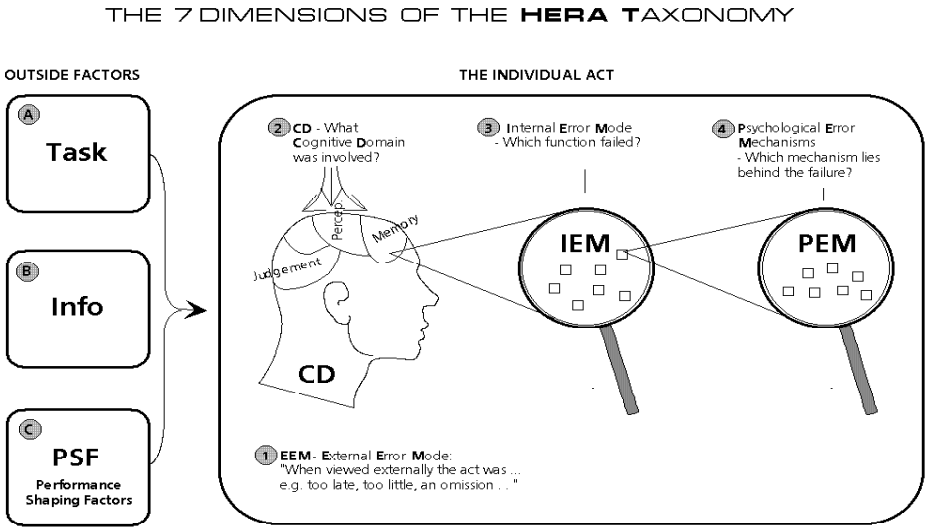


Figure 7: The main dimensions of the HERA taxonomy

2.4.1 The Individual Act

On the basis of the expanded information-processing model each error can be analysed on the basis of four levels of detail:

- External Error Modes (EEM) - the external manifestation of the error (e.g. omission).
- Cognitive Domains (CD) - to be able to describe the error in more detail it is necessary to have some knowledge about the cognitive function that failed (e.g. perception and vigilance). This level of description is based on the cognitive model.
- Internal Error Modes (IEM) - the internal manifestation of the error within each cognitive domain (e.g. late detection).
- Psychological Error Mechanisms (PEM) - the internal mechanism of the error within each cognitive domain (e.g. perceptual tunnelling).

2.4.2 Outside Factors

There are three dimensions associated with the outside factors:

- Task: *What was the controller doing while the error occurred* (e.g. radar monitoring or strip work)
- Information/topic: *What kind of information was associated with the error* (e.g. flight level, heading)
- Performance shaping factors (PSFs) - *What factors may have enhanced the chances of the error?* Often there are factors in the environment that provoke or enhance the risk of errors. Such factors are often referred to as performance shaping factors (PSFs). In this manner some of the blame is moved away from the individual to the broader context in which he/she is embedded. PSFs are also good at pinpointing Human Factors aspects of the ATM environment that contributed to the occurrence of the error (and thereby could be altered to counteract the occurrence of similar errors in the future).

Conclusion	
Advantages	*Reliability.* Studies have been undertaken to evaluate the precursor of the HERA system (TRACEr) and the final version of HERA. In the evaluation of the precursor of HERA the classification scheme could account for about 98% of the identified air traffic control errors in British incident reports and both a good level of inter-analyst agreement and user opinion was revealed (Shorrock et al., 1998). Later evaluations of the HERA system have also revealed a respectable and highly significant level of agreement in relation to the cognitive domains (Isaac et al., 2000). *Diagnosticity.* Insofar as the framework has been developed on the

	basis of a symbiosis of a robust theoretical background and analysis of error events in a huge database of authentic ATC incident reports the focus has been directed towards uncovering the underlying mechanisms and contextual factors.
Disadvantages/ limitations	*Comprehensiveness.* The level of detail in the framework may be too ambitious. This is problematic, first of all, because it is very difficult to make the finer grained distinctions when analysing human behaviour in complex and realistic settings. Furthermore, it will require a very huge database to be able to derive any meaningful statistical information concerning the more detailed cognitive categories. *Reliability.* In air traffic management it may be difficult to distinguish between short-term and long-term memory. Traditionally, short-term memory can only contain a limited amount of information for a very short duration (about 10 to 15 seconds) and can therefore only be used for information that should be immediately recalled. Long-term memory, on the other hand, does not seem to have any constraints in relation to capacity and storage time. In air traffic management most of the task related to maintaining and updating the picture requires remembering information for a duration of 10 to 15 minutes and does thereby seem to lie somewhere in between the short-term and long-term domain (Hopkin, 1995). *Diagnosticity.* It is difficult to see the usefulness of categories of External Error Modes.

2.5 Conclusion

In the previous sections the state-of-the-art of error taxonomies were reviewed. Few attempts have been made to apply the taxonomies to the analysis of errors in complex domains and it is therefore difficult to make firm conclusions about their relative usefulness. Nonetheless, studies indicate that the reviewed taxonomies can accommodate a large part of the observed errors and that reliable classifications can be obtained with these taxonomies. In spite of the fact that the reviewed taxonomies have been relatively successful on the quantitative level (being able to describe most of the identified errors in the reliable manner) there may be some reasons why an information processing model would be the most appropriate framework to analyse human errors in ATM:

- It may be difficult to determine which processing level in Rasmussen's and Reason's models an error occurred. For example, in many ATM tasks the situation is not completely new or completely old and in such situations it may be difficult to determine whether a given error was a rule- or knowledge-based mistake.

- The stages in the information-processing model are frequently mentioned in the ATM literature. This may be a reflection of the fact that errors that occur in ATM, such as hearback errors, visual misidentifications or decision/planning errors, seem to

be most compatible with the information-processing framework. This compatibility is also reflected in the fact that only the reviewed information processing models have been applied to the domain of ATM.

Clearly the HERA technique contains the most detailed taxonomy in itself insofar as the main stages of Wickens' information processing model have been extensively elaborated on the basis of the state-of-art knowledge within the area of human error (including the other human error frameworks previously described). Furthermore, the taxonomy has been specifically adapted to the ATM environment – and been thoroughly validated within this environment - and therefore seems to be a good platform on which to base error analyses.

3 Error management

Throughout the last couple of decades an abundance of research has emerged in relation to understanding the nature of human error and the cognitive mechanisms underlying the production of a variety of errors. An important insight from these studies is that human error is the flip side of human performance and that it is impossible to completely eliminate them (Rasmussen, 1987; Reason, 1990). Furthermore, it has been acknowledged that there are limits to automatic kinds of error detection devices insofar as such machines do not have access to the goals underlying the behaviour (Frese, 1991). Therefore, it becomes important to understand how people manage errors committed by themselves or others. Nonetheless, the understanding of how errors are detected and recovered has failed to keep pace with the understanding of the mechanisms underlying human error. This process following error production - often referred to as error management or error handling - will be the focus of this section.

An increased understanding of the error management process is a prerequisite to improve safety and reliability. A modest, but gradually increasing, amount of studies have emerged concerning the error management process. Some of these have been related to very specific tasks - such as reading (Carpenter & Daneman, 1981) and writing (Hayes & Flowers, 1980) – and been conducted in laboratory settings. Other studies of the process following errors have focused on more complex and realistic tasks such as power plants (Woods, 1984), human computer interaction (Brodbeck et al., 1993), aviation (Wioland & Amalberti, 1996; Sarter & Alexander, 2000; Helmreich, 1999), hospitals (Edmonson, 1996; Cooper et al., 1982), air traffic control (Wioland & Amalberti, 1998), the maritime domain (Seifert & Hutchins, 1994) and everyday tasks (Sellen, 1994). A general insight from these studies from different domains is that people have powerful capabilities for controlling errors committed by themselves or others. Unfortunately, many of the studies have been done by using different conceptual frameworks and therefore it can be difficult to integrate their findings.

The goal of the following sections is to provide a review of the status of existing knowledge and important issues that should be considered in the development of an error management framework.

- In the first section considerations are made concerning the development and content of the taxonomy. These considerations constitute the foundation for the decisions made in the first phase of the taxonomy development.

- An error management model can be used as an organising principle in the development of the taxonomy. Therefore, some of the most promising models will be presented and reviewed.

- The concepts of risk and threat management are introduced. Risk management can be used to describe the dynamic interaction between error production, detection and recovery in dynamic environments which may not be as straightforward as expected

and may be dependent on factors such as workload, meta-knowledge and confidence. Threat management concerns being aware of important factors in the operational environment and dealing effectively with these before they result in errors.

- The taxonomy should be based on a model of human error management and be able to answer four fundamental questions: *who* was involved in the detection and recovery of the error and/or its consequences; *when* was the error or its consequences detected; *how* was the error and/or its consequences detected and corrected; and finally *what* was the behavioural response and outcome*?* These issues will be examined in detail.

3.1 Taxonomic considerations

Before examining specific error management models and taxonomies it is useful to decide on generic principles that should guide the development and structure of the framework. In the following is a review of the main issues considered before starting on the development of the taxonomy (the issues are not presented in any particular order of importance).

3.1.1 The level of analysis

Error management is an ongoing task that can be accomplished at different levels within an organisation. At the highest level are the high level managers who have the overall responsibility of setting and achieving system goals. These managers also play an important role in the error culture and thereby the chances of long-term learning and improvement. A series of line management departments such as training, maintenance, personnel and safety have the responsibility for implementing the strategies of the management. Each of these line departments may contribute to defences to avoid failures and their potential negative consequences. The last line of defence is the front-line staff - namely the controllers - that has the daily responsibility for minimising loss of separation events and for the recovery of failures. In the current context the main focus is on modelling and classifying how the front-line staff controls errors. Nonetheless, attempts at also describing the influence of the broader context on individual performance will be made by the use of the so-called Performance Shaping Factors - as will be described later on.

3.1.2 Error production vs. error management taxonomy

The way human errors and error management are classified and described may involve both similarities and differences, when it comes to the requirements of the taxonomies. Both types of taxonomies should make it possible to pinpoint important mechanisms involved in a specific human performance event and they should be based on a model that contains the important information processing stages behind human performance. An

important difference, however, may be found in which components should be described. When describing human error the goal is to determine at which stage the information processing failed. In contrast, when describing the error handling it becomes relevant to describe the whole process of successful/unsuccessful performance (i.e. from an error production to an error discovery/recovery phase). In order to describe the successful performance we need to examine each of the stages hypothesised to underlie the error handling process and their unique characteristics. On this basis it becomes possible to map the "cognitive route" of the error-handling task. The fact that an error handling model should allow a coherent description of the whole error handling process sets some limitations on how much detail it should contain with regards to the numbers of stages and their potential transitions. By using a simplistic model it becomes easier to compare different error recovery routes. Furthermore, to start out with a simple model is seemingly in good concordance with the current level of knowledge in relation to the error handling process and to suggest a more elaborate model would probably be premature.

3.1.3 Linear vs. circular model

An important issue to consider when developing a model of human error management, and as a corollary of this a procedure for studying error management events, is whether it should be a linear or a circular model. A linear model will require that a fixed sequence of stages can be identified for each error and error management event. Alternatively, a circular model will allow a more free flow between individual stages - i.e. some stages may be omitted and some stages might be repeated. Both the linear and the circular model for describing human error management may be associated with strengths and weaknesses. The circular model has the advantage of being able to describe in a precise way how the recovery events actually unfold during a scenario (see e.g. Kanse and Van der Schaaf, 2000b or section 3.2.1). The approach may, however, have a more restricted practical value insofar as a flexible structure in the sequence of stages will compromise the possibilities for aggregating data across scenarios and studying interactions between stages. Since these characteristics are rather significant with regards to the practical usefulness of the framework, it was chosen to accept a rigid linear description of the flow of events.

3.1.4 Errors vs. detection/recovery failures

A problematic area when analysing error and recoveries within a single framework is to make a clear distinction between errors and unsuccessful detections/recoveries. An example where errors and detection/recovery failures are mixed together is shown below.

ATC Example
Just after the phone call from Stockholm R3 changed the clearance to SAS 483. ***The intention was that SAS 483 should have clearance to FL 260, but instead R3 said "recleared flight level 270"*** *which SAS read back correctly. This may be related to a high stress level due to a high traffic*

load. ***Neither R3 nor D3 reacted to the "erroneous" flight level when SAS read the clearance back*** *[Source: Swedish CAA Report No. 970604].*

In such situations it is necessary to determine whether these missed windows of opportunity should be analysed under the heading of error or error management, or both? It is difficult to make formal distinctions between error and error management phases. Consequently, it is necessary to make a choice between two possible solutions: (1) a detection/recovery failure will only be analysed as an error; or (2) a detection/recovery failure will be both analysed as a part of the error management process and also as a separate error. In the current context the latter solution is chosen because a detection/recovery failure is clearly both an error and an integrated part of the error management process.

3.1.5 Procedural violations

As previously discussed it is not entirely clear whether intentional procedural violations should be conceptualised as a subgroup of errors or not. This issue does also have some implications for the error management analysis because intentional violations are characterised by the fact that they are volitional - that is, the actor knows beforehand that the action or inaction is strictly speaking wrong. An example of this is shown below:

ATC Example
According to the procedures R6 (and instructor) should have opened position R8 at 8.15, but this was not done at the specified time. ***The controller made here a deliberate choice to postpone the opening because of low traffic.*** *[Source: Swedish CAA Report No. 970826]*

In such situations it may seem a bit artificial to speak about detection and correction, because such errors are intentional and therefore less likely to elicit a management response (Klinect et al., 1999). Nonetheless, it was decided that intentional procedural violations should also be included as a subgroup of errors in the error management analysis insofar as they constitute an important group of decision-making failures.

3.1.6 Top-down vs. bottom-up approach

The development of an error management taxonomy may be accomplished on the basis of a top-down approach (i.e. theory-driven), a bottom-up approach (i.e. data-driven) or a combination of these two approaches. In the current context we will start out with a top-down approach and review taxonomies relevant in relation to understanding the error handling process. On the basis of this literature review a preliminary classification scheme will be developed. Since very little research in general has been done in relation to development of taxonomies to describe the error handling process, even less work has been done in relation to examining the error handling process in safety critical domains

such as air traffic control. A consequence of this is that we will have to rely – to some extent – on work/studies from environments that do not share the characteristics of safety and time critical domains. Nonetheless, when exploring and examining the usefulness of different classification schemes we will consider their relevance and applicability to the task at hand. Furthermore, to enhance the chances of their applicability we will focus on studies that have been done in relation to realistic tasks and that consequently have a satisfactory level of ecological validity.

3.2 Error management models

As described earlier on it is very important that the classification system is based on a coherent model. In the following some error management models will be reviewed to establish the most appropriate framework to use in the development of an error management taxonomy.

3.2.1 Failure compensation process model

On the basis of insights obtained through a literature survey and analysis of incidents at a chemical process plant Kanse & Van der Schaaf (2000a, 2000b) have developed a preliminary, general failure compensation model. The model is shown in the figure below:

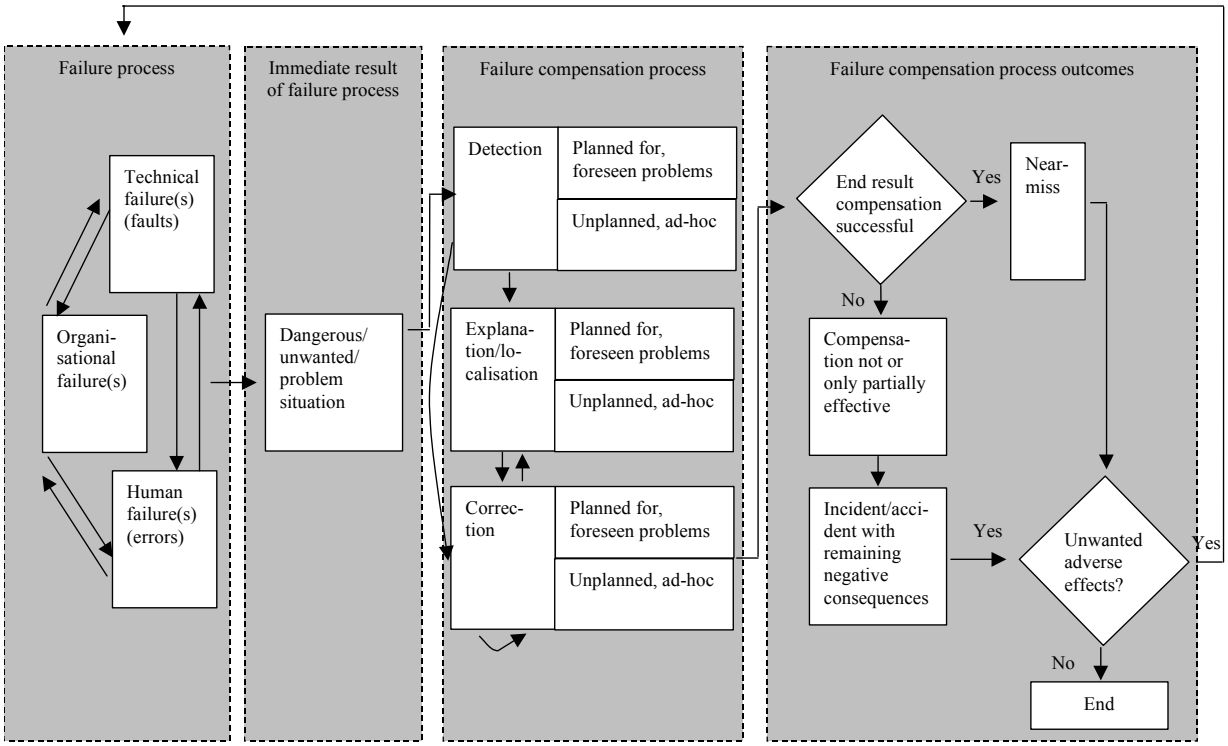


Figure 8: Failure compensation process model

The model is one of the first attempts at depicting the recovery process in details by describing the events beginning with a failure process (either caused by human error, organisational failure or technical failure) leading to a dangerous outcome that is followed by more or less successful compensation behaviours. Of particular interest are the last two boxers in the model, namely failure compensation process and outcome. Inspired by other researchers (e.g. Zapf & Reason, 1994) Kanse & Van der Schaaf distinguish between three different error handling process stages:

- *Error detection*: Discovering or suspecting that an error has occurred without exact knowledge concerning the nature of the error.
- *Error explanation/localisation*: Obtaining knowledge concerning the source of the breakdown. That is, the person knows how the error came about and knows why the error has occurred.
- *Recovery planning*: Initiating a problem solution process to resolve the problem.

It is of interest to note that the transition between the three error handling processes or stages is not simply sequential, but is instead more flexible with regards to the individual transitions. The model suggests a non-rigid flow from error detection to error recovery and some stages may be either omitted or repeated. For example, as part of the planning and problem-solving process a person may return to error diagnosis to obtain more information concerning the underlying error (Orasanu and Fischer, 1997).

The last box in the figure - the failure compensation process outcomes - is concerned with whether the failure compensation process was successful (i.e. a near-miss) or not (i.e. incident/accident). These issues associated with the outcome will be further addressed when reviewing classifications associated with the "what"-question.

Conclusion	
Advantages	*Diagnosticity.* The advantage of the model is that it describes in a rough but intuitive appealing manner the main stages of the error management process. Also, it gives a depiction of the potential interaction between stages and thereby provides an understanding of the potential complexity associated with fault management. *Usability.* The overall structure of the model seems logical and easy to understand.
Disadvantages/ limitations	*Comprehensiveness.* The model does not seem to be applicable to errors that do not lead to unwanted situations. Consequently, the large majority of errors that are caught before any consequences have occurred cannot be analysed by this model. *Comprehensiveness.* Little is known concerning the error identification phase and the studies that exist indicate that the root cause of a problem is rarely sought in high-risk environments (Kontogiannis, 1999). This may, in particular, be the case in air traffic management where problem solving rarely requires insight into the problem's genesis. The exact cause of operational aberrations is often

	discovered post-hoc, if at all, when e.g. listening to radio recordings of the occurrence or talking with the involved pilots. That is, the error localisation happens some time after the resolvement of the problem. *Reliability.* Currently, no data are available to evaluate the reliability. However, it can be speculated that the many transitions in the model might reduce its reliability.

3.2.2 The model of threat and error management

Helmreich et al. (1999) have developed a model of error management on the basis of data about flight crew behaviour and situational factors on normal flight. A slightly modified version of this model is presented below (that is, all terms related to flight crew have been altered to ATCO).

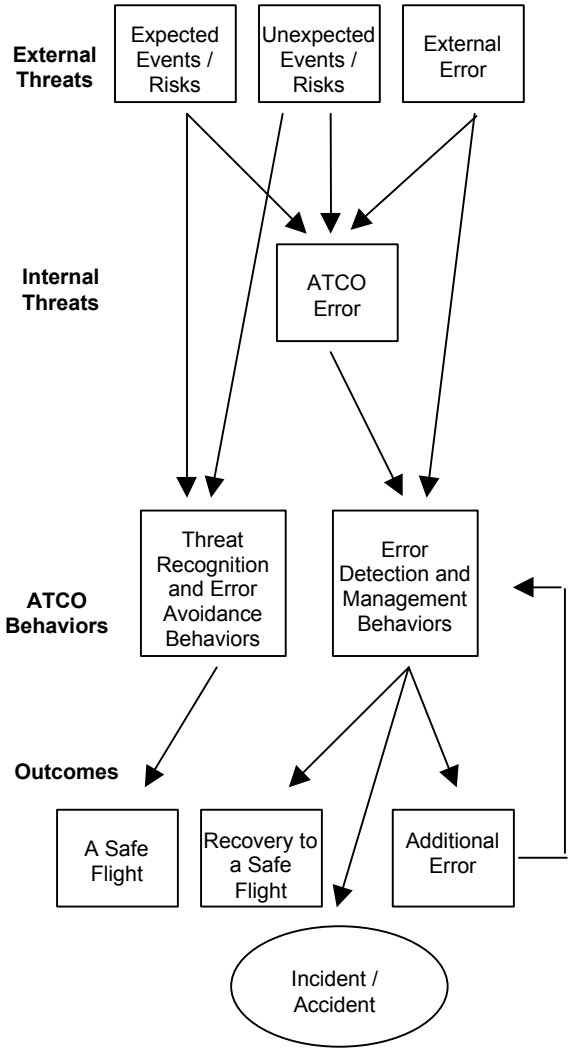


Figure 9: **The model of threat and error management**

According to the model safety risks may come from either expected or unexpected threats (in a previous section referred to as performance shaping factors). Expected events include predicted weather conditions or airport conditions. Unexpected events, on the other hand, include aircraft on wrong flight level, changing weather and equipment malfunctions. External errors include all errors caused by non-ATCO people such as pilots and maintenance crew. If these different kinds of threats are recognised at an early point in time there is a chance to counteract them and ensure a safe flight. If, on the other hand, the threats are not recognised they may lead to an ATCO error. If an ATCO error should occur this may lead to different kinds of error detection and management behaviours. The result of these could either be a recovery to a safe flight, an incident/accident or even additional errors.

Conclusion	
Advantages	*Diagnosticity.* The chief benefit of the model of threat and error management is that it provides a description of the main stages of threats, errors and error management. In this manner error management is placed within a larger context of human behaviour. *Comprehensiveness.* The model is originally derived empirically from observations of flight crew behaviour in line operations (e.g. Klinect et al., 1999), but has also been applied to the analysis of incidents and accidents (e.g. Jones & Tesmer, 1999). In this manner the framework has been useful in the study of human errors and their management in both normal and abnormal conditions. This means that the framework is able to deal with both successful and unsuccessful behaviour. *Comprehensiveness.* The model is unique insofar as it is the only model that incorporates threats as an integrated part of the model. This is an issue that has not previously been emphasised in any other model of error and error management. *Usability.* The model provides an intuitively logical structure to understand the error management process. Furthermore, the concepts do not require any theoretical background and should therefore be easy to understand.
Disadvantages/ limitations	*Reliability.* It is interesting to note that the model distinguishes between error management at the error and the outcome stage. However, in relation to the reliability of classifications made by the use of this framework it may introduce some problems because any disagreement concerning the stage at which the error was detected will also compromise the classifications related to the response and outcome of the error. Therefore, it seems more desirable that the classification of the stage is separated from the classification of the response and outcome of the error. *Diagnosticity.* The classifications included in the model are only of behaviours (i.e. the phenotypical level) and outcomes and not of the underlying cognitive processes (i.e. the genotypical level). That is, the classifications allow a description of *what* happened but not *how* it

	happened. Therefore, the framework should be supplemented with other taxonomies to describe the underlying processes. *Comprehensiveness*. In the model it is implied that threat recognition and error avoidance will necessarily lead to a safe flight. That is, it is possible to be aware of threats without making the necessary preparations if, for example, the risk is underestimated. It is also possible that in spite of making reasonable initiatives to avoid a threat it is not necessarily the case that it is successful (e.g. if a pilot does not comply with ATCO instructions aimed at avoiding the threat).

3.3 Threat and risk management

In this section we will examine two concepts that are essential in relation to expanding the traditional understanding of the role of the human operator in relation to complex systems that has had a tendency to be negatively biased. The human operator is often viewed as a potential weak and vulnerable component of the man-machine system that has a tendency to commit errors. The concepts of threat and risk management are important because they help highlight the positive contribution of the human operator and can therefore be significant to give a more balanced picture. The important distinction between the two concepts is that threat management is about prohibiting that operational threats develop into errors. Risk management, on the other hand, is less concerned with the avoidance of errors as such but is to a larger extent concerned with how the errors committed are kept under control and prohibited from developing into critical situations. In this manner the two concepts complement each other and it seems reasonable to cover these two issues within the same section.

3.3.1 Risk management

Error and error management can be conceptualised as an integrated part of a more global concept, namely risk management. This is supported by a consensus among several studies that show that people as a part of their expertise develop natural abilities to control risk-taking and that they develop protections and defences against their own cognitive deficiencies (Amalberti & Wioland, 1997). Consequently, the determining principle regulating behaviour is not exclusively based on avoiding errors, but is instead based on meta-knowledge and confidence concerning being able to control the situation.

A good example of how risk control and human error is related is the issue of cognitive resource management. It is commonly accepted that the way workload affects performance can be described as an inverted U-curve. A low workload level (e.g. as a result of excessive automation) will lead to degraded performance because of decreased vigilance and a high workload level will lead to degraded performance because of insufficient resources. The highest level of performance is achieved at an intermediate level of workload where vigilance is high and the task demands do not exceed the available resources. In similar vein, the error production and recovery rate may be

constrained by the demands of the context. This idea is illustrated in the figure below (Wioland et al., 1999):

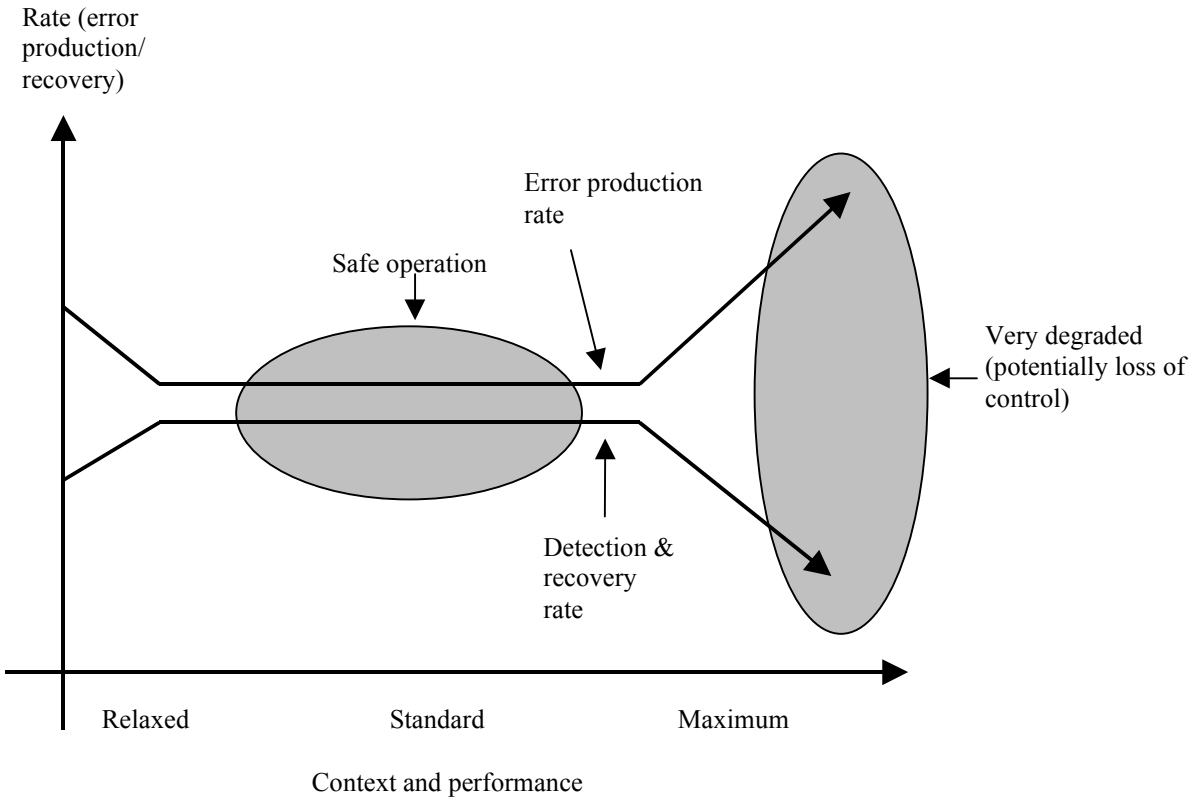


Figure 10: Hypothetical relationship between workload and rate of error production/recovery

As can be seen in the figure the highest level of contained errors can be achieved at intermediate levels of task demands. Here a stable level of errors is committed and most of these are corrected. When the task demands are low there is a tendency to be inattentive and thereby commit more errors and at the same time not be vigilant enough to catch the committed errors. On the other hand when the task demands become too high rapid increases in the error production rate will occur and at the same time the resources for detecting and recovering will become depleted. The result may be loss of control of the situation.

A series of field studies among air traffic controllers have provided support for the notion that individuals progressively adapt their resource management as a function of task demands (Sperandio, 1971; Sperandio 1978). Such changes in control strategies are, in particular, important as the air traffic complexity increases insofar as highly economical operating methods become necessary to avoid that the workload capacity is exceeded. For instance, in many ATM facilities it is common practice during low workload periods to give a shorter than planned route as a general traffic service to the aircraft, but as

traffic load and complexity increases there is a tendency to stick more to the original flight plans. The reason for this is that it becomes exceedingly difficult to maintain the picture as workload increases and by sticking to fixed routes less resources are required in relation to maintaining awareness about the individual flights. It has also been observed by Sperandio that as traffic increases the controllers take into account a smaller and smaller amount of variables for each aircraft and start dealing with clusters of aircraft instead of optimising each individual flight path. By applying such control strategies the controller can handle more aircraft without a high error rate or excessive workload. The disadvantage is, however, that by having all of the flights in-trail and travelling at a uniform speed, the efficiency of the system is temporarily disrupted.

Results from other studies also indicate that people play an active and dynamic role in relation to protecting themselves against the risk of losing control of the situation (Wioland and Amalberti, 1996). These control strategies are based on a compromise between handling the demands of the system in the best way possible and at the same time using a minimum of cognitive resources. The natural consequence of the dynamic cognitive control is that people's meta-knowledge and confidence play an important role in relation to ensuring that the risks stays within acceptable tolerance limits (i.e. the field of safe operations). To have a reliable and well-calibrated meta-knowledge concerning one's abilities to control the situation is important to be tuned into the cues signalising safety boarders being approached and to have necessary skills to recover from these errors. To ensure continuous control of the situation is therefore not the same as total error avoidance, but it is instead dependent on meta-knowledge associated with error awareness and recovery capabilities. This is supported by studies from different domains that show that the errors that are not corrected are also the ones associated with the least risk (Orasanu et al., 1999; Wioland & Amalberti, 1998).

People’s ability to develop reliable and well-calibrated meta-knowledge can inadvertently be undermined by otherwise well-intended safety initiatives. Actually, it has been argued that many existing safety initiatives aimed at improving system safety and efficiency - such as design, training and safety policies - may impede operators’ chances of developing natural and adaptive abilities to control risk (Amalberti, 2001). In particular, the problem is that operators do not experience sufficient possibilities to stabilise their meta-knowledge and confidence with regards to their individual safety abilities. An extreme focus on protections and defences against human errors undermines the possibilities to control the system and ultimately results in an increased risk. Hence, to advance safety within already ultra safe systems it is necessary to focus on "strengthening the ecological mechanisms of cognitive error regulation rather than on fighting them" (Amalberti, 2001).

3.3.2 Threat management

In the following we will examine what is meant by threat and threat management. Several definitions of threat have been suggested:

"A threat is an indication of something coming; a menace; a likely cause of harm"
(Down, 2001)

"Threats are events or errors that originate outside the influence of the flight crew but require active crew management to maintain safety."
(LOSA Rating Form, Human Factors Research Project at University of Texas)

These two definitions vary slightly with regards to which events that might fall under the heading of threat. The first definition treats threats as a very broad concept that, in principle, can cover many different operational factors. Unfortunately, the definition is a bit vague. The second definition is more precise by emphasising that threats should only be events that can be actively handled by the crew, but at the same time puts some unnecessary limits on the concept by constraining it to events that occur outside the influence of the group of actors being observed. Below is suggested a modified definition of threat:

Threats are operational factors that have the potential of jeopardising safety and require active operator involvement to maintain safety.

In the current context, threats are different from errors insofar they are only red flags of potential danger (Down, 2001). If the threats are identified in due time it is possible to initiate actions that will eliminate or reduce their consequence. However, if it is mismanaged or not managed at all the threat becomes an error. In short, a threat might lead to an error, but does not need to. That is, there is a probabilistic relationship between threat and errors. Furthermore, an error can, in principle, occur without any preceding threats (e.g. a slip-of-of-the-tongue can occur without any threats preceding it).

Threats can be subdivided into two independent dimensions:

- **Internal:** These are situations generated at the operational position. This includes, for example, an ATCO that is inexperienced (e.g. on-the-job-training) and inadequate team resource management.
- **External:** These are situations, events or errors that occur outside the operation. This can, for example, be an amateur pilot who is not following the instructions or environmental factors that needs to be taken into account when issuing instructions.

- **Anticipated:** Some examples could be forecasted weather (e.g. a thunderstorm) or a military exercise. In both of these cases the ATCO will have advance information and will thereby be able to incorporate these threats into his or hers plans (e.g. reroute aircraft).
- **Unanticipated:** This can, for example, be an equipment failure or an emergency flight. These kinds of threats are more dangerous seen from an operational perspective because they require an immediately alteration of the existing plans to be able to deal with an unforeseen situation. So, time will have to be spent on both

understanding the new situation and on developing a new plan when time might already be a limited factor.

Even though these dimension seem clear-cut it should be noted that there will be some grey-area cases. For example, a threat might be unexpected at a certain phase in the course of events, but at some point later in time recognised and incorporated into the existing flow of plans. In such cases it might be difficult to determine whether this was an expected or unexpected threat.

Not all threats are equally important and an important task of the controller is to be aware of the threats and to have an adequate understanding of their significance. The potential risk associated with different threats is dependent on both the frequency and the potential severity associated with the threat. Since a human controller does not have access to data about the frequency and the potential severity of a given threat when having to judge and decide about which action alternatives to choose between, they will to a large extent be dependent on the individual controller's experience and training.

A good example of how threat anticipation and management is dependent on the individual operator's background and experience is given in the following authentic story:

ATC Example
The ATCO is working alone at night and the only traffic is a slow-going and light aircraft. The aircraft is flying from South towards Korsa (a navigational fixpoint) at 3000 feet. The standard procedure is to leave Korsa at a certain radial. The pilot is then required to turn to runway and will be at final approach at 12 miles. With a small aircraft like this the procedure is not considered necessary. Based on the current course towards Korsa the ATCO estimates that the aircraft will reach final approach at seven miles. The ATCO instructs the pilot to continue on the current course and promises to give radar vector to final approach at seven miles. The pilot says thanks. Now he does not have to look into procedures to see when to descend to different altitudes. Instead it is the responsibility of the ATCO to ensure that the aircraft is flying at the right altitudes. *Outside Korsa the ATCO instructs the aircraft to descend to 2000 feet. Currently the aircraft is flying 30 degrees and the ATCO intends to give the aircraft turn instructions to 120 degrees to final approach and then the pilot can use the instrument landing system (ILS) to complete landing. The ATCO's wife now calls on the telephone because her car is broken down on the freeway. When the call is finished the phone rings again. This time it is an ATCO from an adjacent aerodrome informing that Rescue 277 (an emergency flight) is starting in five minutes and will perhaps cross the ATCO's airspace. The aircraft has now passed the point at which he should turn to final and is actually close to the sector boarder. The pilot calls the ATCO to ask whether he should not turn to the localiser soon. The ATCO suddenly discovers that he has forgotten to turn the aircraft at the right time and gives instructions to the aircraft to turn right 150 degrees. The aircraft did not get in conflict with any other aircraft, but was about to leave the sector. If the aircraft had left the sector it would not have imposed any conflict factor for any other aircraft.*

On the basis of this episode and similar episodes from low traffic periods the ATCO has learned that he is particularly vulnerable to fatigue and distractions in such working situations and to engage in precautionary initiatives to avoid similar situations in the future. Therefore, the ATCO now gives the pilot instructions to report at certain distances from the VOR. So, even if the ATCO should forget the aircraft a call from the pilot can act as a reminder and thereby an additional safety net. In this manner there is a better chance of breaking the chain of events.
[Source: Personal interview with an ATCO]

Having reviewed and clarified the concept of threat – and some of its dimensions - it is now time to look at what is meant by threat management. A definition is provided below (adapted from Jonker, 2000):

> *Threat management is the act of anticipating and minimising the potential consequences of threats on flight safety.*

In this definition it is emphasised that effective threat management does not only include remaining aware of the critical features of the dynamic environment that vary constantly but does also require that initiatives are made to deal with the threats before they develop into a more serious situation.

3.3.3 Threats and error management strategies

By knowing in advance that certain kinds of errors are more likely to occur in specific threat situations it may become easier to prevent, discover and recover from the error. Some empirical evidence exists to support that experience may play a vital role in relation to being prepared for threat and error situations. In a preliminary study by Mogford et al. (1997) instructors were asked to review five recordings of air traffic sequences containing operational errors. That is, the participants viewed the same information as the controller who originally committed the errors. On this basis they were asked to verbalise any antecedent threats that could lead to an error. Furthermore, they were asked to identify when actions were taken that would lead to loss of separation and make useful suggestions regarding recovery strategies. The results of this study showed that at least one of the four observers recognised an antecedent threat to each error (such as high complexity and similar call signs). Furthermore, in nearly every case the error was identified and useful suggestions were made with regards to recovery strategies. These results indicate that domain expertise supports early recognition of factors that could lead to error, the identification of errors and the generation of useful actions to cope with the problems.

Several other studies have also provided support for the notion that experienced operators have powerful capabilities for dealing with the threats they may encounter. In a study by Klinect et al. (1999) of flight crew behaviour the relationship between threats and errors and their management were examined on the basis of normal flights. Here it was demonstrated that only about 7 percent of the threats influenced the flight crews to

commit errors. This indicates that a large part of the threats are discovered and handled in due time before leading to errors. In another study by D'Arcy & Della Rocco (2001) 100 ATCOs were interviewed about their decision-making and strategic planning. In this study 65 percent of the participants reported that they always try to formulate a (thought-out) back-up plan before sending an initial clearance in case their first strategy for solving a problem did not work. Furthermore, the more experienced the participants were the more likely they reported formulating back-up plans.

As the studies above have indicated control can be maintained on the basis of anticipation of threats by either tackling the threat itself (threat management) or by making the problem constraints explicit and being prepared that the events may evolve in untoward ways (error management). Below are given some concrete examples of how controllers can deal with different kinds of threats:

- **Workload:** Strategies to control threats may be developed as a result of experience. For example, as previously described, controllers start dealing with clusters of aircraft as traffic load increases to minimise the workload (Sperandio, 1978). In similar vein, if the traffic volume becomes high the radar controller in charge of the sector may ask for a data controller to assist in the sector work. In this way an additional set of "eyes and ears" may support situation monitoring and control (Kerns et al., 1999).

- **Memory frailty (ATCO):** Another example of strategies to counteract the occurrence of errors is the way controllers use flight progress strips as an external error memory aid. If the controller has to put something temporarily off (e.g. a request from a pilot which cannot immediately be granted) the controller can offset the relevant strip from the others and use it as a prospective memory cue (Vortac et al., 1995). In this way there is a smaller risk of forgetting the future action.

- **Memory frailty (Pilot):** The STCA (Short-Term Conflict Alert) does not know the ATCO's plan. So, if for example an aircraft is flying at flight level 330 and another is climbing to flight level 310 the STCA may start because within a certain time limit (e.g. 40 sec.) a conflict may occur if the current trajectory is continued. The STCA does not know that the aircraft will not continue through flight level 310. Many ATCOs have made it a good habit to use the STCA as a sort of reminder. It happens that pilots erroneously continue a climb and thereby burst their assigned flight level. The STCA will normally be activated before this happens and the ATCO can therefore make a call to the aircraft to confirm that they will be stopping at the cleared level.

- **Unreliable (amateur) pilots:** The ATCOs warn each other concerning pilots who are less reliable ("if they are given a right turn they might turn left"). This is a way of ensuring to be extra alert concerning these pilots. They are then given a larger safety margin to be prepared for the unexpected.

- **Confusable callsigns:** If a wrong transponder code is used the computer will not be able to generate a label on the radar. It can be a problem if two callsigns are very similar because then a wrong aircraft might enter the transponder code. Here it is particularly important to check the read back to ensure that it was the right aircraft that answered.

The strategies described above are all examples of effective threat management. However, sometimes the adaptive strategies initiated by the controllers may not be equally successful and may in some cases actually be counter-productive. An example is given below:

- **Failed threat management.** In studies of controller-pilot communications it has been revealed that during very busy periods controllers have a tendency to issue longer and more complex messages in a rapid manner (Morrow et al., 1993). This is done to minimise the amount of turn-takings and time on the radio frequency. Unfortunately, this delivery technique imposes heavy memory burdens on the pilots with significant risks of miscommunications. Contrary to the intention the end result may be that a lot of extra time must be spent on clarifications and repair of misunderstandings.

3.4 The "who", "how", "when" and "what" question

After having reviewed the issue of threat and risk management it is now time to turn the focus to what happens after an error has occurred. A comprehensive error management taxonomy should be able to answer following questions:

(1) *who* was involved in the detection and recovery of the error or its consequences;
(2) *when* was the error or its consequences detected;
(3) *how* was the error or its consequences detected and corrected; and finally
(4) *what* was the behavioural response and outcome*?*

In the following we review studies that directly or indirectly deal with these questions in relation to the error capture and management process.

3.5 The "Who"-question

If the detection and correction is not done by the error producer, it may be done by (1) another person in the operational system; (2) automated devices or (3) no one at all - often in spite of recovery opportunities. In the following we review studies related to *who* was the active part in the detection and recovery of the error or its consequences.

3.5.1 Detection and correction by a third party

In many complex environments the safety and efficiency of the system is largely dependent on the resources of all the people involved in the process. This is particularly

evident in relation to detecting and correcting errors. In such situations other people involved in the operational task may constitute an important safety net insofar as they can bring the attention to the problem and may also be helpful in the resolvement of the problem. For example, it has been suggested that detection and correction of operational problems by a third party may be especially useful in situations where a person is unwilling or unable to revise his or her current interpretation of the situation even though data suggests another interpretation (Woods, 1984). This kind of phenomenon is often referred to as fixation error. In such situations the misperceptions are often detected and corrected by a third party entering the situation with a fresh viewpoint.

That a third party can contribute to bringing a person out of his/her cognitive fixation is illustrated in the following description from a British ATM incident report (Airprox 24/96):

ATC Example
A B767 is flying at flight level 280 and when entering a new sector the pilot calls the sector controller. However, the sector controller, when answering him, says "maintain flight level 310" and the pilot replies "up to 310", which is not noticed by the sector controller. Consequently the aircraft climbs toward flight level 310 and is thereby brought into direct conflict with a B747 on flight level 310. ***The support controller notices on the radar display that the B767 is climbing above flight level 280 and tries to bring this to the attention of the sector controller.*** *In spite of the fact that the sector controller had just talked with the pilot and had also ticked the callsign on the flight progress strip (as an indication that the aircraft has contacted the sector), she replies that the B767 aircraft is not on frequency.* ***Only after several attempts by both the support controller and the chief sector controller the sector controller calls the aircraft and gives an avoiding action.*** *It is noted in the report that the high workload may have been a major contributory factor to the occurrence.*

The importance of a fresh viewpoint from a third party has also been highlighted in a study of critical incidents associated with exchanges of anaesthesia personnel during anaesthesia management (Cooper et al., 1982). Such relief procedures are designed to provide the original anaesthetist with either a short break or a relief for the remainder of the operative procedure. There may be advantages and disadvantages associated with the relief procedure. On one hand, the relief practices may be useful because the presence of a new anaesthetist may have restorative effects on fatigue and, at the same time, may support discovery of errors. On the other hand, the relieving anaesthetist may require some time to build a coherent picture of the situation and knowledge of the patient may not be properly transmitted. The study by Cooper et al. demonstrated that the relief procedure was more often beneficial than not. In particular, the relief anaesthetist played an important role in the discovery of an error or causes of an error. Since relief practices are also a characteristic of the ATM environment it can be speculated that similar results could be produced in the area of ATM.

It is also interesting to note how explicit efforts are often made in relation to supporting error detection and correction by a third party in the area of ATM. It is, for example, well known that miss-communications occur frequently between pilots and air traffic

controllers and that such communication breakdowns can have dire consequences for the safety of the air traffic system. Most of the errors are detected and corrected before they have adverse consequences. An important explanation for this is that the system has developed effective and robust cross-people detection and correction mechanisms such as read-back procedures. Another example of how to support cross-people error management mechanisms is by having open and accessible air traffic work spaces which may enhance the chances that a colleague notices or remembers something that the controller may have forgotten (Hopkin, 1995).

To explore in more general terms the detection by a third party process Wioland & Doireau (1995) carried out an experiment where pilots watched a movie of routine commuter flight and were asked to detect pilot errors. The results from the experiment indicated a low average rate of detection by the observers. The errors detected by the observers consisted mainly of rule- and knowledge based mistakes whereas only a small part of the errors detected were slips. A possible explanation of this result is that the observers analyse the situation from a relatively high level of abstraction. This is interesting insofar as those errors most frequently detected by the observers were also the ones that normally are not easily detected by the people committing them. So, even though the *quantity* of errors detected by a third party is relatively low the *quality* of these errors detected may be of considerable importance. There may be several reasons why the amount and quality of errors detected may differ between self induced errors and errors observed by a third party. First of all, the cognitive traces of intentions and actions will only be directly available in the case of self-induced errors. In the case of detection by a third party the observer is deprived of the subjects real intentions (which is, in particular, important in relation to detecting slips) and will have to rely more on indirect criteria, general task knowledge and aspects of context.

3.5.2 The Wioland & Amalberti Taxonomy

An issue that is important in relation to analysing the amount and quality of errors detected is the degree to which the context and goals are shared between the actor and the observer (Wioland & Amalberti, 1998). That is, detection by a third party may be dependent on the relationship between the actor, the observer and the task environment. It may be speculated that if the goals and the context are largely overlapping and compatible there is a good chance of detecting and solving errors by the help of a third party. To describe the relationship between the actor and the observer in relation to the error management process it has been suggested to distinguish between two dimensions: the level of context sharing and the level of possibilities to act on the situation and to share the same goals (Wioland & Amalberti, 1998; Wioland & Doireau 1995). Based on a high and low-level on these two dimensions four types of actor-observer relationships can be produced:

- *Co-actor in context*: the observer and the actor share almost the whole context, goals and actions (e.g. two ATCOs or two pilots).

- *Co-actor outside context*: the observer and the actor share a significant part of goals, but not the context (e.g. an ATCO and a pilot).
- *Outside observer*: the observer and the actor share the context, but have different goals and possibilities of action (an aircraft passenger or a visitor).
- *Excluded observer*: although sharing neither the same context nor the same goals and possibilities of action they are nonetheless related (members of an investigation committee or distance educational situations).

Conclusion	
Advantages	*Diagnosticity.* The nice thing about these distinctions between different types of actor-observer relationships is that it is a generic taxonomy and that the main types of relationships may have different effects on the contribution of a third party in the error management process. *Comprehensiveness.* The categories above seem to cover the main types of generic relations there might exist between an error producer and an error detector.
Disadvantages/ limitations	*Diagnosticity.* Two of the categories seem to be of minor relevance in the current context. "Outside observer" seems to be a highly infrequent detector in the error management process. Furthermore, "Excluded observer" is seemingly a group of actors whose role is post-hoc and therefore not a part of the error management process as such. *Usability.* To agree on what constitutes "context" might be associated with some problems because this might not be a clear-cut-quality. For example, two ATCOs working together in the same position (i.e. a Radar and a Planner controller) will share a significant degree of context. However, two ATCOs controlling two adjacent sectors will share less context, but might still be located closely within the same facility. Therefore, "context" seems to be a matter of degree. *Reliability.* The reliability of the classifications above is currently unknown.

3.5.3 Team related recovery failures

Even if an error has been discovered by a third party it is far from certain that information will be passed on to or received by the error perpetrator. To structure such team related recovery failures Sasou & Reason (1999) have suggested that the recovery process can fall into three stages, namely detection, indication and correction:

- *Failure to detect:* If some member of the remainder of the team different from the error perpetrator had the opportunity but did not notice the error this is a failure to detect.
- *Failure to indicate:* If an error has occurred and has been detected by another team member the recovery may still break down if the error is not brought to the attention of the remainder of the team.

- *Failure to correct:* Even if another member of team becomes aware of the error and indicates it to the error perpetrator it is not certain that the error perpetrator will change his or her mind.

Implicitly in these three stages is the assumption that the error perpetrator should also be the one who corrects the error. This may, in particular, be the case in situations where the error perpetrator is a person of higher competency or if only the error perpetrator has authority to correct the error (e.g. an ATCO may not or can not interfere with traffic in another ATCO's sector).

Conclusion	
Advantages	*Usability.* For team-related recovery activities the above categories seem intuitively understandable.
Disadvantages/ limitations	*Comprehensiveness.* The distinctions are mainly related to team problems where one operator should detect, indicate or correct another operator's error. This restricts the general applicability of the taxonomy (for example, most errors are detected by the people who commit them - see e.g. Wioland & Doireau, 1995). *Diagnosticity.* The categories can only be used to analysing recovery failures and not successful recoveries. That is, they can only be used to redescribe the error in other terms. We do not learn anything about the resources used for catching errors before they lead to serious consequences. *Reliability.* No information about the reliability of the framework is provided.

3.5.4 Detection and correction by automation

Automation is in many complex domains being implemented as a means to enhance system efficiency and safety. In some areas such as aviation a high level of automation is already achieved whereas areas such as Air Traffic Control the introduction of automation is only in an initial phase. However, due to the fact that the current ATC system is in many places stretched to its capacity limits and the prospects of increasing traffic volumes in the near future it is expected that automation will become a more dominant part of ATC. This development may affect the human role in the ATC system and, as a consequence of this, both the errors that will occur and how errors may be handled.

A central concept in relation to understanding the interaction between the human controller and the automated system is levels of automation. Basically the concept refers to the extent to which a task is performed by either the human operator or by machine control. At one extreme, a low level of automation means that a particular function or task is performed with little or no involvement of machine control. At the other extreme, a high level of automation means that most or all of an operation is carried out by automation. Intermediate levels of automation lie between these two extremes.

It has been argued that instead of characterising levels of automation as a uni-dimensional scale it is useful to subdivide it into three dimensions, namely *information acquisition*, *decision and action selection* and *implementation* (Wickens et al., 1998). Information acquisition includes functions such as filtering (for example highlighting highly relevant items and greying irrelevant items) and transformation (for example computing estimated time to contact between aircraft). Decision and action selection are related to the degrees of freedom the human operator has to select different action alternatives. Finally, action implementation is a dichotomous scale and concerns whether an action is carried out by human or machine control.

To illustrate in concrete terms how the three dimensions may affect the error management processes it is useful to take a look at the conflict avoidance task. Many ATC facilities have automated conflict detection systems that constitute an important safety barrier in the case that an imminent conflict is not discovered by the controller. This is a good example of automation at the information acquisitions/integration dimension. Furthermore, automation could also be expanded to the decision and action selection dimension. Similar to automated systems found on-board many flight decks it is possible that the controller is not only warned about a potential conflict, but is also advised about which recovery action to take. Finally, in relation to the action implementation the automation could, in principle, implement the computed most optimal course of action.

Automation can in this way have a role that is comparable to other people in relation to the detection and correction of problems. As in the case with detection by a third party it is also possible that even if a problem is detected by the system it is not necessarily perceived as a problem by the human controller. If, for example, the warning system generates a large amount of false alarms, there is a large risk that the controller will dismiss the warning even though it reflects a genuine problem. In similar vein, the controller may reject an advisory from the system simply because he or she does not trust or understand the rationale of the advisory.

3.6 The "How"-question

In the following section we will explore the potential content of the *error detection* and *correction* stage. Error detection concerns becoming aware of the fact that an error has been committed without necessarily having any specific knowledge about the root cause. This is probably the most analysed stage which is hardly surprising since it is necessary to detect a problem if the error is to be handled. The mental activities associated with overcoming or minimising the consequences of the error is referred to as recovery planning. No specific taxonomies are available for this stage, but since this is an ordinary problem solving or decision-making situation it should be possible to use some of the existing taxonomies within this area.

3.6.1 Error detection

Error detection is probably the part of the error handling process that has been given most attention. This is hardly surprising insofar as error detection is an important part of the error handling process: To be able to recover from an error it is necessary to become aware of the presence of a problem. This awareness can be triggered by a mismatch between the expected outcome and the observed outcome. Furthermore, it may be based on a weak or strong indication of something being wrong. In the following we will review some studies that can shed some light on the different ways at which people become aware of the presence of error. It should be emphasised that the review will not include context-limited experimental tasks. Instead the focus will be on studies that involve tasks and activities with a certain level of realism and complexity.

3.6.1.1 The Allwood-Montgomery taxonomy

One of the first attempts to distinguish between different detection types in relation to solving realistic tasks has been done by Allwood and Montgomery (1982, 1984). This error detection taxonomy was developed on the basis of analyses of think-aloud data from subjects solving statistical problems. The focus was to explore how people detect their own errors before finding the right answer. The categories that emerged from the study consisted of three types of negative evaluation episodes (i.e. types of error detection processes) and these are shown in the figure below:

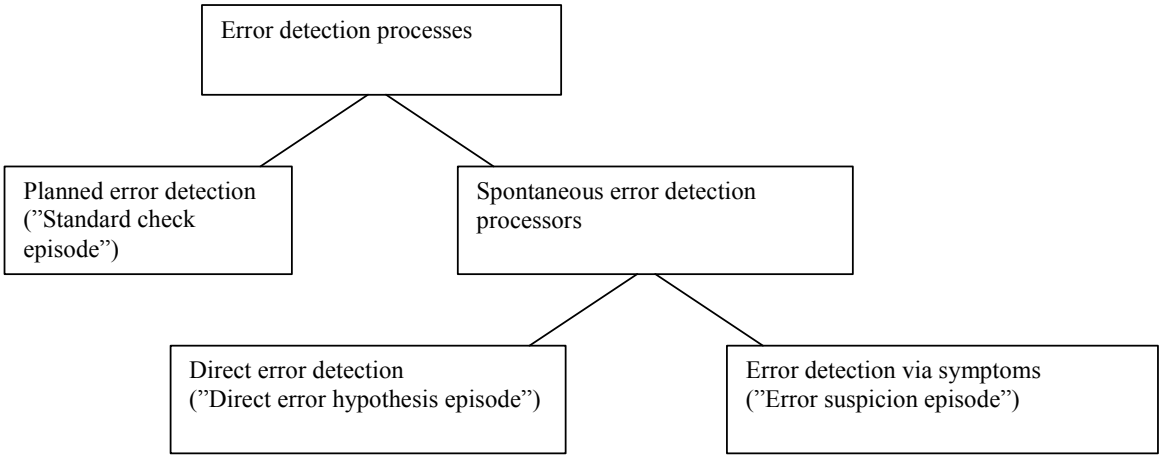


Figure 11: Taxonomy of types of error detection processes.

- *Standard check*: These episodes occur when the subject examines or evaluates the state of the task without having any specific expectation of problems or errors.
- *Direct error hypothesis*: These episodes occur when the subject reacts to a strange result and suspects a particular error to have occurred.

- *Error suspicion*: Even though the subject may not be aware of any specific error having occurred, something in the produced state or solution of the task is strange or unexpected.

The distinction between standard check and spontaneous error detection episodes is interesting insofar as these two categories may be modifiable through different means. The standard check episodes seem to be closely related to the individual problem solving strategies. Consequently, this category may be influenced from training, instruction and experience (e.g. knowing to be error-prone in certain situations). The spontaneous detection episodes seem to depend more on how easy it is to perceive and evaluate the current state of the task. This may critically depend on characteristics of the interface and on the quality of the feedback. In short, the different types of detection can be supported in different ways. Consequently, the taxonomy may have some relevance in relation to the error management framework.

The taxonomy has not only been applied to statistical problem-solving, but also to analysis of subjects think-aloud protocols while using a database system (Rizzo et al., 1987) and to evaluation of an automated process control system of a hot strip mill in a steelworks (Bagnara & Rizzo 1989). Among other interesting results these studies revealed that the different categories of error detection behaviours were associated with different detection rates (the direct error hypotheses have the highest detection rate). Furthermore, the occurrence of different categories of error detection behaviour was dependent on the specific type of error involved. In particular, the results indicated that slips were most closely associated with direct error hypothesis, rule-based mistakes were closely associated with direct error hypothesis and error suspicion, and knowledge-based mistakes were most closely associated with error suspicion.

Conclusion	
Advantages	*Diagnosticity.* The results presented above indicate that the Allwood-Montgomery taxonomy could be useful as a rough way of classifying error detection processes. A positive feature of the taxonomy is the different kinds of errors might be differentially supported by different detection mechanisms. Therefore, by knowing the types of errors that are most likely to occur in a given setting it is also possible to derive the most appropriate detection mechanism to be supported.
Disadvantages/ limitations	*Comprehensiveness.* The Allwood-Montgomery taxonomy is not very detailed and it only describes the kind of behavioural episodes in the error detection, but not the mechanisms behind error detection (Sellen, 1994). For example, we do not have any information concerning the role of memory in detection and we do not know which types of information that were used in the detection. *Usability.* The usability of the taxonomy might be limited in the area of ATC. For example, the concept of Standard Check might not apply very well to the task of the controller insofar as he or she is constantly monitoring and up-dating the mental picture of the situation and the Standard Check can not be distinguished as a separate phase.

	Reliability. No data are available concerning the reliability of the framework, but due to the fact that the taxonomy does not fit very well with domain of ATC it can be speculated that reliability might be jeopardised.

3.6.1.2 The Rizzo et al. taxonomy

Rizzo et al. (1995) have suggested a somehow more refined and comprehensive taxonomy to describe the processes underlying error detection (or as they prefer to call it: mismatch emergence). This classification scheme has been developed on the basis of diary studies where subjects have reported errors they have committed in the everyday live and how they were detected. The main categories are given below.

- *Inner feedback*: This kind of detection is based on information in working memory and not on the consequences on the environment. Rizzo et al. (1995) give an example where a woman has forgotten her book in a pub. Because her bag felt lighter than usual she suddenly remembers that she has left the book at the pub. It can be expected that the reminding/memory retrieval category may be most important in relation to prospective memory failures. Bagnara and Rizzo (1989) suggest that inner feedback also can be associated with mental simulation of activities and of their expected consequences and results. In this case the working memory can be used to make predictions about whether a plan is going to fail or succeed before implementation of any action.

- *Action feedback*: This kind of detection is based on information from the action itself and, again, not on the consequences on the environment. This kind of detection can be triggered by visual, proprioceptive or auditory response-produced information. In this case, the error may be 'caught-in-the-act' (Kontogiannis 1999). This kind of detection is often experienced by skilled typists when they become aware of an error before having seen the output of their action.

- *External/outcome feedback*: Detection occurs due to unexpected consequences on the environment. Sometimes the error information is evident and could, in principle, be detected by any third person even without complete knowledge of the actor's goal (Zapf et al., 1994). At other times - and in particular in complex high-technology environments - the detection becomes cumbersome because the feedback may be delayed and the effects of previous actions may be masked (Kontogiannis 1999).

- *Forcing function*: An action cannot be carried out due to some constraints in the environment. This kind of error detection differs from action and external/outcome feedback by the fact that the correctness of an action does not need to be judged with reference to any internal criterion or expectation. Instead the correctness of an action is determined by constraints and the physical barriers of the environment (Sellen, 1994).

- *Intention uncertainty*: This kind of detection occurs when the subject feels unsure about what do next. This kind of detection is often related to the loss of activation of the ongoing intention. Rizzo et al. (1995) gives an example where a man walks to his office to get a document and on the way gets into a conversation with some colleagues. When he reaches his office he has forgotten what he was there for and had to return to his colleagues to ask why he went to his office.

- *Standard check*: The progress in the task is checked/updated without any specific hypothesis of a problem. Simulations of production planning exercises in a hot strip mill indicated that this self-monitoring strategy was, in particular, useful in detecting slips and to some extent rule-based mistakes (Bagnara and Rizzo, 1989).

There is some overlap, but also some significant differences between this taxonomy and the Allwood-Montgomery taxonomy. In relation to the similarities the most evident overlap is the standard check behaviour. The relationship between the remaining categories in the classification scheme is less evident. It may be speculated that inner feedback, action feedback and forcing function may lead to a direct error hypotheses insofar as these types of feedback will often be available close in time to the error committed (however, direct error hypotheses is not necessarily time-locked to the error). In contrast, error suspicion may arise some time after the error committed and be related to external/outcome feedback or intention uncertainty.

Sellen (1994) suggests an error detection scheme that is very similar to the Rizzo et al. taxonomy. Also this taxonomy was developed on the basis of a large corpus of diary reports. One of the most noteworthy differences between the two taxonomies is the absence of any category similar to the standard check category. A potential explanation for this is that standard check does not describe which kind of information that was used for the error detection. Instead standard check seems to be related to the kind of strategies used for detecting errors. Another category that is not present in the Sellen taxonomy is intention uncertainty. Also this category is a bit problematic. One of the reasons for this is that this category contains a combination of an error (having forgotten what to do) and error detection (awareness of having forgotten what to do).

Conclusion	
Advantages	*Diagnosticity.* The advantage of this taxonomy is that it provides an elaborated list of the underlying processes associated with the error detection. The taxonomy is, in particular, useful insofar as it pinpoints some different types of feedback processes that underlie direct error hypothesis and the error suspicion episodes as described by Allwood and Montgomery (1982, 1984).
Disadvantages/ limitations	*Comprehensiveness.* An important source of information relevant for error detection – namely communication with other people in the operational environment – is not included in the taxonomy (see e.g. Kontogiannis, 1999). *Reliability.* A potential disadvantage of the Rizzo et al. taxonomy is that no studies have been made to validate it. Therefore, it is currently

	not possible to determine its reliability. *Usability.* The level of detail might be too high for the taxonomy to be practically useful. That is, some of the finer details in the taxonomy might be too subtle to be practically useful (e.g. the difference between inner feedback and action feedback).

3.6.2 Error recovery

Having discovered a problem - and maybe identified its cause – the person should consider how to solve the problem. In this case, we do not have any off-the-shelf classification schemes to use. However, since the recovery-planning phase is basically a problem-solving and decision-making situation (with the possible exception of less time and resources available and a higher level of stress), it should be possible to identify a proper decision-making model that can bring about some of the important types of decisions. Since a recovery might be either successful or unsuccessful (i.e. incomplete or flawed) it is necessary that the classification can be used in both cases. Consequently, the classification of the process underlying the decision-making or problem-solving process should not require any normative judgement of what should have been done.

3.6.2.1 The SRK Taxonomy

Previously in the review of error taxonomies Rasmussen's Skills-Rules-Knowledge-model was presented. The SRK-framework has mainly been used in studies of human error. That is, the type of error was determined on the basis of the kind of control that was exercised in the situation. However, the framework could, in principle, also be used to describe the level of performance or behaviour that was exercised in relation to a (perhaps) successful performance (see e.g. Johannsen, 1988). That is, recovery may be associated with different levels of control and experience: (1) Skill-based recoveries are usually routine and automated; (2) Rule-based recoveries use certain types of responses of known and frequently experienced scenarios; (3) Knowledge-based recoveries concern responses to tackle novel and perhaps dangerous situations which require intensive use of cognitive resources.

3.6.2.2 The Bagnara & Rizzo taxonomy.

A potential way to describe the different types of error identification and resolvement processes can be found in Bagnara & Rizzo (1989). Even though this classification scheme can be found under the heading of "Error recovery" the content of the taxonomy seems to be related to both the error identification and recovery process. The main categories of this taxonomy are distinguished on the basis of the amount of resources, if any at all, that are invested in understanding the error. The main categories are:

- *Immediate correction*: "The user, as soon as he detects a mismatch, makes the appropriate correction."
- *Automatic causal analysis*: "The user, as soon as he detects a discrepant outcome, establishes the cause of the error and what should be done for recovery."
- *Conscious causal analysis*: "The user undertakes a typical causal analysis. He carefully evaluates the outcome obtained, the actions previously performed, and, on the basis of these evaluations puts forward the sensible hypothesis about why and when the error has been made and on how to recover from it, plan and execute the actions."
- *Explorative causal analysis*: "The user is able to identify the source of the discrepancy, but finds himself uncertain about the causal chain by which the source can be related to the observed discrepancy. In this case, various hypotheses are usually tested one after the other in an explorative manner."
- *Overcoming of the mismatch*: "The user, when facing a discrepant result, either does not pay any attention to what he has already done or, after an exploration about the discrepancy, realises him unable to reach an adequate hypothesis. In both cases, the user tries either to simply bypass the mismatch, or to fulfil the goals to be reached by looking for alternatives within the same scenario to overcome it, or searches for alternative scenarios."

A central aspect of this taxonomy is that it distinguishes between forward and backward analysis. The first four categories of this taxonomy are associated with backward analysis and describe - in incremental order - the amount of cognitive resources associated with establishing the causal analysis of the error. This kind of taxonomic structure is in good concordance with the Rasmussen skills-rules-knowledge taxonomy.

Conclusion	
Advantages	*Diagnosticity.* The taxonomy distinguishes between different kinds of cognitive resource levels involved in the recovery process.
Disadvantages/ limitations	*Diagnosticity.* A significant part of the framework is devoted to the error identification process. In the current context error identification is not considered relevant for the error management process. *Reliability.* No information is available concerning the reliability of the taxonomy. However, it can be expected that it may be difficult to determine on such a detailed level the amount of cognitive resources invested in the backward analysis. *Usability.* The fact that the taxonomy describes both the error identification phase and the resolution phase within the same dimension might limit its practical usability. *Comprehensiveness.* The fact that many errors might be ignored is not covered by the taxonomy.

3.6.2.3 The Decision Process Model

Orasanu and Fischer (1997) have developed a decision process model in order to describe and understand flight-related decision-making. The model was developed on the basis of both observational studies of pilot crews carrying out critical scenarios in a high fidelity simulator and aviation incident and accident reports. On this basis a number of decision events were identified.

The Orasanu-Fischer model is shown below:

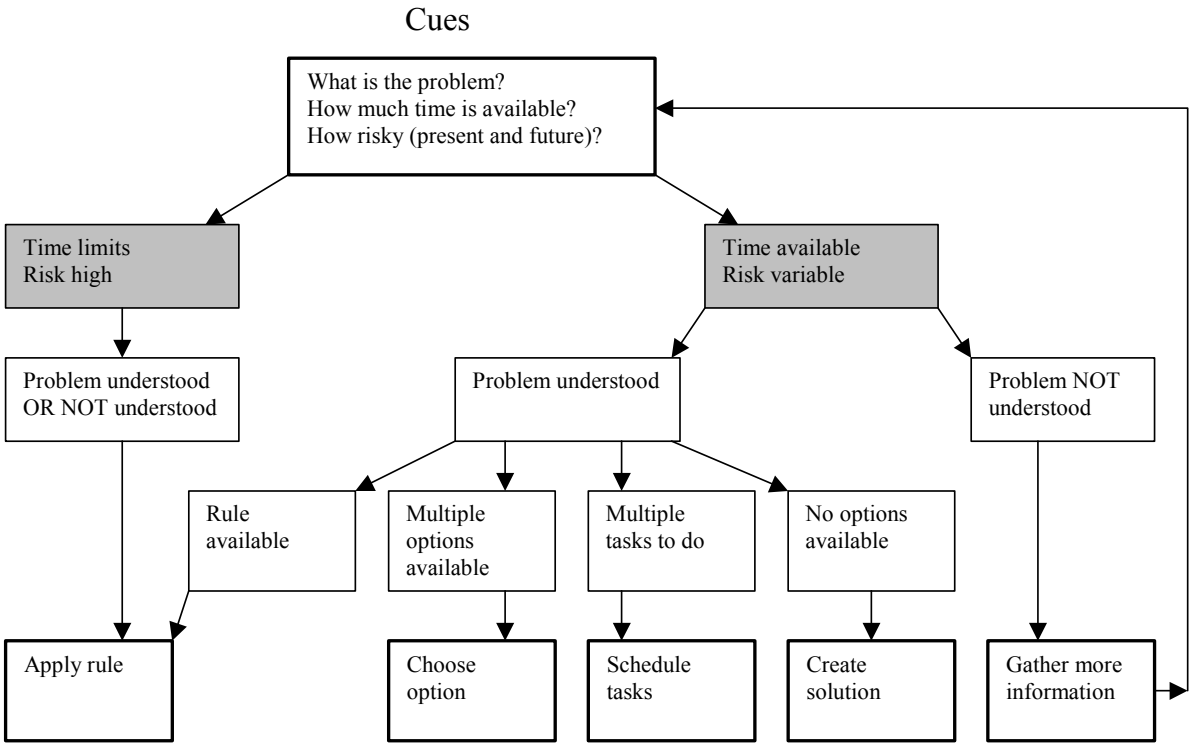


Figure 12: The Decision Process Model.

There are two main components of the model, namely situation assessment and choice of action. In relation to the situation assessment component it can be seen that understanding of the situation as well as the perceived risk and the amount of time available to make a decision play a central role. If the risk is high and the time is limited, it will be necessary to make a fast decision perhaps without having a thorough understanding of the situation. On the other hand, if more time is available several options may be relevant. In the case where a problem is inadequately understood attempts may be made to get more information to disambiguate the situation. If time is available *and* the problem is understood, selecting an appropriate cause of action may depend on requirements of situation. In some situations a procedure may clearly specify the appropriate course of

action. In other cases where no rule prescribes a single procedure it is necessary to either choose between several options, schedule problems or inventing a new course of action.

Conclusion	
Advantages	*Diagnosticity.* The model contains some important dimensions relevant for task domains characterised by high risk, tempo and complexity. In particular, the decision events are derived from critical task components such as risk, time pressure, situational ambiguity (i.e. does the information available clearly specify the problem) and response determinacy (i.e. is there a single prescribed response, several responses to choose and prioritise between; does the person have to create a unique response for the situation). *Reliability.* The model can be directly transferred into the flowchart and it is easy to see the relationship and differences between the individual categories. *Usability.* The concepts seem to be very intuitive and do not require specific theoretical insight.
Disadvantages/ limitations	*Reliability.* A potential problem with the category "schedule tasks" is that it may be difficult to distinguish from “choose option”. *Comprehensiveness.* All of the categories in the model are derived from the assumption that a problem is present and that some remedial action is required (this is related to the fact that the model deals with responses to abnormal and emergency conditions, not with routine decisions). This assumption is not necessarily always true for error situations. In some situations no deliberations of recovery solution is necessary simply because an intervention might exacerbate the situation or because the error problem was considered inconsequential. That such error awareness is of importance has been demonstrated in an experimental study of a micro-world that simulated an air traffic control situation (Wioland & Amalberti, 1998). Here it was shown that the subjects with increased level of expertise (and thereby increased level of knowledge of the system and the risks) tolerated a larger amount of errors without consequences. Therefore, a category labelled "ignore" may be included in the "problem understood"-branch.

3.7 The "When"-question

A framework for the analysis of cognitive reliability that contains some potentially relevant categories has been proposed by Kontogiannis (1997, 1999). Of particular relevance in the current context is the taxonomy associated with the processing stages at which an error may be detected. In the figure below can be seen that a performance sequence starts out by setting high-level goals and formulating plans to achieve the goals. At this initial stage in the performance sequence flaws in the goals and plans might be realised before they are implemented on the system. At the next stage the chosen action is carried out and feedback associated with the action might indicate some deviation from

the desired and intended goal. Finally, actions carried out will after some time delay have consequences. Detection at this stage might be hindered by operator actions being masked by e.g. actions taken by automated safety systems or other persons. Seen from a safety perspective detection should preferably occur before critical consequences have ensued.

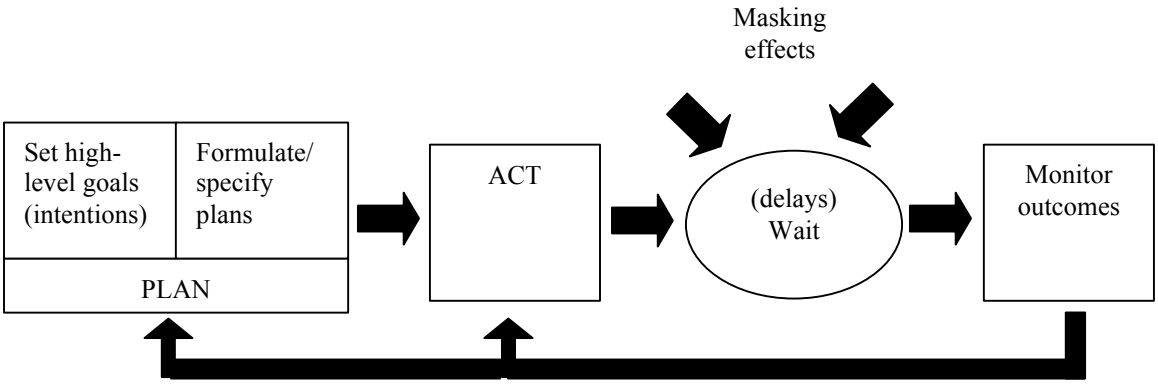


Figure 13: Performance stages at which error detection/recovery can occur.

On the basis of this model it is possible to distinguish between three different kinds of detection stages, namely the planning, execution or outcome stage.

- The outcome stage: A mismatch between expected effects and observed outcomes can trigger error detection.
- The execution stage: Errors are 'caught-in-the-act" and subsequently corrected.
- The planning stage: Operators recognise wrong intentions, or mismatches between intentions and plans or procedures formulated earlier.

Conclusion	
Advantages	*Diagnosticity.* The three stages of error detection are of interest because the required safety initiatives to support error detection may vary for the individual stages. Error detection at the planning stage may e.g. be dependent on the quality of communication and co-ordination between controllers, issues gathered under the general heading of team resource management. Therefore, this kind of detection may be enhanced if operational plans and decisions are properly communicated and acknowledged; if clear roles and responsibilities are defined; if the controllers are open for reviewing, questioning and revising plans. Detection at the execution stage may be supported by controllers actively monitor and crosscheck information from colleagues, pilots and system. *Comprehensiveness.* All major stages seem to be covered by the taxonomy.
Disadvantages/ limitations	*Usability.* Some may object to the concept of error detection at the planning stage insofar as it can be argued that if an error has not been

	carried out it should not be considered error. Nonetheless, as explained above, there can be good reasons for distinguishing between error detection at the planning stage and at the execution stage since error detection can be supported by different means at these two stages. *Reliability.* Finally, it should be mentioned that no formal attempts have been made to apply or validate the classification system to real error situations.

3.8 The "What"-question

Recovery in dynamic environments is in its nature more complicated than in static environments. The important difference between these two kinds of environments is that in dynamic environments the time has direct consequences for the state of the environment. Hence, the implications of errors may alter as a function of time. If, for instance, two aircraft are on a conflicting course due to an error by an ATCO, the criticality of the situation may significantly depend on the time elapsed from error production to recovery.

The potential relationship between error recovery and outcome failures in dynamic and time critical environments is depicted in the figure below (Woods et al., 1994):

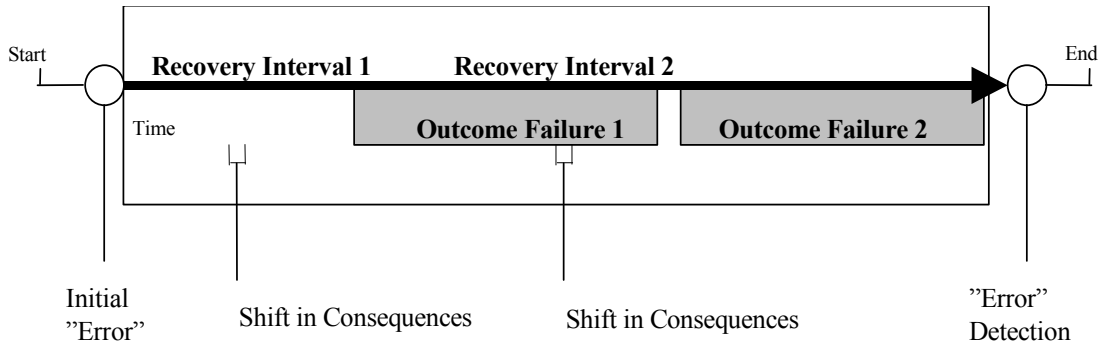


Figure 14: Relationship between error recovery and outcome failures.

The figure describes the hypothetical consequences of an error as a function of the recovery interval (please note that the transition between the recovery intervals may not be clear-cut in many situations). If the error is corrected within the first recovery interval the effects of the error will completely be reversed. However, at some point in time the error will have negative consequences. The consequences may increase in severity as additional recovery intervals are exceeded and at some point recovery may no longer be an option. The time span between the recovery intervals can be seen as indication of the error tolerance of the environments since in an error tolerant environment an error should not have immediate and irreversible consequences. In the following we examine two ways of classifying the response and consequences of an initial error.

3.8.1 Backward and forward recovery

The goal of error recovery is to counteract the negative effects of an error. This may be accomplished with different effects on the environment. Dix et al. (1993) distinguish in this context between two different types of recovery execution.

- *Backward error recovery*: This type of recovery concerns situations where the situation before the occurrence has been restored. In traditional computer systems this can be achieved through commands such as "undo", "cancel" and "stop".

- *Forward error recovery*: In many situations it is not feasible to return to the state of affairs before the error occurrence (for example, breaking a piece of china). Instead an alternative course of action will be necessary to minimise the negative consequences of the action and, as far as possible, restore the situation to normal (for example by gluing the pieces of the china together).

At first glance one would expect that backward error recovery is mainly a possibility in static systems (such as traditional HCI tasks) whereas forward error recovery is the only way to go in safety critical domains such as air traffic control where actions and time can cause irreversible changes to the object of interest. That is, most emergency procedures concern stabilising the situation. Nevertheless, backward error recovery may also occur in many situations. If, for example, a pilot reads back a clearance wrongly, the controller can correct it immediately and the effects of the error are totally removed.

In time critical systems the chances of carrying out a backward recovery may depend on system dynamics and on the time elapsed between the error and the recovery response (Jambon, 1997). That is, within a certain time window it will normally be possible to restore the original situation before the occurrence of the error. After this there may still be chances of achieving a non-optimal, but more desirable system's state (i.e. forward recovery). As time elapses system failures may propagate and at some point in time it may no longer be feasible to achieve a recovery - and a disaster is possible.

Conclusion	
Advantages	*Diagnosticity.* It can be important to be able to distinguish between situations where a complete recovery without any consequences is obtained from situations where it is necessary to make a quick fix to stabilise the situation.
Disadvantages/ limitations	*Reliability.* It may be difficult to make a clear distinction between forward and backward recovery in dynamic environments which may be related to the fact that the framework originally has been developed for static tasks. *Comprehensiveness.* The distinction does not provide any information about the successfulness or unsuccessfulness of the error management – that is, the outcome of the recovery. *Usability.* The distinction between backward and forward recovery

	might be a bit subtle – in particular in dynamic environments. Here the state of the environment will change as a function of time and it might be difficult to determine whether the situation before the occurrence has been restored.

3.8.2 Threat and error management

Previously, Helmreich's model of threat and error management was presented. Below is shown graphically the classifications system associated with the model.

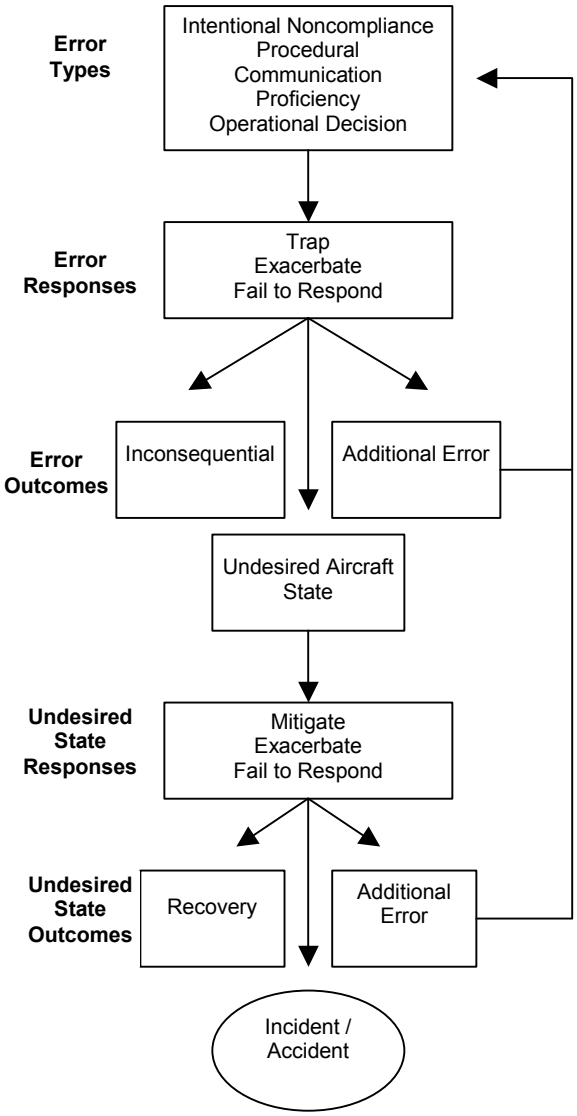


Figure 15: A model of flightcrew error

After the production of an error three different kinds of responses may be produced, namely trap (error is managed before it becomes consequential), exacerbate (the error is detected but the action or inaction leads to a negative outcome) or fail to respond (either because the error is undetected or ignored). On the basis of these three responses there are three possible outcomes: inconsequential, undesired aircraft state or additional error. If the consequence is an additional error this may be the beginning of a chain of errors. If the result is an undesired aircraft state this may be mitigated, exacerbated or not responded to (i.e. "Failure to respond"). Finally, there can be three different possible resolutions of the undesirable aircraft state: recovery, additional error or incident/accident.

The model and associated taxonomy presented in this section is directly related to the model of threat and error management presented earlier on. Therefore, the main conclusions related to the framework can be found in the review of the threat and error management model (see section 3.2.2). However, in the current context it suffices to say that the framework provides a useful and intuitively logical way of classifying both the response and outcome associated with the error management.

4 Performance shaping factors

In the discussion of the nature of human error it has been argued that accidents do not happen narrowly as the result of human errors (as well as failed recoveries), but can instead be seen as instances of human-task mismatches (Rasmussen, 1987). In other words, properties of the task environment - including everything from poor design to bad management decisions - are important to include when analysing performance breakdowns in complex systems. This fact has been highlighted by detailed analysis of tragic accidents both within the area of aviation (e.g. Tenerife) and many other domains (e.g. Three Mile Island, Bhopal, Zeebrugge). A logical consequence of this insight is that features of the context and the work situation should be taken into account when analysing the chain of events in critical scenarios. Only in this manner it is possible to obtain a comprehensive understanding of the aetiology of the events and, ultimately, to be able to enhance system safety. In this section the focus will be on the contextual influence on human performance or, in short, Performance Shaping Factors (PSFs).

One of the first areas where the concept of PSFs was used was in the domain of Human Reliability Assessment (HRA). The purpose of HRA is to make a risk assessment of a given system by analytical and predictive means instead of having to await empirical data from incidents and accidents. The risk assessment is achieved through a combination of logical “tree” models of a system and human error probabilities which allows an estimation of how the system functions might be affected by human error. In this context PSFs have been used to modify the human error probabilities (Swain & Guttmann, 1983). The exact structure, content and number of PSFs has varied as a function of the chosen methodology and the domain in question.

The concept of performance shaping factors was also adopted by Rasmussen (1982) as an integrated part of his multi-facet taxonomy for description and analysis of events involving human malfunction. It was introduced as a recognition of the fact that it is insufficient to only look at the information processing aspects of man-machine interaction when analysing the chain of events in situations involving human malfunction.

A more recent attempt to integrate the influence of contextual factors on the genesis of human errors is found in a model by James Reason (1990), namely the well-known "Swiss Cheese" model of human error causation (see below). In this model there are four levels of human failure that can each influence the next. The model works backwards starting with an accident that was triggered by active failures by the people at the frontline. However, these active failures only constitute the last "holes" in the cheese and before these three levels of so-called latent failures preceded the active failure. The first is psychological precursors of unsafe acts and include factors such as mental fatigue and poor communication and coordination practices. The next level is in the model can help explain why these precursors were present and is referred to as line management deficiencies. Human failures at this level can e.g. be reflected in bad manning practices or deficiencies in the training department that can be manifested in a variety of

preconditions such as the aforementioned. The final level in the model is fallible decisions made at the organisational level. At this level the problems are, in particular, related to the trade-offs between the two not always compatible goals, namely production and safety.

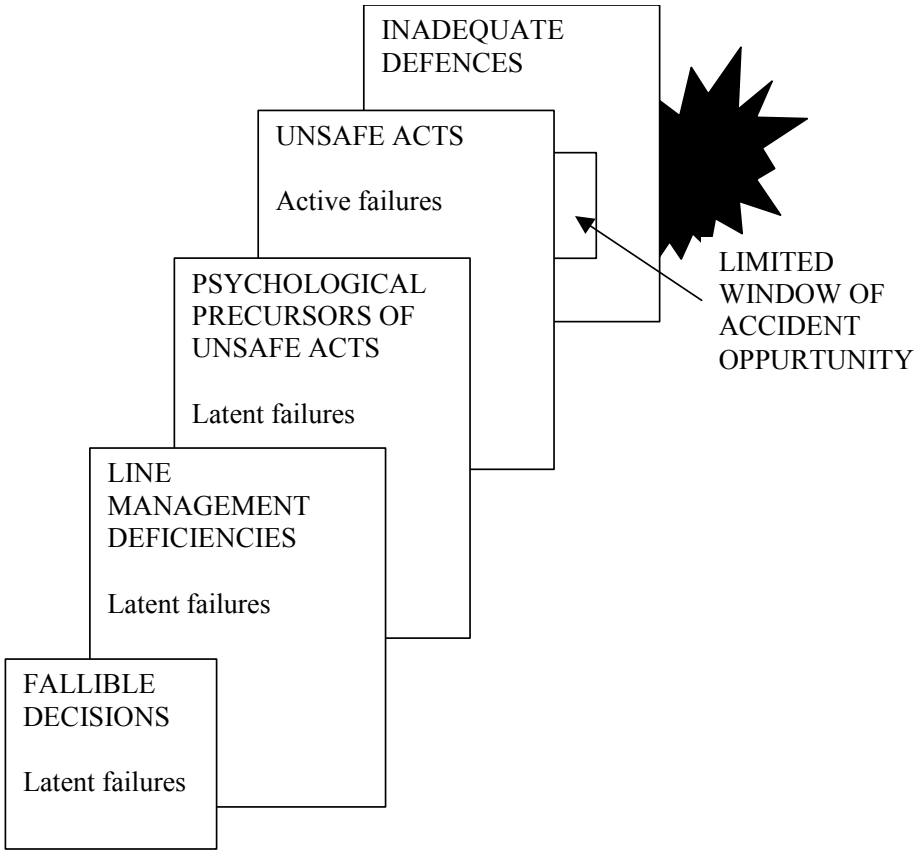


Figure 16: Reason's multi-layer model.

Reason's model is of interest because it encourages investigators and analysts to move beyond the people at the frontline and to expand the focus to all levels within the organisation. A limitation is, however, that it does not define what "the holes in the cheese" really are (Shappell & Wiegmann, 1999). That is, in order to be able to make a detailed analysis of contextual influence on human performance it is necessary to produce a detailed taxonomy that can be used for analysing incidents and accidents.

4.1 Taxonomic considerations

In relation to development of a list of PSFs there are some unique requirements that should be taken into consideration:

Positive and negative factors: Traditionally, the focus has been on the negative side of the PSFs, namely on factors which could adversely affect the operator's performance (i.e. human errors). Even though the PSFs are normally used in this negative context, the

concept itself is, in principle, neutral and should therefore cover all factors (both positive and negative) that are likely to affect operator's performance (from error to recovery). In other words, it should be possible to expand the concept to also encompass *positive* and *negative* factors that influence the *production*, *detection* and *recovery* of human errors. In this context it should acknowledged that there are some difficulties in relation to determining and eliciting positive contributing factors. When analysing, for example, an incident, it is often possible to enlist a series of factors that contributed in a negative way to the events. Here it is possible to use counter-factual logic and state that if these factors had not been present or had been different then the incident would probably not have occurred. On the other hand, it is far more difficult to identify factors that actually had a positive contribution to the course of events and thereby were important in relation to averting an even more serious situation. This is because these factors can be seen as factors taken for granted - they constitute "the background of the picture". Since these positive factors do not stand out in the same way as the negative factors it becomes more difficult to identify and classify these "what saved the day" factors.

Categories not mutually exclusive: In contrast to the previous taxonomies the PSFs cannot always be mutually exclusive. As it has been emphasised by Rasmussen an identification of the root cause to a sequence of events is dependent on the stop rule applied in the after-the-fact-analysis. Even though it is probably not possible to produce a PSF taxonomy that contains categories that are mutually exclusive, it would nonetheless be a strength of the taxonomy if a single category could be chosen in most of the cases being analysed. This is related to the fact that if many equally good candidates are available in many cases this will introduce randomness in the choice of categories as well as the amount categories being picked.

Domain specific issues: Up until now it has been an explicit requirement that categories should be generic and task-independent. However, it should be acknowledged that some groups/categories will be domain specific. The reason for this is that each domain will possess unique characteristics that will be inadequately described by a strictly context-independent PSF-taxonomy.

Interaction between levels: In Reason's "Swiss cheese" model it is implied that features at one level can affect the following level. In principle this opens up for a series of interactions (e.g. between the high-level decision-makers and the line management) that ultimately can have consequences for the chain of events leading to a disaster. If all the possible interactions should be taking into account it would be a rather daunting task. Furthermore, the more removed from the front-line the focus comes the more difficult it becomes to establish a link between error or error management and the PSF. For this reason it was chosen to restrict the framework to only involve the potential interactions between the performance at the frontline and contextual influences.

4.2 Overview of frameworks

A few studies have been carried out that have produced some distinctions that are relevant and useful in the current context. The studies and associated frameworks will briefly be reviewed in the following. Since there exists an extensive amount of such frameworks the current review is in no way intended to be exhaustive. Instead the focus is on frameworks that have explicitly focused on the domain of the Air Traffic Management (or aviation in general) and/or in some way have focused on factors influencing the recovery process. The review of these frameworks will be somehow limited by the fact that little research literature is currently available to determine their utility.

4.2.1 HERA PSFs

One of the most comprehensive PSF taxonomies has been developed in the HERA project (Isaac et al., 2000). This project was carried out as a collaborative effort between Risø National Laboratory and National Air Traffic Services (NATS) for the European organisation for air traffic management (EUROCONTROL). The purpose was to develop an approach to ATM analysis to determine how and why humans were contributing to incidents. An important part of the HERA approach was the PSF taxonomy that was used to capture the context surrounding the ATCO's task. The taxonomy was developed on the basis of a huge database of ATC incident reports (more than 50 actual incident reports from different European and non-European countries), a literature survey and input from domain experts and, finally, a review of current and future ATM systems to ensure that the aspects that are particularly relevant in the ATM system were included. The main groups that have been generated on this basis can be seen in Appendix B.

Conclusion	
Advantages	*Diagnosticity.* The PSF framework has been directly aimed at the ATC environment and is based on an extensive database of ATC incidents. It should therefore be highly relevant for the current project. *Comprehensiveness.* The list of PSFs has been determined on the basis of an extensive review of factors influencing incidents in several different countries. Furthermore, about 60 per cent of the subjects who participated in an evaluation of the framework indicated that they felt that the level of coverage within the PSFs was about right (Isaac et al., 2001). *Usability.* The overall structure of the PSFs is fairly logical and easy to understand.
Disadvantages/ limitations	*Comprehensiveness.* The taxonomy has mainly been developed to structure the analysis of factors that enhance the potential of errors and not error management as such. Therefore, it should be transformed and adapted to become applicable to the broader definition of PSF. To do this it is necessary to rethink many of the identified categories so that

	the negative phrasing is converted into a neutral phrasing (e.g. "abnormal time pressure" can be altered to "time available and degree of urgency"). *Comprehensiveness.* Even though the subjects who participated in the evaluation of the framework indicated that they felt that the level of detail was about right the list of PSFs might be too comprehensive and the level of detail too ambitious. The problem could, for example, be that several categories might apply to the same factor and consequently it might be a bit arbitrary which category or categories that will be chosen. Furthermore, the framework runs the risk of producing many missing values as well as requiring the classifier to make some very subtle distinctions. *Reliability.* An extensive validation of the PSFs revealed that analyses of errors in incident reports on the basis of these PSFs did not provide robust results (Issac et al., 2000). This might partially be a result of the high level of detail within the system.

4.2.2 ADREP-2000

ADREP-2000 is a classification system that has been proposed by the International Civil Aviation Organisation (ICAO) for structuring the data analysis of aviation accidents (Cacciabue, 2001). This highly detailed taxonomy provides the opportunity to subtract important human factors insight from accident databases. The foundation of the framework is the classical SHEL-model that has been suggested by Edwards (1972). In this model the focus is on the human component (i.e. the liveware) and its interaction with other main components within a socio-technical system. The components are given in the SHEL-acronym: Software, Hardware, Environment and Liveware. The main groups available within the ADREP-2000 system generated on the basis of these components can be seen in Appendix B.

Conclusion	
Advantages	*Diagnosticity.* Elaborate efforts have been made to ensure that all possible contextual factors are included in the framework and it has with success been tested out on a number of accident reports (Cacciabue, 2001). *Comprehensiveness.* All conceivable factors that could influence safety within the area of aviation seem to be included in this very comprehensive framework. *Usability.* The fact that the framework is based on a traditional and well-accepted overall structure for the framework (namely the SHEL-model) makes it easier to "navigate" within the taxonomy.
Disadvantages/ limitations	*Comprehensiveness.* The level of detail within ADREP-2000 seems to be too ambitious. *Usability.* The high level of detail also compromises the usability to some extent because it is necessary to walk through a very long list

	before finding a relevant category. At the same time the classifier might have to check a lot of other places within the taxonomy to see whether some other category might be more appropriate. *Reliability.* The reliability of ADREP-2000 is currently unknown. Since many similar categories can be found in the taxonomy this might compromise the reliability. For example, it might be difficult to distinguish between several categories under the heading of "Psychological limitations" such as Perception, Attention, Monitoring (attention) and Vigilance.

4.2.3 Recovery influencing factors

A preliminary list of factors that can influence the recovery process has been proposed by Kanse & Van der Schaaf (2000c). The list has been developed on the basis of findings from a literature survey and an exploratory study that involved incident data from a chemical process plant. The factors identified are able to affect all phases of error management (which in the current context involve error detection, explanation and correction) even though some factors might only affect parts of the error management process. The main groups of recovery influencing factors can be seen in Appendix B.

Conclusion	
Advantages	*Diagnosticity.* The framework does provide some unique categories relevant for the recovery process and most of the main groups fit directly with the structure of the HERA PSFs.
Disadvantages/ limitations	*Usability.* Some of the main groups (e.g. "Factors relevant for prioritisation of recovery related tasks" and "Occurrence-related factors") are not very intuitively understandable. Furthermore, some of the specific categories are probably a bit difficult to directly apply to an error management analysis (e.g. "Feeling of responsibility with regard to recovery" and "Pride regarding a job well done"). *Diagnosticity.* The list of factors is only designed to describe factors that affect the recovery process and not the error production process. *Comprehensiveness.* It should be noted that the list of factors is not as comprehensive as the HERA PSFs (and, of course, no Air Traffic Control specific groups are included). *Reliability.* The reliability of the categories is currently unknown. However, it does contain several categories with subtle differences. For example, under the Person related factors, there are several closely related factors: "Competency in task concerned", "Competency with regard to specific problem occurrence" and "Competency in problem-solving tasks in general".

4.2.4 ASAP contributing factors

The ASAP (Aviation Safety Action Programme) incident reporting form has been developed by Helmreich and co-workers to obtain both structured and unstructured information about events that have the potential of negatively impacting aviation safety (Helmreich & Merritt, 2000). In the form there is a list of factors that can influence the events in either positive or negative direction. Most of the main groups are specifically related to the aviation domain (e.g. operational tasks, aircraft and auto flight system) and therefore not relevant in the current context. However, one of the groups, namely "Cockpit Crew Factors", could be considered highly relevant in the current context. These factors are of interest in so far as they concern crew resources management (CRM) issues. The list of ASAP cockpit crew factors is provided in Appendix B (source: ASAP Report Form Version 2/24/00, The University of Texas Human Factors Research Project).

Conclusion	
Advantages	*Diagnosticity.* The CRM factors seem highly relevant and concern social factors that could play a vital role in relation to error and error management. Furthermore, these categories are of such a nature that they can contribute in both a positive and negative direction. *Usability.* The categories seem easily applicable and refer to issues that have been well documented within the research literature.
Disadvantages/ limitations	*Comprehensiveness.* It is mainly categories related to social factors that are of relevance within this framework. *Reliability.* The reliability of the framework is currently unknown. Some of the categories cover issues that share many features and, consequently, might not be easily distinguishable.

4.2.5 BASIS

The Human Factors Reporting (HFR) programme is a component of the British Airways Safety Information System – in short, BASIS (O'Leary, 1999). The purpose of this confidential and voluntarily reporting system was to obtain information concerning why flight crew related problems occurred and also how effectively the crew coped with the problems. The main groups of categories within the framework are shown in Appendix B.

Conclusion	
Advantages	*Diagnosticity.* This is one of the few frameworks that in a comprehensive manner deals with a wide range of factors that can contribute positively and negatively to safety. *Comprehensiveness.* The framework contains a series of main groups that are relevant for understanding aviation safety. *Usability.* The overall structure of the taxonomy is easily

	understandable and in good concordance with other frameworks.
Disadvantages/ limitations	*Reliability.* No information is available about the reliability of the framework. *Comprehensiveness.* A large amount of the categories are specific to the domain of aviation and do therefore not apply to ATC.

4.3 Conclusion

As can be seen from the previous sections there already exist different classification systems that could be of relevance in relation to developing a PSF-taxonomy. Even though there are some variations in the structure of the main groups and the level of detail in the individual frameworks their overall structure is fairly similar. The HERA taxonomy is the only one of the previously described frameworks that has been extensively validated and that has been directly aimed at air traffic control. It is also characterised by the fact of having been derived from an extensive amount of incident reports. For these reasons it seems useful to start out with this framework. However, since the focus is limited to negative factors it could benefit from being modified by other PSF-classification systems such as the other considered frameworks since they deal directly with the issue of recovery. Another issue that should be taken into consideration is the fact that the HERA taxonomy has in an empirical validation yielded only a very modest level of inter-rater reliability. The exact reason for this cannot be determined with certainty, but a likely explanation is that the system is too comprehensive. Consequently, it might be beneficial to reduce the level of detail in the taxonomy to a more modest level.

5 Enhancing error management

Error management is not just a coincidence but is instead something that can be reinforced through different kinds of human factors initiatives (Van der Schaaf, 1995). In particular, team training and implementation of new technology seem to be promising ways to strengthen the systems' defences against human errors. In the following we will explore how research within these two areas could have some beneficial effect in relation to enhancing error management within ATC. More specifically, the focus will be on the training concept referred to as Team Resource Management and the design concept referred to as Interactive Critiquing.

5.1 Training for error management

In spite of the fact that many errors happen on a daily basis in Air Traffic Control facilities the rate of loss of separation remains relatively low. This indicates that safety nets embedded in the system, in addition to a portion of sheer luck, play a significant role for the safety and for breaking the chain of errors that may lead to an accident. Some of these safety nets are technologically based whereas others are based more on human resources. If improvements in error capture and thereby safety is to be achieved it seems obvious to focus on these two mechanisms which are amenable to change. Error management promotion based on technological improvements is limited by being both time-consuming and costly. On the other hand, human performance and teamwork is more adaptable and amenable to change. In the following we will review the relationship between effective teamwork and error management - and how controllers through training can become better at helping each other in anticipating, detecting and resolving potential problems.

Some of the most disastrous accidents in the history of aviation - such as the Tenerife-accident in 1977 and the Potomac-accident in 1982 - involved situations where at least one other person was aware of (or suspicious of) the problem but failed to share critical information or generate a response from the appropriate person (Hawkins, 1987). A less dramatic but illustrative example of how incomplete teamwork can adversely affect safety is given in the following authentic air traffic control incident report.

ATC Example
The R8-position was at 13.42 split into R8 and D8. When the strips for KLM and SAS arrived to D8 he placed the strips in the FPB. The strips arrived approximately at the same time. R8 and D8 discussed whether a conflict would occur and agreed that this was not going to be the case, but decided to follow up on situation. How this should be done was not discussed. *At 14.04 EKDK Planner A co-ordinated with D8 "KLM direct SORLA" which D8 accepted and marked SOR on the strip. R8 did not become aware of this information. D8 did nothing to bring this information to the attention of R8 since this information was considered "routine". In this manner R8 did not achieve a full picture of the traffic situation.*

At 14.06 SAS was on TRENT and was cleared ALM direct LEKSI which gave SAS a shorter flight route to the east of the planned route. At 14.12 KLM called on the frequency and was cleared directly to SORLA by R8. KLM had gotten this route by EKDK, but as earlier described R8 did not know about this. When KLM called he was North of the traffic route that R8 had expected. R8 did not know this since he focused on the other traffic situation to the Northeast. The ATCO in the R8-position said (according to the tape recordings) that he had radar contact with KLM. After the incident R8 could not recall he had said radar contact or whether in this situation he had made any estimation of conflict. In the mental picture of R8 there was no conflict in this situation.
R8 was engaged in a conflict to the Northeast. R8 therefore changed the radar picture to be able to see situation better. When this traffic situation was solved R8 changed his radar picture back. Some minutes before D8 was relieved. According to the procedures D8 should now relieve R8. When the relieved D8 stood beside the R8-position to relieve, R8 discovered that there was a conflict between KLM and SAS. At 14:18:08 SAS was given order to descend immediately to FL 270. SAS received traffic information and discovered immediately afterwards KLM. The separation was violated.
[Source: Swedish CAA report # 981115]

In this incident R8 and D8 detected at an early point in time that there was a risk of conflict between KLM and SAS, but decided that it would probably not be a problem. Incomplete communication and co-ordination between these two controllers - combined with the distraction by the other traffic situation that demanded attention - created the foundation for not catching this erroneous judgement and thereby not avoiding the conflict. In short, if the teamwork had been more effective the incident would probably have been avoided.

One approach to improving error management is training programs united under the heading of crew resource management (CRM) or team resource management (TRM). The goal behind these training programs is to make better use of the human resources with specific focus on enhancing inter-personal aspects such as communication, group decision-making and leadership. Such issues have been identified in many system breakdowns to play a crucial role for safety within the area of aviation as well as other safety critical domains. Several generations of crew resource management have evolved as new and refined insights have emerged concerning the relationship between teamwork and safety. For the most recent generation of crew resource management it has been suggested that the overarching rationale for the training should be error management (Helmreich & Merritt, 2000). This implies that crew resource management should provide countermeasures against error in the form of avoidance, detection and management techniques.

Nagel (1988) has concluded that over half of aircraft incidents are the result of communication breakdowns. This makes the link between crew resource management and effective error management behaviour an important issue seen from a safety perspective. Therefore, it is of interest to determine what kind of crew resource issues are critical factors for ensuring the control of errors. In a study conducted by Jones (1998) team issues related to successes and failures in an air traffic system were investigated. More specifically, events related to mishaps, normal operations and exemplary

performance were contrasted on the basis of three team related scales, namely *task management* (e.g. contingency planning and workload distribution), *information exchange* (e.g. offering and encouraging sharing of information) and *interpersonal relationships* (interpersonal sensitivity and receptivity). Preliminary results from the study revealed that the team issues as reflected in the three behaviour scales played a critical and significant role in relation to operational errors. In other words, positive scores on the scales were associated with an absence of mishaps.

Even though no firm conclusions can be drawn from the study concerning whether the behavioural markers of team skills were positively associated with mainly error avoidance or error management (or both) these results highlight in general terms the importance of team skills in error containment. Furthermore, it is not difficult see how these three team dimensions may relate to error management. The task management dimension is primarily related to ensuring a clear workload distribution and preparing for a multitude of contingencies. This is exemplified in the incident described above where the lack of contingency plans impaired the team's ability to cope with their initial erroneous judgement of the situation. That is, even though they knew that the judgement of situation could be wrong they did not consider precautionary initiatives. The other dimension, namely information exchange, concerns e.g. passing information to appropriate persons without being asked, asking questions to clarify ("take nothing for granted") and providing periodic updates which summarises the picture. In the incident above the data controller could have been more effective in the information exchange to ensure that the radar controller became aware of the altered course of one of the aircraft. If this have been done the radar controller would have had a more accurate picture of situation and perhaps have discovered the emerging conflict at a much earlier point in time. The final dimension is interpersonal relations and concerns issues such as acceptance of critique and listening actively to ideas and opinions of others. An example of how failure to listen to comments from colleagues can jeopardise safety is given below.

ATC Example
A newly checked out ATCO was sitting together with two colleagues in approach. Since the ATCO was newly checked out the colleagues were extra attentive to him. There was very little to do so the ATCOs were talking with each other. The ATCO had two aircraft. The weather was good so one of the aircraft from the West was allowed to fly a visual approach. ***He told the aircraft to pass the coastline at 2.500 feet and that he could now fly to the final on his own (error #1).*** *The aircraft was then handed over to tower. Both of the colleagues of the ATCO knew that to shift an aircraft to visual approach so early (the aircraft had 15 to 20 miles left to go) then it is almost guaranteed that he will fly a very long visual approach. Another aircraft from North which had arrived through Sveda was also allowed to fly directly (it is not yet visual but is on radar vectors). This aircraft was flying very fast. The two colleagues told the newly checked out ATCO - a bit for fun - "Wow - that was early you shifted him to visual approach." So, at first it was just considered an undesirable strategy (during training/check-out you learn that you should not shift the aircraft to visual approach – you should have the aircraft on your own frequency so you can control it). When the aircraft got closer to runway 22L the two colleagues started hinting to the ATCO that now it was about time to do*

something. ***He did not intervene because he was checked out as an ATCO by now (error #2).*** *So, he did not respond and the two aircraft continued getting closer and closer. He seemed convinced that everything would be all right in spite of receiving these disconfirming information from the colleagues. The colleagues now started saying to the ATCO: "You really have to do something now - this is not going well!"* ***Still no response was made (error #3).*** *In the end the colleagues were almost shouting to the ATCO to turn one of the aircraft to the left. No other solutions were considered. Not until this point he turns the aircraft from North to the left. It might be due to his pride that he did not respond to his colleagues earlier on: now he was a fully trained ATCO and he did no longer have to listen to others. The aircraft came very close to each other and separation standards were violated.*
[Source: Personal interview with an ATCO]

These issues associated with team dynamics are important seen from an error management perspective because analyses by National transportation Safety Board of commercial aviation accidents in which crew performance was a factor revealed that in almost 3/4 of them one crew member made an error and the other either failed to detect or correct it (Orasanu et al., 1998). Also in an experimental context it has been demonstrated that many errors are less likely to be detected and challenged when they involve a person with a higher skill, judgement and competency (e.g. a captain) and the risk is high (Orasanu et al., 1998). A consequence of such insights is that the role of social dynamics can be significant in relation to whether errors are caught or not.

As shown be the previous paragraphs the ability of individuals to work together as a team is vitally important for the containment of human errors. Below are given some examples of how team skills and error management capabilities of crews or teams can be enhanced:

Attitude Change. Studies have shown that professional groups such as pilots, controllers and physicians have unrealistic self-perceptions concerning their invulnerability to stressors such as fatigue and workload (Helmreich & Merritt, 1998). This denial of vulnerability can result in a failure to use teamwork as a countermeasure to errors and stress. Focus should therefore be on enhancing the realisation that human errors are an inevitable fact of life and that teamwork is important for trapping and mitigating the consequences of errors. Actually, empirical evidence is available that demonstrates that such attitudes concerning own limitations can be changed through training (Helmreich & Merritt, 1998). A potential benefit of this is that people to a larger extent will rely on the redundancy and safeguards that can be provided by other team members instead of individual actions.

Team Self-Correction. A way of improving error management of teams is also by enhancing through exercises the general understanding of generic factors that affect the effectiveness of the team process. The goal of team self-correction is that teams should be able to self-correct co-ordination breakdowns. This requires that the team members become able to identify which specific team processes that function well and which do not. A systematic approach for developing generic team work skills has been suggested by Smith-Jentsch et al. (1998) and consists of four stages: (a) focus team members' attention on critical teamwork dimensions (such as information exchange,

communication, supporting behaviour and team initiative/leadership) during an exercise pre-brief; (b) observe team performance during an exercise; (c) diagnose strengths and weaknesses on the basis of the critical crew dimensions after exercise; and (d) guide the team through a self-critique of the performance which can then by used as a goal for a new round of the training program. Seen from an error management perspective it is interesting to note two things. First of all, the researchers behind the program state that errors should not be prohibited or corrected by an instructor, but should instead by allowed to unfold naturally without interfering with the team co-ordination (or lack thereof) during the exercise. Secondly, many of the team process skills that are being reinforced within this training program are closely associated with avoiding and controlling the effects of errors. Actually, one of the components within this program deals explicitly with monitoring and correcting team errors.

Cross Training. Another way of improving team's error management skills is by enhancing the understanding of other team member's tasks through a training strategy referred to as cross-training (Blickenderfer et al., 1998; Volpe et al., 1996). The idea behind this kind of training is that effective co-ordination and communication between team members is dependent on individual team members' interpositional knowledge - that is, their knowledge about the rules, responsibilities and requirements of other positions in the team. To achieve this goal cross-training can be used which means that each team member is trained in, or at least provided with knowledge about, the duties of his or her team-mates. Dependent on the degree of inter-dependency between team members different kinds of cross-training can be used. Several empirical tests have supported the benefit of cross-training interventions on variables such as teamwork behaviour, communication and task performance (Blickenderfer et al., 1998). In similar vein, other researchers have suggested that interpositional knowledge has an important function for a team's ability to detect and correct errors within a system (Seifert & Hutchins, 1994).

5.2 New technology

It is clear that new technology and automation will alter the controller's tasks in many ways (Wickens et al., 1998). This evolution will inevitably have many implications for the types and amount of errors that will be committed and will also most likely affect the opportunities of managing the errors that will occur. To examine in detail the potential effects of new technology will not be relevant in the current context. Instead we will limit the focus to a specific type of concept that directly deals with the issue of error detection and correction, namely what is referred to as interactive critiquing.

The concept of interactive critiquing can be seen as a specific and innovative form of decision support system. Traditionally, decision support systems have been designed in such a way that a computer tries to solve a given problem for its user and gives its results for the user to review. In this manner the human operator is given the role as the one to critique the results generated by the computer and decide whether he or she agrees with them.

There may be several problems associated with this traditional type of decision support system. All kinds of automation will be brittle in some situations – in particular, in situations that they have not been designed to handle. This can lead to problems if the operator blindly trusts the system and the user may be adversely biased by the computer in cases where it exhibits brittle performance. The more reliable a given system is the more the operator will tend to trust it and the less the chances are that the operator will be able to effectively critique the system. In addition, the operator will easily get out-of-the-loop concerning the underlying processes and assumptions leading to the (potentially flawed) results generated by a decision support system and, as a consequence of this, may not be able to effectively critique the system.

The traditional decision support tools create basically a dichotomous situation for the joint system: Either the machine does the job fully automatic or the operator does it fully manual. Interactive critiquing is a concept that has been proposed to overcome the problems associated with the traditional cooperative architecture. Instead of having the human to critique the computer the computer system will be assigned with the role of critiquing the system user's problem-solving. In other words, it should be able to detect and correct human errors without inducing any new errors.

A few empirical evaluations of decision support tools based on the principle of interactive critiquing have been made. One of the more recent and rigorous ones has been done in the domain of medicine (Guerlain et al., 1997). More specifically, the focus was on the design of a decision support system aimed at assisting blood bankers in identifying antibodies in patient's blood. The study produced several interesting results. In scenarios where the system was fully competent the participants who used the critiquing system did not produce any errors whereas subjects who did not have the system misdiagnosed cases 33% to 63% of the time. What is even more interesting is that in the cases where the system was not fully competent (e.g. brittle) the group that had the system available still displayed a superior performance. In short, the system was useful in helping users by catching errors and helping users to recover from these errors irrespective of whether its knowledge of a given case was complete or not.

To what extent interactive critiquing can be successfully transferred to the domain of ATC is currently an open question. Nonetheless, the current trend in the area of ATC is towards mid-term and long-term decision aid tools and it seems obvious to consider how these technological innovations could be integrated with the interactive critiquing concept. In this manner the ATCO could at an early point in time receive critique about plans related to the future traffic. The end result might be a higher degree of error tolerance within the air traffic system.

5.3 Summary

As can be seen on the basis of the previous sections enhancing the controller's error management capabilities may be achieved through several means such as team resource management training and by making intelligently use of technological opportunities.

These two solutions have in common that they have the opportunity to strengthen the interactive and cooperative resources within the man-machine system. Even though both of these solutions could play an important role in relation to enhancing error management this issue has been inadequately explored and elaborated in the extant research literature. Consequently, insufficient knowledge is available concerning how these concepts can be implemented in the domain of ATC and to what extent they might prove useful in supporting error management.

PART THREE

CONSTRUCTION OF THE TAXONOMY

6 The framework

6.1 Introduction

In the following pages an analytical framework based on an error management model is presented. The focus of the framework is Air Traffic Control, but the core of the framework is generic in nature and should therefore be applicable to many different domains such as aviation, process control and the maritime domain. With the framework it is possible to analyse in detail both the cognitive failure behind the error and the way it was managed. In addition, it is possible to identify from a list of Performance Shaping Factors (PSFs) the positive and negative contribution of generic contextual factors. The framework has been developed on the basis of the previously described literature review and it has been further refined and tested on the basis of incident reports, critical incident interviews and simulator studies (the results of which will be described extensively in later chapters).

6.2 A model of error management

To be able to develop a classification system of the error management process it is useful to the have a model that can be used as an organising principle. Currently few models are available to describe the generic structure of the error management process. Some of the most promising frameworks are to be found in Helmreich (1999) and Kanse & Van der Schaaf (2000b). Below is presented a model that tries to incorporate the advantages of these models.

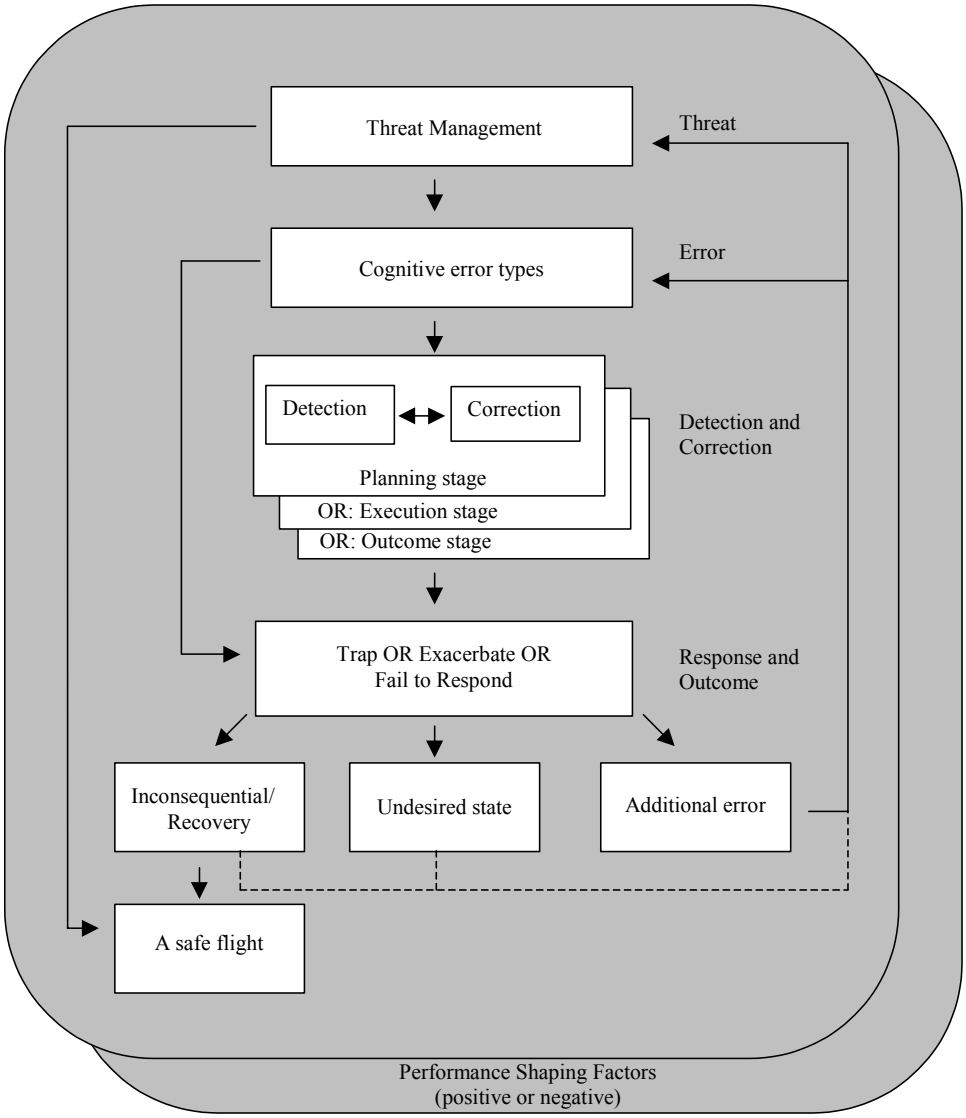


Figure 17: A model of error management

The model starts out with a threat management section which concerns the subject's awareness of aspects of the operational environment that might lead to errors and operational problems (e.g. thunderstorm). The subject may try to avoid threats leading to problems and errors resulting in a continued safe flight. If the threat is not discovered an error might be the result. The error can be analysed on the basis of the cognitive mechanisms underlying the error. The error might not be detected, but if it is the detection and/or the recovery may happen at different stages in the evolution of the error. Different kinds of responses might be produced and the result may vary from being inconsequential to an undesired state or a new error. In the case where the outcome is an additional error a new error analysis can begin. Even if the outcome is inconsequential or an undesired state new errors might still occur in the event sequence.

As described in the previous sections, the analysis unit within the framework is the individual actor involved in the error production and management. Even though the analysis unit is the individual level it is possible – and frequently is the case - that different actors are associated different stages in the model. Furthermore, a list of contextual factors – so called Performance Shaping Factors - constitutes an integrated part of the framework. These factors can be used to expand the analysis beyond the individual level to include team and organisational factors that are relevant to gain a comprehensive understanding of why the event occurred and how it was prevented from developing into an even more serious situation.

A general problem with modelling the relationship between error, recovery and outcomes is that these individual components of the error management process are ordered in a neatly and simple fashion. Instead it is often case that several errors can occur after each other and that it is only on the basis of the total outcome that a problem is discovered. That is, several errors and error recoveries might be present at the same time. How to structure the analysis of such complex scenarios will be elaborated later on (see section 8.2.2)

6.3 The main dimensions of the framework

Based on the model presented above an error management taxonomy has been developed. The taxonomy provides an opportunity to analyse the cognitive and behavioural activities of the individual actors involved in the error management process and the influence (positive and/or negative) of a series of contextual factors. The dimensions and classifications associated with the individual actions are shown below.

THREAT & ERROR							
Threat & Error	**Threat Preparedness**	No Anticipation			Anticipation		Unknown
	Cognitive Error Type	Perception	Short-term memory	Long-term memory	Decision	Response	Unknown
	Procedural violation	Yes			No		Unknown
DETECTION & RECOVERY							
***Who:* Actor**	**Error/state detected by**	No one	Producer	Co-actor in context	Co-actor out-side context	System	Unknown
	Error/state corrected by	No one	Producer	Co-actor in context	Co-actor out-side context	System	Unknown
***When:* Processes**	**Detection Stage**	Planning		Execution	Outcome		Unknown
***How:* Processes**	**Detection source**	External communication		System feedback	Internal feedback		Unknown
	Error/state correction	Ignore	Apply rule	Choose option	Create solution		Unknown
RESPONSE & OUTCOME							
***What:* Behaviour & outcome**	**Error/state Response**	Trap/ mitigate		Exacerbate	Fail to respond		Unknown
	Error Outcomes	Inconsequential/ recovery		Undesired state	Additional error		Unknown

Table 1: The analysis framework

6.3.1 Threat management

Little is known about how operators use their experience to control threats and risks of errors. A threat can be seen as a part of the operational environment which might evolve into a problem if not handled in due time. Threat management concerns being prepared for these threats and is important insofar as by knowing in advance that certain problems might occur it becomes easier to respond in a timely and efficient manner. Below are described two types of threat management.

- **No Anticipation:** In this case no indication of recognition of any threat(s) was made by any of the involved ATCOs before it developed into a real problem.
- **Anticipation:** A threat (or several threats) in the environment is known by one or several ATCOs before it leads to a problem. In some situations no concrete attempts are made at prohibiting the threat from developing into an error. In other situations explicit attempts are made at controlling it by either preventing it from leading to an error or by contingency planning (if-then).

6.3.2 Cognitive domain

To analyse the mechanisms behind the individual errors it is necessary to build upon a recognised conceptual framework that will allow both analytically useful and consistent classifications. An extensive review of existing human error frameworks has indicated that the information processing models - such as the one presented by Wickens (1992) - seem to correspond conceptually with the tasks of the air traffic controller (see also Shorrock & Kirwan, 1998; Isaac et al., 2000; Isaac et al., 2002). Below are shown the cognitive failure types that were chosen for the analysis of mechanisms behind individual errors.

- **Perception:** This cognitive domain concerns issues related to picking up and understanding information. A typical kind of error associated within this domain is hearback error. That is, a controller fails to pay attention to the content of a pilot's read back and hears instead what he/she expects to hear.
- **Short-term memory:** This domain concerns short-term storage or retrieval of information. For practical reasons it has been decided that short-term memory errors are associated with information received during an operational shift[5]. For example, a controller may forget to follow up on a potential conflict between two aircraft in spite of having intended to do so.
- **Long-term memory:** This domain concerns long-term storage or retrieval of more permanent information based on the person's training and experience. For example, an ATCO may forget to carry out a specific procedure because he has not been using it for a long time.
- **Judgement & decision-making:** Controllers are constantly required to make projection of trajectories, plan future actions and to make decisions. These activities may all be associated with errors. For example, the controller may mis-project the future position of two aircraft and consequently not consider any need to monitor them further.
- **Response execution:** Sometimes people carry out actions that they have not intended. A typical example is when a controller gives a clearance to one flight level but had intended to give clearance to another flight level. This is often referred to as a *slip-of-the-tongue*.

6.3.3 Procedural violation

Procedural violations are included as a part of the error section and they constitute within the framework a subgroup of the decision-making errors. That is, only in the case where a decision-making error has been made should the classifier determine whether it was also

[5] Please notice that the definition of short-term memory in the current context varies slightly from the one found in the research literature where it is normally said that information can only be maintained in the short-term memory store for about 10-15 sec (see e.g. Wickens, 1992). However, in the current context this time span is limited and not useful insofar as ATCOs are normally expected to maintain task relevant information in the memory for a much longer period – e.g. 10-15 minutes (Hopkin, 1995).

a procedural violation. It should be emphasised that even though procedural violations and errors within some frameworks are considered mutually exclusive (e.g. Reason, 1990) procedural violations are in the current context viewed as a subgroup of errors insofar as intentional violations usually are carried out as a short-cut as to what is seen as unnecessary procedures and regulations (Helmreich et al., 2001). The distinction between these two types of decision-making errors is of practical interest because a high degree of procedural errors might indicate too many or too complex procedures whereas a high degree of non-violation decision-making errors may indicate too few procedures (see e.g. Helmreich et al., 2001 and Reason, 1997).

6.3.4 Error discovery and recovery

The taxonomy for structuring the analysis of the error discovery and recovery is based on following questions, namely the "who"-, "when"-, "how"- and "what"-questions. More specifically these questions concern:

1) *who* was involved in the detection and recovery of the error and/or its consequences;
2) *when* was the error or its consequences detected;
3) *how* was the error and/or its consequences detected and corrected; and finally
4) *what* was the behavioural response and outcome*?*

6.3.5 The "who"-question

Research by Wioland & Doireau (1995) has demonstrated that fellow team members play a significant role in the detection of errors. An important finding in this context is, for example, that the errors people detect themselves are qualitatively different from the ones that are detected by others. On a more theoretical level it has also been suggested that detection by others is dependent on the amount of context sharing between the error producer and the error detector (Wioland & Amalberti, 1998; Hutchins 1994). In the ATM domain this would mean, for example, that an error made by a controller is more likely to be detected by a colleague than by a pilot. The following are the different possible actors involved in the detection and correction of the error or its consequences:

- **No one:** No one discovered the problem while it was still possible to solve.
- **Producer:** The person who produced the error was also the one to discover and/or recover the error (or its consequences).
- **Co-actor in context:** An observer sharing almost the whole context, goals and actions (e.g. two ATCOs or two pilots) discovered and/or recovered the error (or its consequences).
- **Co-actor outside context:** An observer sharing a significant part of goals, but not the context (e.g. an ATCO and a pilot) discovered and/or recovered the error (or its consequences).
- **System:** Some kind of automated defence – e.g. TCAS or STCA - discovered

and/or recovered the error (or its consequences).

It should be noted that the categories “co-actor in context” and “co-actor outside context” are meant as generic categories that should be adapted to the specific study that is being carried out.

6.3.6 The "when"-question

From a safety perspective human errors are not a problem in themselves as long as they do not have adverse consequences on the system (Reason, 1990). Consequently, it is of interest to determine the time of which an error was detected. Kontogiannis (1999) has suggested three different stages of performance during which error detection may occur.

- **The outcome stage:** The error is not caught until it has produced some consequences on the environment. This does not necessarily mean that an error has had *serious* consequences on system safety but only that an action has now been carried out. An example of this is that the planner discovers that he has forgotten to update a strip some time after this action should have been carried out (e.g. right after a coordination).
- **The execution stage:** An erroneous action has been carried out on the system, but the error is caught before any consequences have ensued. Typically, detection at this stage happens when the controller gives instructions to pilots or when pilots read back instructions. An example could be a controller who makes a slip-of-the-tongue (e.g. giving a wrong flight level) and then immediately corrects the error.
- **The planning stage:** Detection at the planning stage is usually associated with information-pick-up necessary for later actions or discussions and deliberations about what to do (e.g. between radar and planner controller). In other words, detection at the planning stage normally occurs before any instructions have been given or coordinations have been made.

6.3.7 The "how"-question

The how-question covers both how the error and/or its consequences were detected and how it was corrected.

Detection

As a corollary to understanding "when" an error was detected it is also of interest to obtain knowledge about the cues or mechanisms of the detection. Various researchers (Sellen 1994; Rizzo et al., 1995; Kontogiannis, 1999) have suggested a number of partially overlapping classification systems. The mechanisms relevant in the current context can be subsumed under the following categories.

- **External communication:** Interaction with other people can provide information to detect an error. That is, a problem is discovered because another person says something that is either wrong or that reveals the presence of a problem. This kind of detection requires that the cues for error detection can be found mainly in communication. E.g., a controller gives a wrong instruction to an aircraft and discovers this when the pilot reads back the clearance.
- **System feedback:** This kind of feedback relies on cues directly found in the operational environment. System feedback includes information from the radar screen and also visual sighting from either the tower or the cockpit.
- **Internal feedback:** This kind of feedback refers to error detection that requires no direct feedback from the environment. In the current context internal feedback covers following three sub-categories: (a) *Error Suspicion*. An ATCO is aware or has a suspicion of having missed something (e.g. if he fails to hear something but knows he did not hear it or recall not having heard or done something). (b) *Standard check.* An ATCO can detect an error without having any prior hypothesis about the presence of an error (e.g. one ATCO asks another ATCO if he has carried out a required action without knowing whether this is the case or not). (c) *Spontaneous memory recall.* An example could be if an ATCO gives an erroneous instruction to an aircraft and discovers it before feedback has been received from any external source of information such as the radar screen or pilots.

Recovery

In addition to identifying the processes underlying the detection, it is also of interest to understand the processes underlying the problem solving and decision-making associated with the recovery. Below is presented some distinctions, which are inspired by a classification system developed by Orasanu & Fischer (1997) to distinguish between different kinds of decision events.

- **Ignore:** Even though an error has been detected - while there still is a chance to do something about it - no response to correct it is chosen. This might be because the error is considered irrelevant or because an intervention is expected to exacerbate the situation.
- **Apply rule:** In many situations there only seems to be one thing to do in order to resolve the problem. In retrospect, several potential solutions might be available, but in situ only one solution was considered. This corresponds to what is referred to as Recognition Primed Decision-making (Klein, 1989).
- **Choose option:** In this case several options were considered before deciding on a specific solution and more conscious resources are required than the "Apply rule" category. In other words, the response is less automatic and does require some degree of deliberate resources.
- **Create solution:** This group of recovery processes is concerned with situations where a completely new response has to be generated since such situations have not occurred previously. This is the most resource demanding of the possible recovery processes.

6.3.8 The "what"-question

The final question concerns what was the behavioural response and the outcome of the error. These issues are based on directly observable phenomena and do not require any inferences about the underlying cognitive processes. Based on a model of error management developed by Helmreich et al. (1999) the following classifications have been derived:

Error/state response contains three groups:

- **Trap/mitigate:** Error is detected and managed before any consequences have developed or the consequences of the error are diminished.
- **Exacerbate:** The error is detected but the recovery action worsens the situation. This could, for example, be the case if a controller, having discovered an emerging conflict, gives avoiding instructions that actually brings the aircraft closer together rather than bringing them apart.
- **Fail to respond:** No response is produced because the error is either not detected, detected too late or simply ignored. Error may be ignored if it is considered inconsequential such as not providing traffic information to the involved aircraft after the resolvement of a conflict.

Error/outcome contains the following categories.

- **Inconsequential/recovery:** No negative consequences were observed and recovery attempts were successful.
- **Undesired state:** The end result was a potentially critical situation, an incident or accident. In the current context the most frequent undesired state is violation of the prescribed aircraft separation standards.
- **Additional error:** Sometimes errors pave the way for new errors and this may be the beginning of a chain of errors. The general characteristic of these errors is that they negatively affect workload, situation awareness or other task related factors. It should be emphasised that additional error refers to *a causality not a chronology*. Therefore, additional error should only be used in the case where there is an explicit causal relationship between two errors and it is not enough that two errors follow each other. An example of an error leading to an additional error is if a controller does not set up the radar in an optimal manner which later on enhances the risk of the controller not noticing an emerging conflict between two aircraft.

6.3.9 Performance Shaping Factors

Performance Shaping Factors are generic factors that can have a positive or negative influence (or both or none) on the course of events. They can be used to give an answer to the *why*-question – namely why did the error occur and why was it successfully or

unsuccessfully managed. The main groups of contextual or Performance Shaping Factors are shown below[6]. Please note that some of the factors are domain-independent whereas others are specifically related to Air Traffic Control.

Performance Shaping Factors		
What was the influence of these factors (positive, negative, both or none)?		
1. Traffic, airport and airspace	*Pos.*	*Neg.*
a) Traffic load/ traffic mix/ R/T workload		
b) Time available and degree of urgency		
c) Call sign similarity		
d) Air space design characteristics		
e) Airport design, facilities, or conditions		
f) Visibility of A/C and vehicles on aerodrome		
g) Temporary sector activities - military, parachuting, student pilot		
h) Weather - clear weather, snow/ice/slush, fog/low cloud, thunderstorm, windshear		
i) Other traffic, airport and airspace factors		
2. Ambient Environment	*Pos.*	*Neg.*
a) Sterility of environment (noise, distraction - supervisors, colleagues, visitors)		
b) Lighting – illumination, glare		
c) Other ambient environment factors		
3. Procedures and Documentation	*Pos.*	*Neg.*
a) Procedures (availability, compatibility, quality and usability)		
b) Operational materials – checklists/advisory manuals/charts/notices		
c) Regulations and standards		
d) Other procedure and documentation factors		
4. Workplace design, HMI and equipment factors	*Pos.*	*Neg.*
a) Radar display (interface properties)		
b) Radar coverage		
c) Transponder factors		
d) FPS (Flight Progress Strips) factors		
e) Communication equipment		
f) Warnings and alarms		
g) Automation		
h) Other workplace design, HMI and equipment factors		
5. Training and Experience	*Pos.*	*Neg.*
a) Knowledge/experience		
b) Quality of training		
c) Time since last (re)training in task		
d) Informal work practice		
e) Other training and experience factors		
6. Person Related Factors	*Pos.*	*Neg.*
a) Vigilance (fatigue, boredom, alertness)		
b) Risk-assessment/short-cuts		
c) Error coping strategies		
d) Confidence and trust in self/others		
e) Confidence in equipment and automation		
f) Emotional state (calm, chock, panic)		
g) Pride regarding a job well done/feeling of personal responsibility		

[6] The PSFs are based primarily on TRACEr (Shorrock, S.T. and Kirwan, B., 1998), HERA (Isaac et al., 2000), ADREP2000 (Cacciabue, P.C., 2001), ASAP (Helmreich et al., 1995), BASIS (O'Leary, 1999) and research on Recovery Influencing Factors (Kanse & Van der Schaaf, 2001).

Performance Shaping Factors		
h) Other personal factors		
7. Social and Team Factors	*Pos.*	*Neg.*
a) Quality of hand over /take over		
a) Language/phraseology/culture issues		
b) Brevity, timing, accuracy and clarity of communication		
b) Team climate		
c) Authority gradient		
d) Monitoring/cross-checking		
e) Assessing safety threats and planning countermeasures (if-then)		
f) Verbal statements of plans/challenging plans		
g) Review status/modification of plans		
h) Procedures selected		
i) Procedural compliance		
j) Task planning: Prioritisation/task allocation		
k) Other social and team factors		
8. Company, Management and Regulatory Factors	*Pos.*	*Neg.*
a) Company/commercial pressure - unsafe ops, failure to correct problems		
b) Regulatory – planning, decision making, feedback		
c) Management/Organisation - planning, decision making, feedback		
d) Organisation of work and responsibilities		
e) Training plan		
f) Personnel selection plan		
g) Supervision		
h) Shift patterns and/or personnel planning		
i) Management attitudes towards human error and safety issues in general		
j) Other company, management and regulatory factors		

Table 2: Performance Shaping Factors

An elaboration of the individual PSF-dimensions is given in the following

Traffic, airport and airspace
This is the only main group that is only concerned with domain specific factors. Some examples of relevant factors within this group are *traffic load*, *air space design characteristics* and *weather*. Weather is a good example of a factor that can have both positive and negative contribution. On the one hand, a strong wind can compromise the ATCO's chances of making reliable predictions of aircraft trajectories. On the other hand, a very clear weather can play a significant role for an ATCO's ability to monitor and perhaps recover a situation where the aircraft have gotten too close to each other.

Ambient Environment
This includes *lighting* (illumination, glare) and *sterility of environment* (noise, distraction - supervisors, colleagues, visitors). The latter may be particularly important insofar as many controllers feel that they "have to put up with unnecessary sources of distraction while controlling air traffic" (see Air Traffic Management, p. 26, March/April, 2001).

Procedures and Documentation
This group includes the *availability, quality and the usability of procedures and rules*. An example of how procedures can support error management is in the case with two aircraft

with very similar call signs. In this case a procedure makes it possible to alter one of the call signs and thereby minimise the risk of call sign confusion.

Workplace design, HMI and equipment factors
Traditionally the design of man-machine systems has been considered one of the most important Human Factors issues in relation to both avoiding and provoking human errors. Only to a far lesser extent has there been focus on how to promote a more error tolerant design (but see e.g. Rasmussen, 1984; Rouse & Morris, 1987; Hutchins et al., 1985). Some important issues within this group are *interface-properties of the radar display* (e.g. visibility and reversibility of actions) and *warnings and alarms* (e.g. are they reliable or do they generate false alarms and false projections).

Training and Experience
Knowledge, *experience* and *time since last (re)training* can play an important role for an operator's ability to respond in a potential critical situation. Many ATCOs feel that they have had insufficient training in handling very rare, but potentially very critical scenarios (see e.g. Air Traffic Management, p. 26, March/April, 2001). Experience with such situations can be crucial for an operator's ability to respond in a timely and effective manner.

Person Related Factors
This group contains a series of factors related to characteristics of the individual ATCO. They include *vigilance (fatigue, boredom or alertness)*, *error coping strategies*, *confidence* and *emotional state* (ranging from calm to chock and panic). Some examples of error coping strategies can be found in the section about threat management (see section 3.3.3). Trust and confidence is another important issue (Bonni et al., 2001). Self-confidence is important insofar as it concerns believing in your own abilities which helps making fast decisions. Self confidence should, however, not lead to a macho attitude (e.g. "they will not catch me saying 'no' to those aircraft" or "I do not need any help from others"). Similarly, confidence and trust in others is essential to the job. At the same time it is also important doubting and double-checking and sometimes disagreeing with decisions made by other controllers. In short, both trust and mistrust in others play a significant role in the work of the controller.

Social and Team Factors
"Social and team factors" is a very important group of PSFs - in particular in relation to error recovery – and is concerned with exchange of information between two or more people (in the current context team should be understood in the broad sense including both ATCOs and pilots, see Wickens et al., 1997). Many of these factors are issues that are given attention during Team Resource Management (TRM) training programs such as *monitoring and crosschecking* each other, clearly *stating plans and challenging potential flawed plans*. Manifestations of good and bad teamwork can be found in critical situations. In some cases ATCOs are not very good at receiving critique. It might be all right if they are sitting two together in the situation, but they do not want any interference from a third person (for example, an ATCO from a neighbour sector) unless the situation is very extreme and dangerous. In other situations a much more positive attitude toward

co-operation is displayed. This includes monitoring and crosschecking each other. In particular if an ATCO is not so busy and a colleague is busy the ATCO might monitor the situation in the colleague's airspace (D'Arcy & Della Rocco, 2001). So, in the case of an emerging conflict an approach ATCO might contact an area ATCO to check whether he or she is aware of a specific situation: "Do you have control of these two?" In a constructive team climate the attitude would be appreciable of such helpful comments.

Company, Management and Regulatory Factors

This group is perhaps the most abstract group of the PSFs and their influence on the course of events is often more subtle than the other groups, but can nonetheless be very important. This is because they can affect the whole working climate in which the controller operates. Some good examples of how organisational factors can influence the error and error management process are given in the following.

Management attitudes towards human error and safety issues in general

Different ATM organisations vary in the extent to which they assign blame to individual ATCOs. In those organisations which are characterised by having a punitive attitude towards human errors many attempts will be made at covering up the errors that invariably will happen. The consequence of this is, first of all, that the organisation will not have the opportunity to learn from these errors and the result might be that similar errors will occur again in the future and perhaps this time with a less fortunate outcome. Another problem is that people might engage in initiatives to cover up errors that might actually have negative consequences on safety. A good example of this can be found in the reluctance of ATCOs to use the term "avoiding action" on the radio (in similar vein, pilots are also reluctant to call mayday or pan-pan when having problems). The purpose of this statement is to maximise the chances of the pilots responding immediately to avoid an upcoming conflict. The problem is, however, that if this term is used on the radio then an official investigation will be made and the ATCO is likely to be blamed (and perhaps even sanctioned) for the incident. As a consequence of this ATCOs are willing to go very far to save a situation without the pilots discovering that they were close, because then no report will be made. When saving the situation in a non-dramatic way (e.g. "descent immediately" instead "avoiding action") no one will know. This is unfortunate because by stating "avoiding action" the pilot knows immediately what to do and that there is no room for arguing. If, for example, the pilot is just instructed to descend to 2000 feet the pilot might ask: "Confirm 2000 feet - we are established". By then it might be too late.

Management/Organisation: Safety-efficiency tradeoffs

Safety is normal stated as the primary goal in any ATM organisation. Even though safety is the primary goal, efficiency is also important and these two goals do not always correspond with each other. In an increasingly competitive environment it is the management's - and as well as the customer's - wish that as many aircraft as possible are started and landed. This message can be conveyed to the people at the frontline in different ways. If, for example, the traffic level is below a certain desired level (for example, due to weather conditions or the fact that airspace might be lent out to military or other purposes) questions might be raised by the administration. Another example is

the reduction in the required safety standards (e.g. to go below the three miles to 2½ miles when commencing on final approach) that has been possible due to increased precision in the tracking of aircraft. This can be seen as an indirect pressure to land more aircraft. The end result of such pressures to be more effective is that the safety margins are gradually made smaller and smaller. This pressure reinforces a tolerance among ATCOs to let the aircraft get closer and as a consequence of this the reaction time and the possibilities to counteract errors have become smaller.

Organisation of work and responsibilities

Personnel planning is another example of how organisational factors can affect the safety at the frontline. It is, for example, regulated how many days in a row that an ATCO is allowed to work. To follow these regulations can be important to avoid that ATCOs become excessively worked-out and fatigued. However, in periods with a shortage in manpower (e.g. due to the illness, holidays or insufficient amount of employees) then it might be necessary to call in staff members who have already reached their limit concerning how many days they are allowed to work. The results of this can be a decrease in vigilance that might negatively affect both the error likelihood and also the chances of detecting errors.

6.4 Analysis of a case – an example

To get a more concrete impression of how the framework can be applied to an analysis of a concrete event we will in the following do a walk-through analysis of an authentic ATC episode.

ATC Example
This event occurred at a big international airport in the Middle East. A local trainee was being checked out. When there was less activity at the airport - that is, at noon and in the afternoon - IFR-training was being carried out. The radar coverage was not very reliable. The trainee had just given a clearance out of the airport to a Gulf Air aircraft. The aircraft would climb straight ahead of runway 31 and climb to 2,000 feet. A helicopter was flying in a holding pattern at 3000 feet. The helicopter disappeared from the radar display because the radar was located on the field that causes a cone of silence. ***The Gulf Air aircraft called and the ATCO gave clearance to continue climb to flight level 160 (Error #1).*** *The instructor was alert and knew that the helicopter was still there even though it could not be seen on the radar. He therefore told the pilot to disregard the instruction and maintain 2,000 feet. When the helicopter re-emerged on the radar display the two aircraft were very close to each other. The aircraft did not pass 2,000 feet (2,000 feet is a low altitude for such a big passenger aircraft - so the aircraft would have liked to have continued the climb directly) so the vertical separation was never violated. The instructor was particularly alert in the situation due to several circumstances. First of all, he was alert to the fact that the helicopter disappeared from the radar display and had been used to working with a radar with a much better track. Secondly, the fact that it was a local controller being checked out made the instructor much more vigilant since they in general were much less skilled and qualified (instructors were*

directly told by colleagues and supervisors that when you have one of the locals on checkout you should be extra alert - many of these were totally 'green' and had had very little training before check-out). If the aircraft had continued the climb the two aircraft would most likely have gotten very close to each other.
[Source: Personal interview with an ATCO]

The analysis of threat, error and error management is described in the following (please notice that the free text included in the individual boxes below is an excerpt from the event description).

ERROR EVENT # 1																			
THREAT & ERROR																			
"The instructor was particularly alert in the situation due to several circumstances. First of all, he was alert to the fact that the helicopter disappeared from the radar display and had been used to working with a radar with a much better track. Secondly, the fact that it was a local controller being checked out made the instructor much more vigilant since they in general were much less skilled and qualified (instructors were directly told by colleagues and supervisors that when you have one of the locals on checkout you should be extra alert - many of these were totally "green" and had had very little training before check-out)." … "The Gulf Air aircraft called and the ATCO gave clearance to continue climb to flight level 160."																			
Threat & Error	*If threat(s) present:*																		
	Threat Preparedness	No anticipation						Anticipation								X	U/K		
	Cognitive Error Type	Perception		STM	X	LTM		DM		Response							U/K		
	If decision-making error:																		
	Procedural violation	Yes						No									U/K		

In the current case there were two threats that preceded the error. The first one was that the radar coverage was incomplete and the second threat was the lack of qualification of the local controllers. Both of these threats were anticipated and the instructor was ready to react immediately in the case one of these threats would lead to an actual problem. The error was committed by the local controller and the error type was a short-term-memory (STM) failure insofar as the controller forgot about the aircraft that currently could not be seen on the radar display. Since the error was not a decision-making error no selections should be made in the procedural violation row.

DETECTION & RECOVERY															
Detection															
"The instructor was alert and knew that the helicopter was still there are even though it could not be seen on the radar."															
Recovery															
"He therefore told the pilots to disregard the instruction and maintain 2,000 feet."															
Outcome															
"When the helicopter remerged on the radar display the two aircraft were very close to each other. The aircraft did not pass 2,000 feet (2,000 feet is a low altitude for such a big passenger aircraft - so the aircraft would have liked to have continued the climb directly) so the vertical separation was never violated."															
Who	**Error/state detected by**	No one		Producer		ATCO	X	Pilot		System			U/K		

	Error/state corrected by	No one		Producer		ATCO	X	Pilot		System		U/K	
	If detector ‹› "No one" or "System":												
When	**Detection stage**	Planning			Execution		X	Outcome			U/K		
How	**Detection source**	External communication		X	System feedback			Internal feedback			U/K		
	If corrector ‹› "No one" or "System":												
	Error/state correction	Ignore		Apply rule		X	Choose option		Create solution			U/K	
RESPONSE & OUTCOME													
What	**Error/state Response**	Trap/ mitigate		X	Exacerbate			Fail to respond			U/K		
	Error Outcomes	Inconsequential/ recovery		X	Undesired state			Additional error			U/K		

The instructor was the active part in the detection and correction of the error. Therefore, the detector and corrector is "ATCO" (i.e. an ATCO different from the error producer). The error was detected immediately when it was carried out (that is, "Execution" stage) on the basis of the communication made by the local controller to the pilot (that is, "External communication"). Since no other option was considered the decision-making type is "Apply rule" and the response is "Trap/mitigate". The error did not have any consequences insofar as the instructor intervened immediately and the outcome is therefore "Inconsequential/recovery".

Below is shown the different kinds of Performance Shaping Factors identified in the incident. For each of these it is possible to determine a specific PSF category based on the previously described list and to determine the type of influence (positive or negative).

DESCRIPTION OF PSFs (WHOLE INCIDENT)						
1. "The helicopter disappeared from the radar display because the radar was located on the field which causes a cone of silence."						
PSF code	**4B:** Radar coverage					
Influence	Positive		Negative	X	U/K	
2. "The instructor was alert and knew that the helicopter was still there even though it could not be seen on the radar."						
PSF code	**6A:** Vigilance (fatigue, boredom, alertness)					
Influence	Positive	X	Negative		U/K	
3. "… the fact that it was a local controller being checked out *made the instructor much more vigilant* since they in general were much less skilled and qualified (instructors were directly told by colleagues and supervisors that when you have one of the locals on checkout you should be extra alert - many of these were totally "green" and had had very little training before check-out)."						
PSF code	**6A:** Vigilance (fatigue, boredom, alertness)					
Influence	Positive	X	Negative		U/K	
4. "… the fact that it was a local controller being checked out made the instructor much more vigilant since *they in general were much less skilled and qualified* (instructors were directly told by colleagues and supervisors that when you have one of the locals on checkout you should be extra alert - many of these were totally "green" and had had very little training before check-out)."						
PSF code	**5B:** Quality of training					
Influence	Positive		Negative	X	U/K	

PART FOUR

VALIDATION

7 Validation and methodology

In the previous chapters the state of knowledge with regards to error management was described. A conceptual framework was developed on the basis of this literature review and analyses of error events in different kinds of data material. The next stage is to validate the framework on the basis of authentic data. This is important because even though the framework may appear theoretically consistent and comprehensive, it might be difficult to apply to real-life situations. In the following we will elaborate on the concept of validation and present the methodology used to evaluate the framework.

7.1 Validation

Validity is related to the degree to which a framework accurately reflects or assesses the specific concept that a researcher is attempting to measure. Validity is therefore an important concept in any research study: Without validity the results of a study become meaningless. A key concept in relation to validity is the issue of "truth" which is particularly important in qualitative and quantitative studies within the area of social sciences. In many of these studies the purpose is to build a bridge between theoretical concepts – e.g. classical psychological concepts such as memory, attention, motivation and attitudes – and observable manifestations of these concepts. One main problem is that these concepts cannot be directly observed and the scientific challenge is therefore to find a way to obtain good-enough observable manifestations of these not-directly observable constructs.

In the current context, the goal is not to operationalise a theoretical construct, but instead to develop a conceptual framework. As a consequence of this the issue of validity has here a slightly different meaning compared with many other studies in the area of social science. Since there can be developed many different kinds of conceptual frameworks to describe a given phenomenon it also becomes difficult to say whether or not a given framework is close to the "truth". A practical example can be useful to highlight this point: Does Reason's classical distinction between planning- and execution errors reflect the "true" nature of errors or would it, for arguments sake, have been more appropriate to distinguish between errors on the basis of which day of the week they occurred? Each of these classifications could, in principle, be accepted as an appropriate reflection of different types of errors. Nonetheless, common sense would lead most people to conclude that the first of these two ways of classifying errors is the one that provides most utility.

Several types of validity have been suggested. Some of these can be analysed quantitatively and others qualitatively. A research study should, in particular, be concerned with two types of validity, namely *external validity* and *internal validity*. External validity is related to the extent to which a study can be generalised to other contexts. Internal validity, on the other hand, concerns the extent to which the study is valid within a particular setting. The most important types of internal validity are:

- **Face validity:** Does the framework seem reasonable using common sense? This can be determined by having "experts" to review the contents of the framework to see if they find it useful and relevant on the basis of its face value (Reber, 1985). This issue is directly related to the evaluation criterion referred to as "usability".
- **Content validity:** The important issue in relation to content validity is the comprehensiveness of the framework. In other words, are all the major and important issues within a given research topic covered. This issue is directly related to the evaluation criterion referred to as "comprehensiveness".
- **Criterion Validity:** A criterion can be seen as an already validated and accepted standard to which a measurement or methodology can be compared. This issue is directly related to the evaluation criterion referred to as "diagnosticity".
- **Construct Validity:** Construct validity is probably the most difficult type of validity to establish and cannot be done within a single research study[7]. It refers to the extent to which evidence points to the construct or concept being useful in a scientific endeavour.

In addition to these issues it is also very important that the framework satisfies some reliability standards. Reliability refers to the extent to which a framework or measuring instrument yields the same result on repeated trails. There are basically two kinds of reliability that are important in a scientific enquiry, namely inter-rater- and intra-rater reliability. Inter-rater reliability is the consistency across judges or classifiers. Intra-rater reliability is consistency of the same judges or classifiers over time. It can be expected that a slight variation may occur in both cases, but in general the confidence in the results increases as the stability increases. It should be noted that inter-rater reliability is considered the most crucial type of reliability and is therefore also the one that will be given most credence in the current context.

Steps towards validity

In the model below is illustrated three important steps in relation to achieving reliable and valid analyses in a given classification study, namely the information elicitation, the segmentation and classification stage. For each of these stages there are some factors that can affect the results from the given stage. The model is not meant to portray the development of a conceptual framework, but instead the stages of importance when conducting an analysis on the basis of an already developed framework. In the development of a conceptual framework it is not necessarily such that one stage is completed before the second stage is initiated. Instead it will often be necessary to jump back and forth between the stages, because results from one stage often will have effect on the other stages.

[7] The concept of construct validity and criterion validity are closely related to each other. In the current context criterion validity can be seen as an operationalisation of construct validity. That is, even though construct validity cannot be measured as such it is possible to enlist some criteria concerning how the framework should behave on the basis of theory and existing research.

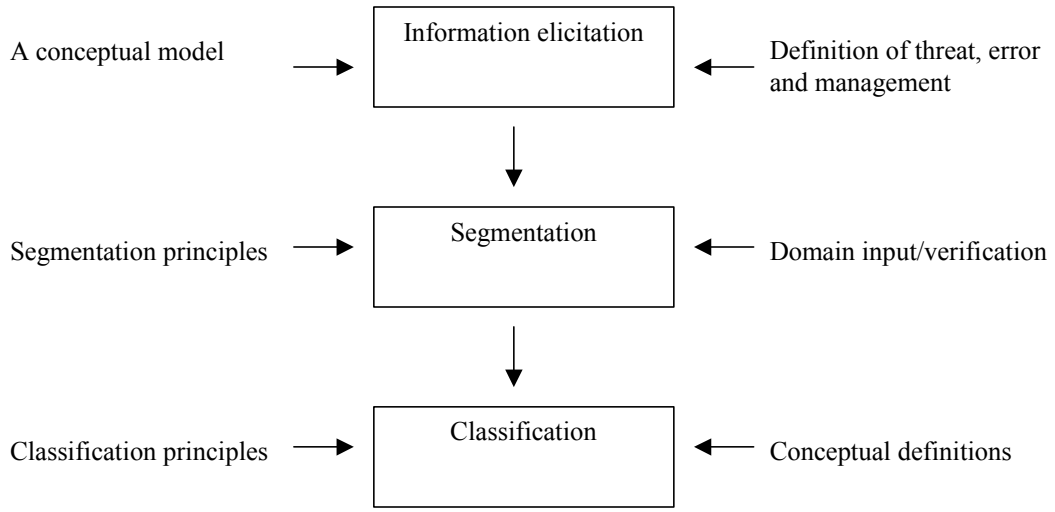


Figure 18: Main steps toward validity.

- *Information elicitation:* In this phase it is important that sufficient information is derived from the available data sources (e.g. interviews, video-recordings and think-aloud-protocols) so that later detailed analysis can be carried out. To achieve this goal it is important to have some sort of conceptual model that can help guiding the search for information. It is also important to have a clear definition of the central concepts within this framework – in the current context in particular what is meant by the concept of threat, human error and error management.

- *Segmentation:* This phase is about breaking the event description down into groups that fit into the conceptual model. The importance of the segmentation phase is related to the fact that complex real-life events are often characterised by having several people involved in both creating and resolving a potential critical situation (actually, in some cases different people might have divergent influence on the recovery). If such events should be segmented in a way that permits carrying out statistical analyses it is desirable that the events can be adapted into a workable format and at the same time avoid loosing information. It is important that the segmentation of events is done on the basis of logical and consistent principles. In this context it can be useful to obtain input from domain experts to ensure that, for example, the identified errors also are errors according to their judgement.

- *Classification:* It is important to have principles that can be used in the analysis of the segmented events. The definitions of the individual concepts within the framework (as well as examples of their application) are also important in this context.

7.2 Methodology

There are several potential ways of obtaining authentic data to develop and refine the taxonomy:

- *Incident/accident reports.* Such reports contain detailed information of critical events and analysis of why they happened. They can be useful in relation to analysing human errors and factors that may have provoked their occurrence. However, only limited information concerning error management can be derived from such reports (in particular accident reports, because they to a lesser extent will contain error management initiatives). Therefore, incident and accidents reports can, in themselves, only be of limited value in the development and evaluation of an error management taxonomy.

- *Critical incident technique.* This technique can be described as a retrospective interview strategy with the goal to elicit information about non-routine incidents (Flanagan, 1954). This knowledge elicitation technique is done by the use of a semi-structured format to probe different aspects of the process and circumstances of an incident of interest. A variation of this technique has been applied in the study of human errors (Jensen, 1997). The advantage of this method is that it is possible to obtain naturalistic data derived from everyday life and that it is possible to elicit detailed information about the error management events. Even though the method can be useful in obtaining detailed information about error management events, it would probably require a significant amount of resources to produce a comprehensive database on the basis of this technique (in particular because the interview should ideally be carried out shortly after the occurrence of the incident to obtain reliable information).

- *Real-time studies.* Real-time studies consist of simulator experiments and real operational task. Such studies can be useful in relation to studying error management in realistic circumstances and it is normally possible to obtain many different sources of information to support the analysis (such as a/v recordings, recordings of eye movements, etc.). The disadvantages are: (1) a considerable amount of domain knowledge is required to be able to analyse the errors and recoveries and (2) a high level of resources must be invested in relation to data collection and analysis.

- *Diary studies.* By having people to do self-reports of errors committed and recovered it is possible to obtain a rich source of descriptions that can be used in the development of an error management taxonomy. Furthermore, it is easy to administer and does not require as many resources. A potential disadvantage is that the quality of the data obtained is to a large extent dependent on the conscientiousness and willingness of the people filling out the diary reports. Furthermore, we cannot be sure with regards to the representativeness of the frequency of the errors and recoveries reported in diary studies. Even though the quantities of different categories should be treated with some caution, diary studies can be useful as "wide-gauge trawl nets" to

catch a qualitatively representative sample of important distinctions (Reason & Lucas, 1984).

Clearly, the different methodological approaches described above have both advantages and disadvantages. In the table below is provided an overview of some of the important differences in strengths (i.e. '+') and weaknesses (i.e. '-') of these approaches.

	Many types of data material available	**Represen-tativeness of errors and ma-nagement types**	**Details about error manage-ment**	**Realistic setting**	**Limited resource require-ments**
Incident/ accident reports	+	-	-	+	+
Critical incident technique	-	-	+	+	-
Real-time studies	+	+	+	(+)	(-)
Diary studies	-	-	+	+	+

Table 3: Advantages and disadvantages of different methodological approaches

The individual items in the table are elaborated below:

- *Many types of data available:* Some types of data material are based on a series of different kinds of information sources whereas other kinds are only based on a single source of information. By having several sources of information available it is much easier to verify information and to get a comprehensive understanding of the episodes being examined. For example, incident reports are normally based on several types of information such as radar and voice recordings, interviews with ATCOs and sometimes also pilot reports. In similar vein, a real-time study can e.g. include video recordings and post-hoc interviews. In contrast, both the critical incident technique and a diary study rely heavily on memory recall. Since human memory is unreliable there will be some limitations concerning how accurate information that can be obtained.
- *Representativeness of errors and management types:* Different kinds of data material might be associated with different kinds of biases and therefore conclusions from such studies should be considered with precautions. In particular, incident reports and the critical incident technique might not contain a representative sample of error and error management events. On the other hand,

real-time studies should provide a more accurate picture of error and error management events as they occur during normal operational practice.

- *Details about error management:* This item is concerned with the extent to which information about error management is available or can be obtained. In this context incident reports might not contain detailed description of the processes underlying the error management. On the other hand, in the three other types of studies it is possible to obtain a more detailed description of these processes.
- *Realistic setting:* The more realistic the setting is in which the study is carried out the better are the chances of generalising the results outside the context being studied. All of the studies contain a high degree of realism. The only type of studies where the realism can be slightly reduced is in the case of simulator studies (a subgroup of real-time studies).
- *Limited resource requirements:* This item is related to the amount of resources required for conducting a study based on the specific type of data material – in particular in relation to obtaining the basic data. In the case of incident reports and diary studies other people carry out the data production. In contrast, in the case of real-time studies far more resources must be invested in obtaining the data material.

An initial attempt to use diary studies was attempted but failed because no one filled out the diary report forms. As a consequence of this it was decided to focus on the three other types of studies. Real time studies, incident studies and the critical incident technique could all be useful in the development of an error management taxonomy, because they allow gathering of a large corpus of error management events. In the current context it was chosen to start out with incident reports, because this method allows gathering of naturalistic data from operational activities over an extended time frame. At the same time the method does not require a large amount of resources in the basic data collection process. Therefore, this approach seems most appropriate in this initial stage of the taxonomy development. However, since the completed incident reports will not allow additional inquiries concerning specific issues of interest, the incident approach will be supplemented with a real-time study and some interviews based on the critical incident technique.

In addition to these studies aimed at applying and developing the framework it was also decided to do a questionnaire study where the goal was to get some input from human factors experts concerning the relevance of the individual dimensions and the overall structure of the framework. Hereby it would be possible to get an indication of the face and content validity of the framework.

7.3 Hypotheses

As mentioned before criterion validity concerns whether the framework is able to produce results that is in conformity with pre-established hypotheses. Some a priori hypotheses about error and error management *and* the relationship between dimensions within the framework are given below. These are mainly based on existing empirical

results but to some extent also on theoretical inferences. The empirical studies in this thesis will be used to explore whether the framework can verify these statements.

Error:

- Hypothesis 1: *Long-term memory errors will be more frequent among novices*. The argument for this is basically that long-term memory (LTM) errors most frequently are the result of insufficient experience with a certain task. Therefore, experienced controllers should be less susceptible to LTM errors compared with novices.

Error and error detector:

- Hypothesis 2: *Response Execution errors are most frequently self-detected.* A series of studies have previously established this relationship (see e.g. Nooteboom, 1980; Woods, 1984). For example, in a study from an ATC microworld it was shown that the errors most frequently recovered were slips and to a far lesser extent rule- and knowledge-based mistakes (Wioland & Amalberti, 1998). The explanation for this is that the criteria for successful performance are to a large extent directly available in the head of the error perpetrator. The chances of discovering response execution errors may, on the other hand, be more difficult for external observers because they do not have access to the intentions underlying the behaviour (Wioland & Doireau, 1995).

- Hypothesis 3: *Decision-making errors are either not detected at all or detected by others.* In an experimental study of emergency scenarios in a nuclear power plant by Woods (1984) it was determined that none of the diagnostic errors (i.e. a subgroup of decision-making errors) were noticed by the operators themselves. On the other hand, the diagnostic errors that were detected were discovered by external agents with "fresh eyes". In a study by Wioland & Doireau (1995) where pilots and instructors viewed video recordings of scenarios where actor pilots committed errors it was demonstrated that only a small part of the inserted errors were discovered by the observers. However, those errors discovered had a tendency to be associated with problem solving and decision-making (i.e. rule- and knowledge-based) rather than slips. In short, both of these studies indicate that decision-making errors will be difficult to discover and if they become discovered this will frequently happen through the assistance of others.

- Hypothesis 4: *Long-term memory errors are either not detected at all or detected by others*. LTM errors share some of the characteristics with decision-making errors and are therefore also expected frequently to either not be detected or be detected by others. However, LTM errors are probably easier to discover (and agree upon) than decision-making errors by a trained observer insofar as a standard for determining successful performance might be more readily available.

- Hypothesis 5: *Error detection by others depends on the amount of context-sharing.* As previously described, several researchers have suggested that the amount of context-sharing is critical for the chances of discovering errors committed by others (Wioland & Amalberti, 1998; Hutchins, 1994; Seifert & Hutchins, 1994).

Error and detection stage:

- Hypothesis 6: *Response Execution will be more frequently detected at the execution stage*. Basically, if people have a clear expectation concerning what action they intended to carry out it should be easy to detect their execution errors by comparing the action carried out with what they felt, saw or heard. Studies have shown that response execution errors are frequently caught and corrected by a direct feedback-checking (e.g. Rabbitt, 1966).

- Hypothesis 7: *Errors found in incident reports will have a tendency to be more frequently detected at the outcome stage compared with errors committed in normal operations*. Errors that are detected at the planning or execution stage will tend to be omitted from incident reports, because they are not considered relevant for the investigation. That is, since the focus is on factors that directly or indirectly affected the incident – and not factors that could have affected the situation if not caught at such an early stage – they will not be described in the incident report. Instead, these fast and effective corrections will only be apparent when observing normal operations.

Error correction and problem solving:

- Hypothesis 8: *The problem-solving process associated with error recovery will vary in such a way that 'Ignore'/'Apply rule' will be most frequent and 'Choose option'/'Create solution' the least frequent*. The reason for this expectation is that the taxonomy is here very similar to Rasmussen's SRK-model. Within the SRK-framework it is postulated that the behaviour of experienced operators will most of the time be controlled at the lower resource demanding levels (skill- and rule-based level) and only rarely it is required to move up to the resource intensive level (knowledge-based level). In the current framework the "Ignore" and "Apply rule" are the cognitive processes that require the least mental resources – that is, a straightforward recovery solution is available in the situation. On the other hand, the categories "Choose option" and "Create solution" are associated with an increasingly more cognitively demanding recovery situations.

- Hypothesis 9: *The errors that are ignored and tolerated are frequently inconsequential.* In a study based on an ATC microworld by Wioland & Amalberti (1998) it was demonstrated that with increased expertise – and thereby better knowledge of the system and its risks – the subjects tolerated a larger degree of errors without consequences. This is most likely related to the fact that the subjects learn that certain errors are without consequences and consequently

they can save resources by not correcting them. In similar vein, in a study by Orasanu et al. (1999) carried out in a full-mission flight simulator it was demonstrated that more errors were missed (i.e. not corrected) by both captains and first officers when risk was low than when risk was high. It does not follow directly from these studies that errors ignored will have a larger tendency to be inconsequential than errors responded to insofar as the results from these studies only concern the errors not-responded to. Nonetheless, it is the expectation that errors (including procedural violations) that are ignored will tend to be less serious than errors that are judged to require some action to maintain control of the situation.

Errors and their consequences:

- Hypothesis 10: *Response execution errors (including speech or action errors) should be easier to detect than other errors (e.g. lapses and mistakes).* This is related to the fact that there is a direct discrepancy between intention and corresponding action or outcome and, consequently this should be easy to discover.

- Hypothesis 11: *Decision-making errors are more often associated with undesired states*. This is, for example, supported by a study by Wiegmann & Shappell (1997) that showed that decision or response selection errors (in the current context just referred to as decision-making errors) were more frequently associated with serious accidents. Conversely, minor accidents were associated more with response-execution errors than with major accidents. Also Klinect et al. (1999) have shown that operational decision errors as well as proficiency errors were the most difficult for the flightcrews to manage and, consequently, were the ones that most often had consequences. The explanation for this is that for decision-making errors (including reasoning, judgement and diagnosis) the criterion for detection is not directly available in the head of the individual, but instead the correct solution is only available in the external world and is often not clearly recognisable in advance (Reason, 1990).

- Hypothesis 12: *Most errors in everyday-life situations will be inconsequential*. For example, in an observational study of pilot crew errors during normal operations it was found that about 85% of the crew errors were inconsequential (Klinect et al. 1999). Therefore, a larger amount of consequential errors is expected to be found in incident reports compared with real-time observation.

- Hypothesis 13: *Procedural violations will frequently be inconsequential.* A study by Klinect et al. (1999) based on real-time observation of pilot's behaviour showed that intentional non-compliance errors (i.e. procedural violation errors) were the most frequently committed and also the least consequential. It can be speculated that the reason for this is that people develop a meta-knowledge based on experience concerning which violations that are consequential and which are not. This would be in agreement with studies indicating that people develop

natural risk-talking abilities and that their main goal is not to avoid errors, but instead to maintain cognitive control (Wioland & Amalberti, 1996). Consequently, many "minor" violations might be accepted, because the risk is considered small or absent.

8 Study 1 – Incident reports[8]

In order to expand on the existing knowledge about human error capture a series of Air Traffic Control incident reports from the Swedish CAA (Civil Aviation Authorities) containing both consequential and non-consequential controller errors have been reviewed and analysed. The reason for focusing on incident reports as a basis for the analysis is, first, that a relatively large number of *fairly detailed* reports are available which is important in relation to statistical analysis and, second, that the reports provide a high level of operational fidelity compared with, for example, laboratory based research (Wickens & McCloy, 1993). Furthermore, since some error compensation behaviour to hinder an accident is normally present in incident reports, the recovery aspect will often be conspicuous in such reports (van der Schaaf et al., 1991). Even though the incident reports are also associated with inherent problems and biases (as will be discussed in more detail later on) they constitute a useful starting point for analysing human error recovery events.

Since little research has been done in relation to using a comprehensive error capture analysis framework to dynamic and complex real-life scenarios, the current study is explorative in nature. Being explorative, the goal of the study is, first, to get an indication of the robustness of the core of the classification system (that is, can consistent and reliable classifications be obtained by different judges?) so as to improve the taxonomy. Second, the goal is to determine whether the framework can be used in relation to uncovering error and recovery patterns in the database material.

The agenda for the remainder of this chapter is as follows. First, a presentation of the dimensions of the framework that will be used in the current study is given. Only the dimensions of the core of the framework that can be usefully applied to the current data material are included. The description of the framework is followed by an empirical study aimed at applying and evaluating the usefulness of the taxonomy. Finally, a discussion of the results and the chosen methodological approach will be made.

8.1 The analysis framework

There are two main dimensions in the core of the framework: the first main dimension concerns classification of the error itself (analysis of threat anticipation/management and recovery-planning is not included in this study due to insufficient information in the incident reports but will be explored in study 4). The second dimension concerns what happens after the error, namely the detection and recovery of the error. An overview of the part of the framework used in this chapter is shown in the table below containing an

[8] The chapter is based on Bove, T. & Andersen, H.B. (2000): "Types of Error Recovery in Air Traffic Management". 3rd International Conference on Engineering Psychology and Cognitive Ergonomics.

example from one of the reports analysed[9].

DESCRIPTION OF ERROR AND RECOVERY # 1						
"When SAS asks for clearance to flight level 310, R1 could have discovered the risk for a conflict insofar as the strips for SAS and SCW were available. The conflict was detected by a relieving ATCO when the separation standards between the two aircraft were violated."						
ERROR						
Error	Cognitive Domain	Perception	Memory	Decision	Response	
DETECTION & RECOVERY						
Who:	Error/state detector	No one	Producer	Colleague	Pilot	System
	Error/state corrector	No one	Producer	Colleague	Pilot	System
When:	*If detector is "No one" or "System" then go to "What":*					
	Time of detection	Planning		Execution		Outcome
How:	Detection cue(s)	External communication		System feedback		Internal feedback
RESPONSE & OUTCOME						
What:	Error/State Response	Trap/mitigate		Exacerbate		Fail to respond
	Error/State Outcomes	Inconsequential/recovery		Undesired state		Additional error

Table 4: The analysis framework (study 1)

Please notice that if "No one" or "System " is chosen in the identification of the detector then the *When*- and *How*-questions should be omitted (i.e. they are cognitive classifications and are only applicable to situations where human actors are involved in the process).

8.2 Method

In the following is reported an empirical study aimed at applying and evaluating the usefulness of the taxonomy. More specifically, the goals of the study are (1) to determine the reliability of classifications made by the use of the framework; and (2) to apply the framework to the analysis of a database of ATM incident reports to uncover error and recovery patterns.

8.2.1 The data material

Altogether 45 Swedish Air Traffic Management incidents (1997-98) were used for the study. Each of the incidents has been investigated and reported by the Swedish CAA (Civil Aviation Authorities, Air Navigation Services Dept.). The Swedish reports are particularly informative not least because the Swedish ATM provider is regarded as having largely succeeded in developing a *no-blame culture* and, therefore, these incident reports are often rich in detail and appear to be candid. In general, an incident will first be

[9] Please notice that in this study it was chosen to collapse short-term memory and long-term memory. Hence "Memory" concerns both short-term and long-term storage or retrieval of information. The reason for this was that the data material used in this chapter would frequently not allow determining whether the person had the right intention but forgot to carry out a task (i.e. short-term memory) or if the person could not recall more permanent information based on training and experience (i.e. long-term memory).

reported to the Air Navigation Services branch on a reporting form by one or several controllers involved in the incident. On the basis of this report it is decided whether or not an investigation should be carried out. If so, the report along with radar and voice recordings, interviews with the involved controllers, and possibly pilot reports, form the material on the basis of which the investigator generates a narrative description and an analysis of the course of events, draws conclusions about the involved precursors of the incident (and, of course, the recommendations that follow from the results of the investigation). Normally, an incident report will describe several human errors as well as human and organisational factors that may have negatively affected human performance. Most of the incidents involved violation of separation standards, but only in very few cases there was an imminent risk of collision.

8.2.2 Procedure

The 45 Swedish Air Traffic Management incident reports were reviewed and a total amount of 144 controller errors were identified[10]. Even though many pilot errors were also observed they were not included in this analysis because the focus was on controller errors (for comparable studies focused on the pilot side please refer to Degani et al., 1991 and Sarter & Alexander, 2000). The analysis procedure was divided into two phases: (1) a calibration trial where the incident reports from 1997 were coded by two judges (independently) and afterwards any problems and disagreements in the classification principles were clarified and resolved; (2) a test trial where the incident reports from 1998 were independently coded by two judges on the basis of the lessons learned from the calibration trial. The results presented in this chapter concern, first, the reliability of applying the error management framework (and the data behind this inter-rater reliability derive from the 81 events of the 98-incidents); and second, the frequency tabulations that derive from applying the error management framework (the data behind the frequency tabulations derive from the 144 events of the 97- and 98- incidents).

When analysing error and error capture in a complex domain such as air traffic management some difficulties will inevitably arise that require general decisions concerning how the analysis of the data material should be carried out. Some important questions and issues revealed during the calibration trial (and previous experiences with analysing human errors) are how the *segmentation* of the events should be carried out and which principles should underlie the *classification* of identified error and error management events. First, however, we will briefly examine the limitations associated with the *information elicitation*.

Information elicitation

The data material used in the current study is already completed incident reports. Consequently, it was not possible to obtain any additional information. This puts some limitations on the analysis that can be done on this basis. In particular, little information

[10] The identification of errors was done by the author of the thesis on the basis of a set of pre-determined principles.

will be available in relation to the processes underlying the error management since the investigators have had the main focus on errors and their causes rather than what occurred after the error and they have not had any conceptual error management framework to guide their information elicitation.

Segmentation

In this section we will examine some principles related to ensuring a consistent segmentation of the incidents. Some important questions are: (1) Which principles should govern the identification of error events? (2) Which principles should underlie the segmentation of error management events when errors have different causal effect on the course of events? (3) How should the error management be segmented when several interacting errors are only discovered when leading to an undesired outcome?

Principles for error identification

The following are issues that should be clarified in order to make a consistent identification of human errors.

- *How many errors in one action?*

Problem: A single act or non-detection (an inaction) may well consist of a chain of failures - for instance, a controller may have a strong expectancy about an aircraft's flight level (FL) and then overhear FL from pilot (wrong hearback), fail to check strips and fail to monitor radar. Thus, the controller misses several opportunities for correcting the erroneous FL.
Solution: Each single failure - to conduct right hearback, to check strips, to monitor radar - involved in a single inaction should be classified as an error; so in this case, three individual errors are involved, each of them influenced by "expectancy bias".

- *How many errors should be counted when two controllers share an error?*

Problem: What should we do when two controllers share an error? For instance, if the co-ordination between two controllers in a relief situation has been carried out in a hasty manner and each of them, had he or she followed procedures and good working practice, would have carried it out much more thoroughly - is this one or two errors?
Solution: We count errors by individuals since the framework used is cognitive in nature - so if two controllers are involved in the transmission of information and there is no evidence that one of them made a single mistake that explains the flawed information sharing, we count this as two instances of error.

The principle of causal relationship

Another segmentation principle of a more complicated nature is related to the fact that incident scenarios often involve several errors with variable effects on the course of events. Some errors have a direct causal effect on the course of events whereas other

errors have a more indirect effect or no effect at all on the situation. For these different kinds of errors it is necessary to determine the stopping rules for the individual error management analysis and thereby also which actor (if any) who should be associated with the error capture. Below three kinds of causal groups and their respective stopping rules in relation to the error management analysis are described. The general principle behind these stopping rules is that the relevance of an error event depends on the membership of the different causal groups.

- *Direct causes:* If the error had not been made the incident would most likely not have occurred or at least been less serious. Therefore, even though mitigating actions are initiated the end outcome is normally an undesired state. An example could be a controller not detecting a conflict between two aircraft on the flight strip board. *Stopping rule:* If an error has direct causal consequences for the situation it is reasonable to continue the analysis until the point where either no recovery is any longer possible or where recovery initiatives have been successfully implemented.

- *Indirect causes:* The error enhanced the potential for additional errors. This kind of error is normally not discovered and corrected by anyone until additional errors have occurred. An example is a radar controller who does not request assistance from a planner in spite of an increasing workload and thereby enhances the risk of new errors occurring. *Stopping rule:* The natural stopping point in the error management analysis for errors whose main function is to pave the way for new errors seems to be when these errors have had their probable consequences, namely provoking additional errors.

- *Non-causal:* The error did (most likely) not play any role in the incident. Such errors may, for example, be caught at the planning or execution stage (and thereby be prohibited from affecting system safety and performance) or simply be of a noncritical nature. An example of the latter is in the case where a controller does not provide the aircraft with traffic information immediately after a conflict is over. *Stopping rule:* Since errors of this kind do not affect the course of events in any detectable manner the error management analysis is continued until the end of the incident description.

The problem of multiple causes

Most incidents and accidents are the results of multiple errors interacting with each other. This fact has direct consequences on the analysis of error management. If only a single error occurred in each error scenario it would be a simple matter to link the error with the consequences, but when several errors are present it becomes more difficult to untangle the effects of the individual errors and recoveries because they are often tightly intertwined. This is especially the case when errors are not discovered until they have adverse consequences. In such situations several causal errors may share a common error management history. This may impose some problems when adding up error management data because the same error management events may be included in the

material several times[11], but it seems to be the only viable solution.

Classifications

In this section we will examine some principles that can enhance the chances of using consistent classification principles in the error management analysis. In particular we will examine two questions: (1) What should be done in those cases where more than one classification seems appropriate for a specific error event? (2) Which principles should underlie the classification of error detector and corrector?

Dual classifications

The categories in the analysis framework have been selected on the basis of being mutually exclusive. There are important reasons for this. First of all, it is difficult to produce a statistical index of inter-rater reliability if categories are not mutually exclusive, simply because the statistical test relevant for this task - Cohen's kappa - requires that the categories are mutually exclusive. Second, if the categories within a given dimension are not mutually exclusive it becomes more problematic to make trend analyses and analyses of interactions between variables. Even though the categories are mutually exclusive some cases occur where the classifier may feel that several categories can apply and it is therefore necessary to decide what to do in these situations. In the current context it was decided that all dual classifications would be omitted from the data material analysed. Even though this means losing some information it was considered the best solution in order to avoid the above-described problems.

Classification of error detector and corrector

A consistent identification of the error detector and corrector is actually a very significant part of the error capture analysis process. This is related to the fact that all the following steps in the error capture analysis are directly dependent on who is assigned to these two roles. If, for example, disagreements between different analysts should occur due to some inherent ambiguities in the data material or inconsistencies in the applied classification principles, this will directly undermine the reliability of the classifications, because the choice of actor determines which perspective the error management analysis will be made from. The identification of actors associated with the error detection and correction has therefore been governed by some pre-determined situationally driven classification principles.

[11] It is interesting to note that among those studies that have focused on classifying error detection and recovery episodes, there has been a tendency to use scenarios that involve a single error and behaviour aimed at mitigating the error (e.g. Sellen, 1994; Degani et al., 1991; Sarter & Alexander, 2000). Furthermore, in those studies where several errors could occur within a single scenario (e.g. Bagnara et al, 1989; Wioland & Amalberti, 1998) no explicit explanations or comments are made in relation to this problem concerning multiple causes and error management analysis.

- *The principle of first involvement:* First of all, the error detector and corrector are, in general, assigned to the one who was the first to discover and recover the error. For example, if the error producer discovers an error and afterwards a colleague also (independently) discovers the error, it is the error producer who will be chosen as the error detector.

- *The principle of involvement importance:* Different parts of the consequences may be recovered (or attempted to be recovered) by different actors. In some situations the first response is not necessarily the most relevant one (for example, if an ATCO gives an avoiding instruction to a pilot and the pilot responds in the opposite way). In cases where the principle of involvement importance is in conflict with the principle of first involvement, the first overrules the latter.

- *The principle of active involvement:* The third principle is that the actor should be actively involved in the process. If, for example, a controller detects an emerging conflict on the basis of a conflict alert this detection would be attributed to the system and not the controller, because the latter is only a transducer in the process.

- *The principle of involvement opportunity:* Finally, an error detection and correction can only be attributed to a system or human actor as long as it is discovered in due time to be able to initiate some recovery initiatives. Therefore, if an error induced problem is not detected until, say, two aircraft have passed each other and are no longer in conflict, it will be analysed as not detected by any one.

8.3 Results

8.3.1 Reliability analysis

Kappa is a statistical measure that is commonly used to determine the reliability of classifications made by independent judges and which is corrected for chance agreement (Cohen, 1960; Fleiss, 1971, 1981). In this section we report the results of the kappa analysis for each of the main dimensions in the framework. The data material used for this analysis is those classifications of error events where each of the raters has used one and only one category in the classification system. Only the reports from 1998 (that is, "81 error events") were used for the reliability analysis, since the reports from 1997 were used as a means to attune and refine the classification scheme.

During the analysis it became clear that for two of the dimensions - namely the when- and the how-dimension - it was necessary to collapse categories to be able to make consistent classifications. Finer grained distinctions were not possible to make in relation to these process dimensions because insufficient information was available in the incident reports. Therefore the following categories were collapsed: (1) For the when-dimension

the planning and execution stage were collapsed since these stages concern detection before any consequences have ensued. (2) For the how-dimension the categories external communications and system feedback were collapsed to external feedback so that the main distinction is between internal and external feedback. The following table displays the results of computing the chance corrected coefficient of agreement between two independent raters[12].

Error		
Cognitive domain		
Kappa	0.81	P-value<0.001
Detection and recovery		
Who – detection		
Kappa	0.64	P-value<0.001
Who – correction		
Kappa	0.62	P-value<0.001
When		
Kappa	0.56	P-value=0.005
How		
Kappa	0.62	P-value=0.025
Response and Outcome		
What – response		
Kappa	0.45	P-value<0.001
What – outcome		
Kappa	0.51	P-value<0.001

Table 5: Inter-rater kappa coefficients and P-values for each of the main dimensions in the framework (study 1)

The interpretation of the level of agreement (above chance) obtained by independent raters is, by convention, nearly always stated along the lines suggested by Fleiss (1981) or Landis & Koch (1977), who differ only slightly. They suggest that levels below 0.40 show poor or merely fair agreement, and this figure remains a conventional cut-off point (rather as the interpretation of a p-value at or below 0.05). Fleiss proposes that levels above 0.75 show strong agreement, and Landis & Koch suggest that levels between 0.41 and 0.60 indicate moderate agreement, between 0.61 and 0.80 substantial, and above 0.80 almost perfect agreement. As can be seen from the table each of the dimensions in the framework produced kappa values that lie between 0.45 and 0.81. Therefore, all the dimensions in the framework produced results ranging from a fair/moderate to a strong or substantial level of agreement.

It is interesting to note that the kappa values for the cognitive domains are significantly higher than the kappa values from the error management analyses. There may be several

[12] The two independent judges were the authors of the paper Bove & Andersen (2000) who rated the target material - the Swedish ATM incident reports - independently for the reliability data (all 98-reports).

reasons for this. The most important reason is probably that analysis of the mechanisms behind the individual errors can be carried out by examining a very limited part of the incident description. On the other hand, analysis of the error management process requires to a larger extent integration of the whole incident description to be able to derive the error capture classifications.

8.3.2 Pattern analysis

For the analysis of patterns in the incident reports only the part of the data material where consistent classifications were independently obtained by the two raters are used. All the reports from 1997 and 1998 were used in this analysis (that is, 144 "error events"). For the statistical analysis of the distributions is used an exact Goodness-of-Fit test based on a uniform distribution (i.e. the observed distribution is compared with a distribution where each of the categories has an equal likelihood of occurring).

Below is shown the distribution of categories for the cognitive domain. As can be seen in the figure, a large majority of the errors were either decision and judgement errors (61.5%) or perception errors (26.9%). The differences are highly significant, $X^2(3, N=104)=86.54$, $P<0.001$.

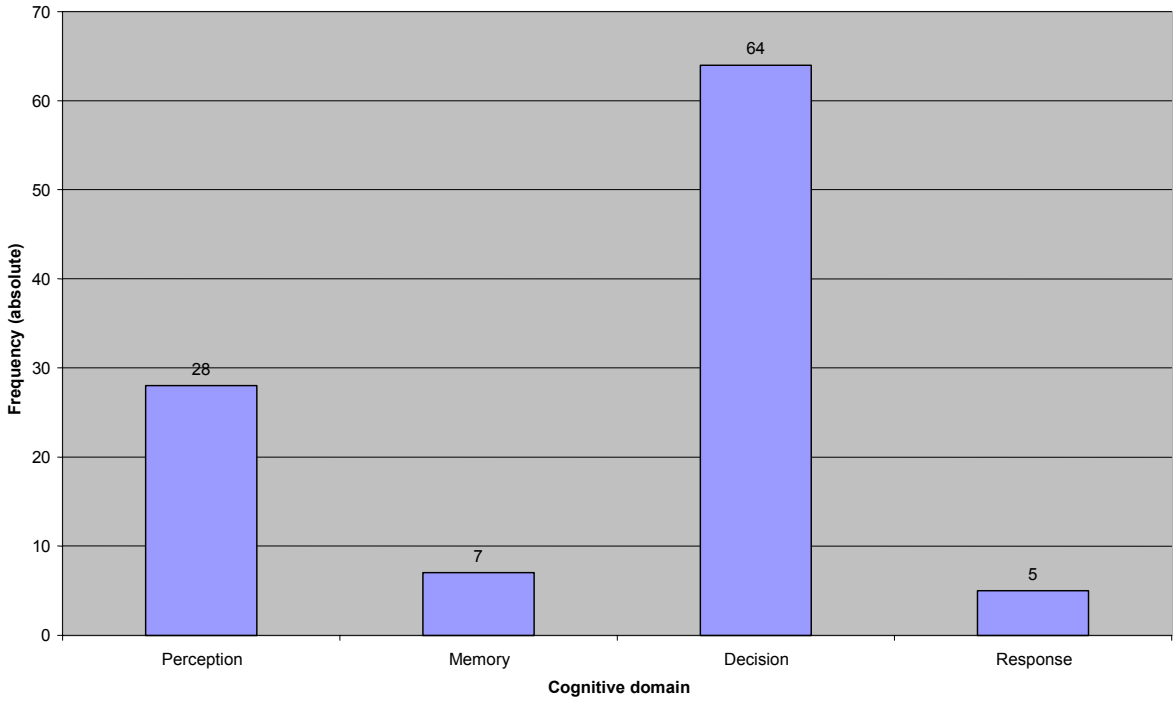


Figure 19: Distribution of cognitive domains

The results in relation to the actor associated with the error detection ($X^2(4, N=80)=42.89, P<0.001$) and error correction ($X^2(4, N=86)=87.00, P<0.001$) are shown below. It is interesting to note that a very large part of the errors are not detected at all (42.5%). Among those errors that are detected it is primarily the producer of the error who eventually also detects the error (30% of all the errors or 52.5% of the errors detected). Nonetheless, colleagues also play an important role in the error detection (18.8% of all the errors or 32.6% of the errors detected). These results are in good concordance with other studies (e.g. Wioland & Amalberti, 1998) and goes to highlight the importance of other people in the error management process. It may appear surprising that pilots are only involved in a small part of the error detection, but this is related to the fact that pilots only have access to a small part of the larger traffic picture and therefore only have a limited possibility for discovering errors. Another interesting result of the analysis is that warning systems do not play any significant role in relation to the error detection. A somewhat similar picture emerges when examining the actors associated with the error correction (the reason why the graph shows that no corrections at all were initiated by pilots is that the numbers are based on agreed classifications and each of the classifiers only identified, but did not agree on, a couple of errors which where corrected by pilots). This is a reflection of the fact that the one who detects the error or problem is usually also the one who is the active part in the correction (87% of the time the error detector was also the error corrector).

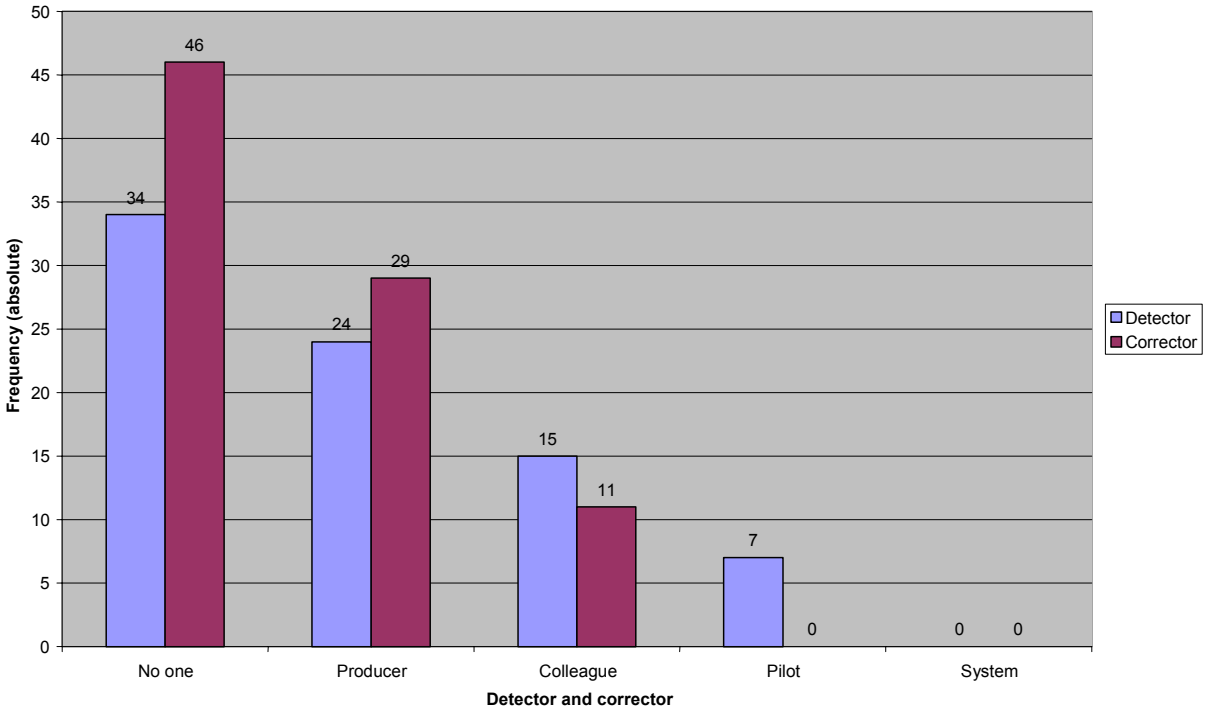


Figure 20: Distribution of error detector and corrector

As mentioned previously it is desirable that errors are caught at an early point in time. However, as can be seen in the chart below, a large majority of the errors is not caught until the outcome stage, $X^2(1, N=46)=32.14$, $P<0.001$. Seemingly, this is in conflict with the often-cited statement that most errors are caught before having any consequence. This high number of errors that are not caught until at the outcome stage is to a large extent the product of the data material used in this study, namely incident reports. These reports are only written in the case where the system safety has been jeopardised and normally by domain experts who have en tendency to only consider something as an error if it has consequences. Consequently, there is a strong bias towards consequential errors.

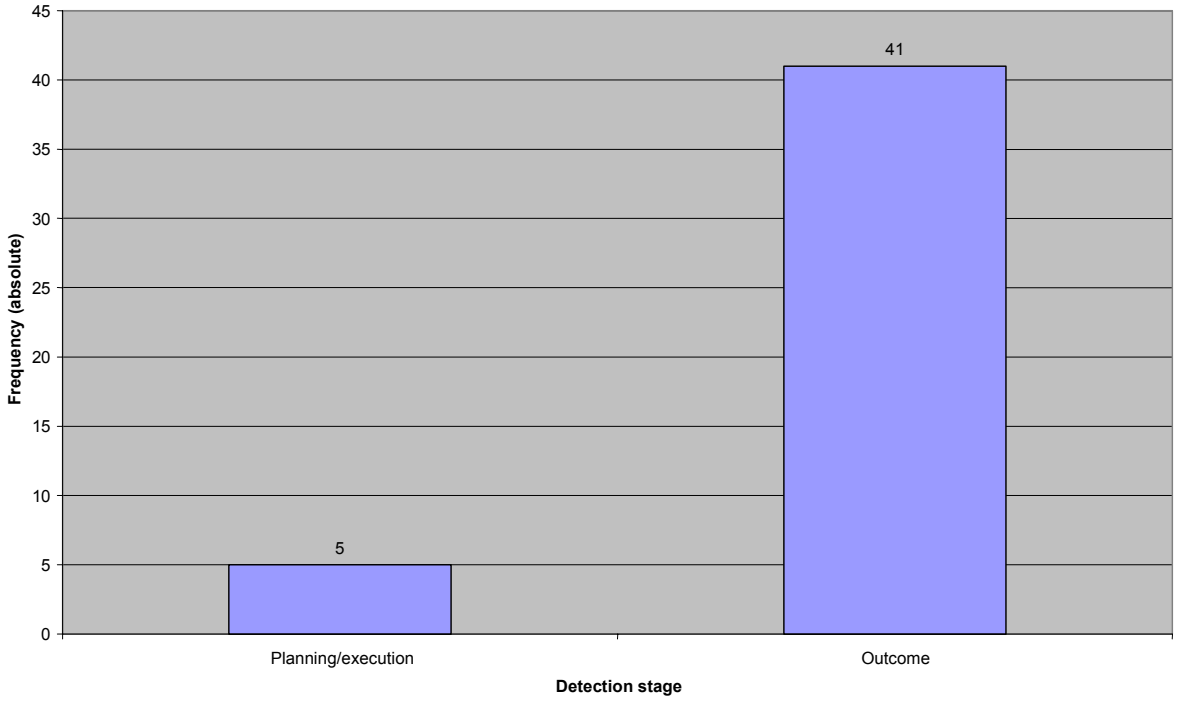


Figure 21: Distribution of detection stages

Below can be seen whether the error detection was cued internally or externally. In this context external feedback (i.e. system feedback and external communication) is clearly the most significant source in the error detection process, $X^2(1, N=58)=46.62$, $P<0.001$. This is not surprising because internal feedback is mainly (but not exclusively) relevant in relation to memory failures and response execution failures - and since these two groups of errors were the least frequent in this study it follows that also the number of errors detected by the use of the internal feedback mechanism should be low.

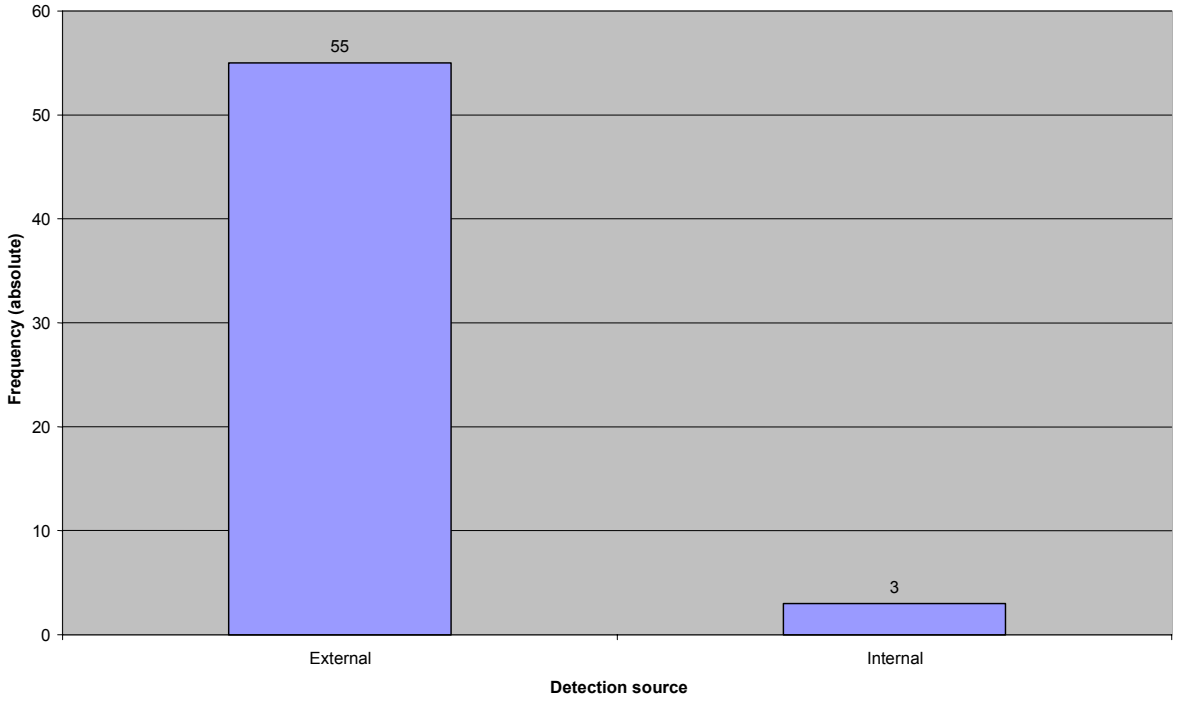


Figure 22: Distribution of detection sources

The final two tables are related to what occurred on the external level, namely the error response and outcome. In relation to the first group the most frequent response was to trap or mitigate the error. So, even though many errors are not detected until the outcome stage - that is, when some consequences have ensued - many errors are nonetheless averted from developing into an even more serious situation (please notice that no cases of exacerbation were observed in the incidents). The differences in the frequencies were significant, $X^2(2, N=94)=50.86$, $P<0.001$.

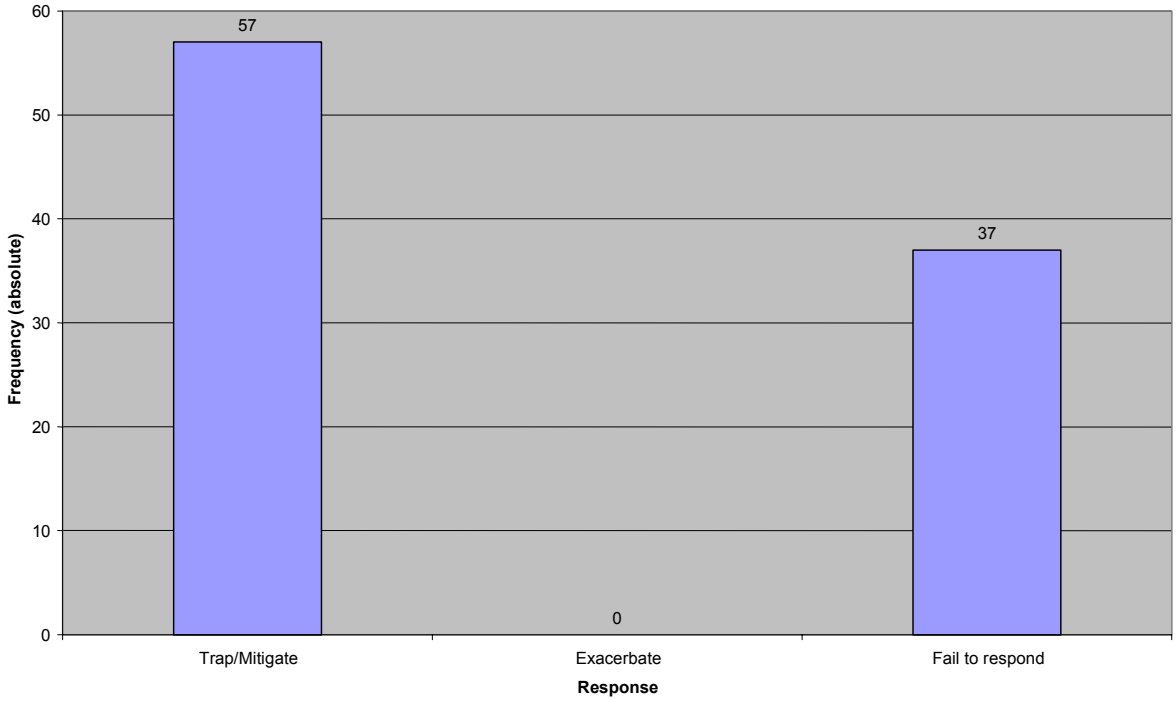


Figure 23: Distribution of error responses

As can be seen in the table below the errors identified in the data material contained a very small part of inconsequential errors. Almost 90% of the errors led to either an undesired stage or an additional error. Again, this can be seen as a product of the particular type of data source used in this study. The differences in outcome were significant, $X^2(2, N=70)=34.40$, $P<0.001$.

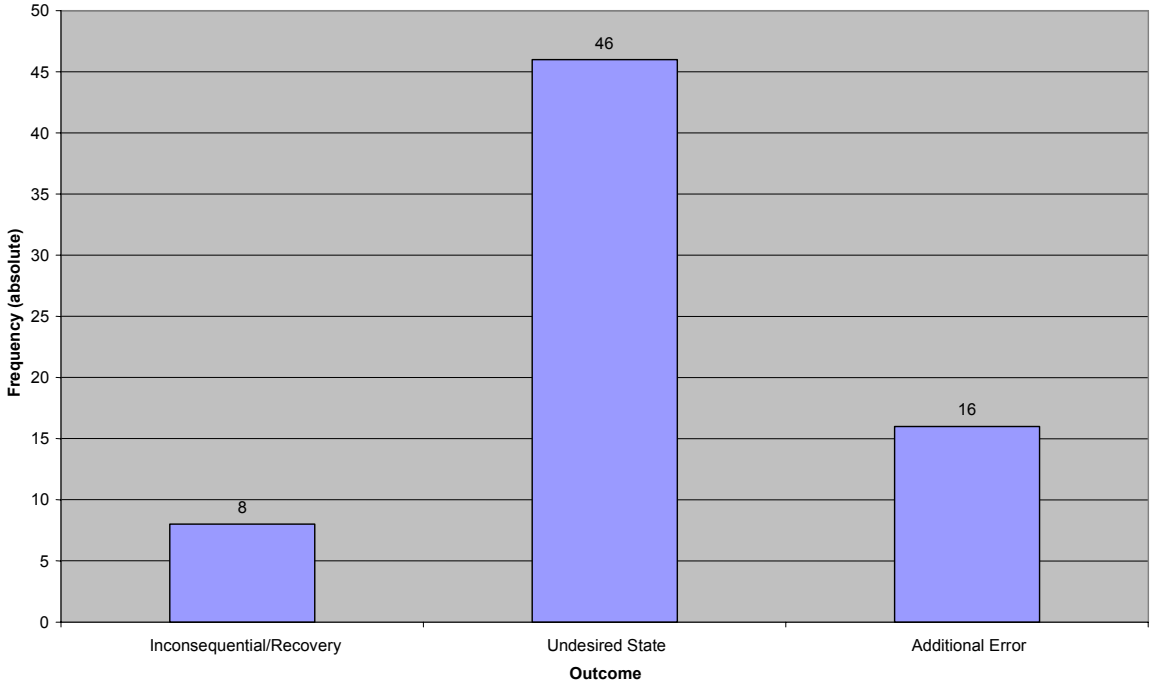


Figure 24: Distribution of error outcomes

8.4 Conclusion

In this chapter part of the conceptual framework has been applied to the analysis of ATM incident reports. The goal of this study was, first, to get an initial indication of the robustness of the proposed error capture taxonomy. The results, based on the kappa analysis, indicated that all of the dimensions in the framework produced reliability ratings that, on the standard interpretation, lie between fair/moderate and strong/substantial. These results are promising because this is the first attempt to formally test the framework on authentic ATC error events.

A second goal of the study was to uncover trends in the identified error events. Several interesting conclusions were made in this context such as the fact that most of the errors discovered are detected by the error producer and that most of the error detection is based on externally derived cues. It should be noted that in the current context only patterns related to the individual dimensions were analysed. Additional interesting trend information may be uncovered by examining interactions between individual dimensions (for example, is there any relationship between the type of errors and the cues used for detecting the errors). These issues will be explored in more detail in the following

empirical studies.

The current study used incident reports as a data material for studying error and error management. In spite of the many advantages of incident reports there are also some inherent limitations in using incident reports as data material:

- The errors found in incident reports are normally not described in psychological terms or in sufficient detail to derive the psychological mechanisms. This disrupts the chances of obtaining highly reliable classifications. The problem is even more prevalent when analysing the error capture processes since the error resolvement is often described in vague terms (and consequently it was necessary to collapse some of the process categories and no classifications of the problem-solving could be made). This may be related to the fact that historically human factors safety initiatives have mainly focused on avoiding the occurrence of errors and not controlling the consequences of errors.

- Since the incident reports are to a large extent based on self-reports, the reports may be associated with inherent biases and, consequently, statistical conclusions should be treated with caution. For example, practitioners have a tendency to only accept something as an error if the error has direct consequences for the course of events. Human factors specialists, on the other hand, consider errors as equally important irrespective of the outcome of the errors (Amalberti & Wioland, 1997). Since ATM professionals write the incident reports it may be expected that inconsequential errors will be underrepresented in the data material.

A potential solution to overcome these two limitations is to apply the analysis framework to other data sources such as interviews based on the critical incident technique (Flanagan, 1954) and real-time studies. Interviews are of interest because they will allow obtaining detailed information about critical incidents and thereby provide a more solid foundation for making inferences about the underlying error and recovery mechanisms. Similar advantages can be obtained with real-time studies. In addition, these may contain a large (and unbiased) variety of categories of the framework and thereby be better suited for testing aspects of framework that were less prevalent in the incident reports.

9 Study 2 – Real-time study

Several researchers have suggested that there has been too little emphasis on normal operations when analysing error and error management (Helmreich et al., 2001; Maurino, 1999). This is unfortunate because studies of normal operations might be a significant and vital source of information to obtain knowledge about safety critical issues involving human performance. As a part of the Ph.D.-project it was therefore decided to do a small-scale study of how errors are captured in everyday operational situations. There are in the current context in particular two arguments for focusing on normal operations in the evaluation of the error management framework. First, it was a general requirement to the framework that it should be applicable to both normal and abnormal situations. Second, it can be expected that other types of error management might be more prevalent in this kind of material compared with, for example, incident reports. In particular, most errors from everyday situations will have a tendency to be captured much sooner and perhaps also by different mechanisms.

Real-time studies can be accomplished by either using a real operational situation (e.g. checkout of controllers) or using a simulated setting. There are advantages and disadvantages with each of these possible approaches. The advantage of using real operational situations is that it becomes possible to the study the phenomena in a situation with a high degree of ecological validity and contextual richness. The disadvantage is, however, that such a study would need to be carried out over an extended amount of time to achieve the required corpus of events. Furthermore, practical constraints related to obtaining permission to recording and analysing errors from real operational situations can be a problem due to the sensitive nature of this subject.

The study presented in this chapter was carried out on the basis of a simulator study of En Route Air Traffic Control with video recordings. The simulator study was done as a part of the ATCO trainee's education referred to as the Radar Module. This module constitutes the last part of their basic training and at this point in time they have acquired a basic level of skills necessary for carrying out the controller task (afterwards they will start on on-the-job-training). The recordings from the simulator trials were carried out as a part of another Ph.D. independently of the current Ph.D.[13] After having reviewed a series of tapes from these scenarios it was decided that they could constitute an interesting foundation for the current project insofar as they contained a number of minor errors and it was therefore decided to study these events in detail by the use of the error management framework.

[13] Hauland, G. (2002): Measuring Team Situation Awareness in Training of En Route Air Traffic Control – Process Oriented Measures for Experimental Studies. Ph.D. Thesis. Risø National Laboratory, Roskilde, Denmark.

9.1 The analysis framework

An overview of the framework used in this study is shown in the table below. Some minor changes have been made in relation to study 1. First of all, different kinds of error producers are included: *Radar* and *Planner* are the radar and planner controller involved in the experiment (for a description of their respective tasks see section 1.3); *Pilot* is the "ghost"-pilot involved in the experiment and *Other* is an ATCO that is not directly a part of the experiment. Second, *Instructor* has been added to the detector and corrector dimension insofar as an instructor was sitting behind the radar and planner controller in the simulator scenarios.

DESCRIPTION OF ERROR AND RECOVERY # 1							
"When SAS asks for clearance to flight level 310, R1 could have discovered the risk for a conflict insofar as the strips for SAS and SCW were available. The conflict was detected by a relieving ATCO when the separation standards between the two aircraft were violated."							
ERROR							
Error	Producer	Radar	Planner	Pilot	Other		
	Cognitive Domain	Perception	Short-term memory	Long-term memory	Decision	Response	
DETECTION & RECOVERY							
Who:	Error/state detector	No one	Producer	ATCO	Instructor	Pilot	System
	Error/state corrector	No one	Producer	ATCO	Instructor	Pilot	System
When:	*If detector is "No one" or "System" then go to "What":*						
	Time of detection	Planning		Execution		Outcome	
How:	Detection cue(s)	External communication		System feedback		Internal feedback	
RESPONSE & OUTCOME							
What:	Error/State Response	Trap/Mitigate		Exacerbate		Fail to respond	
	Error/State Outcomes	Inconsequential/Recovery		Undesired State		Additional Error	

Table 6: The analysis framework (study 2)

9.2 Method

9.2.1 The data material

The simulator recordings were obtained from an experiment that was conducted using a real-time ATC simulation facility at the Danish CAA in Kastrup. In total, 56 unique team combinations of air traffic controllers consisting each of one Radar and one Planner controller participated in training exercises in an En Route Centre simulator. In addition, "ghost pilots" participated in the scenarios when communications were made between air traffic control and the aircraft.

The trainees were required to carry out some different scenarios that could be expected to elicit some of the main error categories in the taxonomy. Each scenario lasted between 30 and 45 minutes and consisted of three probes. These are inserted events that are typical of normal everyday tasks of a controller and involved coordination requirements:

- **Probe 1:** This event occurred right after the hand-over where an unknown fighter calls Radar controller. The fighter pilot requests to change from VFR to IFR pick-up to do ILS training at Brande (IFR is necessary to be able to make an ILS).
- **Probe 2:** This event occurred about 10 minutes after the hand-over. A fighter requests to change flight level (FL) from sector B to A as a part of a test flight. The pilot would like to do an IFR test flight at FL 280 which involves following a route with 3 fixpoints and then after the test flight to return to VFR.
- **Probe 3:** This event occurred about 20 minutes after the hand-over (in the case of abnormal scenarios just after abnormal) where an aircraft requests a diversion in sector B. The reason for the diversion is either (a) company request or (b) change of flight plan.

In addition, some of the scenarios included one of two types of abnormal events:

- **Emergency descent:** Emergency descent starts at a high level in sector B. Scenario example: shortly after hand over from A to B emergency descent begins to FL 100 (with max. Rate Of Descent 4.000-5.000 ft/min). The aircraft turns 30 degrees to the left. SSR is shortly afterwards set to 7700. No reply is made at the first call from the ATCO, but a little later it is reported that: "(callsign) -- executed emergency descent due to loss of cabin pressure". At FL 100 it is reported that: "(callsign) -- request clearance to (EKBR/EKDA) at this level. Request priority landing due to smoke in the passenger cabin. 1 passenger has been injured. Request medical assistance at arrival".
- **Fuel dump:** Fueldumping due to hydraulic failure. Starts at a high-level in sector B. Scenario example: shortly after hand-over of frequency from A to B or shortly after hand-over of frequency from B to A the pilot calls the ATCO and says: "Due to hydraulic failure request clearance to an area for fuel dumping and then clearance to EKBR". The pilot will be able to make a normal landing and the radio transmission is normal during fuel dump. The duration is 6 min. and the amount is 30 tons. At landing the fuel will be 100 tons.

Each of the scenarios began with the instructor giving a hand-over and finished about 30 minutes afterwards. The human activities were videorecorded to provide continuous record of events (incl. head-mounted camera on radar and planner controller, overview of the scene, and radio communication) and strips from the simulator scenarios were obtained.

A total number of about 60 scenarios were initially available. The recordings included helmet-mounted video recordings from both the radar and planner controller and all audio communication (both controller-controller and controller-pilot). Among these 10 scenarios were selected to be used for further analysis. The scenarios were selected on the basis of the fact that for this subset of scenarios an instructor was present which was considered important because they would often comment on things that were not clearly visible on the video recordings. It should be noted that the scenarios were not chosen on

the basis of any theoretical considerations, but it was expected that the scenarios would reflect a broad range of errors and recoveries from the taxonomy.

9.2.2 Procedure

In the following we will review the procedure associated with the information elicitation, the event segmentation and the classification. The information elicitation and segmentation phase are partially overlapping in this study and are therefore described under one heading.

Information elicitation and segmentation

The scenarios were segmented into a number of errors, each of which was relatively self-contained descriptions. For these errors both the information related to the communication between actors and, when relevant, contextual information was added to give an understanding of what was going on. It should be emphasised that there were some limitations in relation to this initial identification of errors:

- We do not have access to mental activities underlying the behaviour of the controllers and it therefore puts some limits on the precision with which we can analyse the cognitive foundation of the errors and their recovery. Nonetheless, since there was both Radar and Planner controller (and an instructor) present in the position in all the scenarios it meant that there was a larger chance of externalising their plans and considerations.

- Even though the helmet camera from the radar controller could give a fairly accurate picture of the traffic situation it was not possible to read the labels of the individual aircraft due to the resolution of the camera. Nonetheless, it was possible to get rough picture of the location and callsign of the main aircraft insofar as the eye-tracking made it possible to see which aircraft the controller was looking at when talking about a specific aircraft. All the original strips from the individual scenarios were also available which made it easier to get a picture of the traffic.

In many cases the video recording did not give sufficient information for an observer to determine the complete context of the error or the underlying mental activities. In particular, it was necessary to have a high level of domain knowledge to be able to determine deviations from an optimal performance. In order to get a more complete understanding of the scenarios – and in particular the error events - it was decided that it would be useful to have some "ATC-experts" present - i.e. instructors - to review the video tapes and comment on central episodes within each of the scenarios. The instructors were associated with the same facility as the air traffic controllers used in the experimental sessions. Hereby it was possible to avoid that variations in local procedures, practices and knowledge would influence the error identification. Also, the instructors were, of course, highly familiar with what the trainees were taught and could therefore easily identify errors committed.

Hence, the instructors were given the opportunity to comment on the trainee's performance and the thoughts lying behind the behaviour. All of their comments were recorded on one sound channel on a separate videotape containing the original pictures from the simulator trials so it was hereby possible to preserve both their comments and the associated simulator context. For each of the selected scenarios the instructors were before reviewing the videotape given a paper with a short transcript of the identified errors and a copy of all the strips that had been used in the scenario. The main focus was on the error events and the discovery and recovery of errors or, in general, the handling of any problems encountered. During the review of the simulator scenarios some new errors were discovered and some of the previously discovered were abandoned as being errors because the instructors did not consider the events erroneous. In this manner the identification of errors was calibrated and at the same time additional information concerning the pre-identified errors was obtained. On the basis of this process a total of 250 errors were identified.

An example of the product of the above-described procedure is shown below. In the example the Radar controller (R) creates a conflict between two aircraft that is immediately discovered by the instructor sitting behind the ATCO.

Communication
R: *Birdsong 17, left turn inbound ODIN* Pilot: Birdsong 17, left turn inbound ODIN. Confirm R: Birdsong 17, left turn inbound ODIN Instructor: Won't it lead to two aircraft hitting each other? R: Yes. Birdsong 17, turn left heading 270 immediately. SAS 633, Turn right heading 360 at once (traffic information is then given to SAS 633)
Contextual Information
The pilot must now have initiated the turn since it is necessary to make an avoidance response and provide traffic information.
Instructor's Comments
She (i.e. the Radar controller) creates a conflict. It is birdsong that was flying towards Brande that suddenly wants to do some airwork at ODIN. She probably wants Birdsong to go south of the holding, but as a consequence of her instruction it heads directly towards an aircraft in the holding (i.e. SAS 633) which she has probably not seen.

As can be seen in the example three types of information were elicited. The first is "Communication" and is a verbatim transcription of the relevant communication associated with the error (the "Instructor" in the text is the instructor who was sitting behind the ATCOs in the scenario). The second is "Contextual Information" and is elaboration of what is going on in the situation. The final field is "Instructor's Comments" and is the comments from the instructor who reviewed the videotape after the experimental trials (and is therefore not the same instructor as the one who actually was present in the scenario).

Classification

The classification part of the study was divided into three sub-phases: (a) *Initial classification* – First, all the error events were classified twice by the author with one month's interval and on this basis a consensus classification was produced; (b) *Calibration phase* - The second classifier was trained in the use of the taxonomy. In order to achieve reliable and valid categorisations the classifier was given feedback concerning the "correct" classification as a part of the training (based on the author's classifications) and potential misunderstandings in the use of the taxonomy were calibrated. The first four scenarios containing 98 error events were used for this purpose; (c) *Test phase* - On the basis of the transcripts from both the simulator episodes containing errors and instructors' comments to these errors the trained observer was asked to classify the remaining errors and error recoveries observed in the simulator trials.

In some cases it was difficult to choose between two seemingly equally good candidates when classifying the error events. In order to maintain as much information as possible, and at the same time avoid statistical problems associated with analysing dual classifications, it was decided to use a rank-based dual-classification principle. The idea is here that it is allowed to use dual classifications in those situations where more than one category may apply. However, to be able to single out one particular category as the chosen classification they should be ranked according to their estimated applicability[14].

For the study it was considered useful to develop guidelines to enhance consistency in the classifications of error events. The guidelines based on the experience from the analysis of the simulator scenarios are briefly described in the following.

Cognitive domain

In some situations the choice between some of the cognitive domains may be associated with difficulties. The most prevalent ones are elaborated below.

1. *Long-Term Memory vs. Decision-Making errors*. Long-Term Memory (LTM) errors are closely related to Decision-Making (DM) errors. An important distinction is, however, that LTM errors occur when learned information is not triggered in memory. So, if we can reasonable assume that the ATCO had learned what to do – e.g. a procedure or some standard phraseology - but did not do it then it is a LTM error. DM errors are, on the other hand, less standardised and are more a question of bad judgement, reasoning or prioritisation.
2. *Long-Term Memory vs. Short-Term Memory errors*. If an ATCO forgets to carry out some learned procedure this is a LTM error. If he/she forgets information obtained during shift then it is a Short-Term Memory error.

[14] Please notice that this procedure is slightly different from the one used in Study 1 where all dual classifications were omitted altogether from the analysis. The procedure was modified in this study to maintain as much information as possible.

The Who-Question

In the choice of detector and corrector there are some rules-of-thumb that are useful to make a consistent selection:

1. *Error suspicion.* If an ATCO knows that something is wrong but not the exact nature of the problem he/she is the detector. So, you might only have a suspicion that you misunderstood or misheard something.
2. *Overlap between detection and correction.* Sometimes the detection is also the correction. If, for example, an ATCO points out to his colleague that something has not yet been done, he thereby also tells what should be done. So, in this case he is both the detector and corrector.
3. *An active 'no-choice'.* If a deliberate choice to do nothing is made - while it is still possible - it is still considered an active choice. The person deciding to do nothing is here the corrector. If an error is detected selection of "No one" in the corrector field will only occur if the correction is forgotten (so it is unintentionally not carried out) or information not adequately transferred from person A to person B (perhaps due to some team dynamics).

The How-Question – Detection

1. *Slips.* In many cases when making slips during communication people correct their own errors immediately without any help from the environment. In such cases the detection can only occur through internal feedback. In other cases when interacting with equipment people discover and correct their own errors immediately because they cannot carry out the action (often referred to as a "Forcing Function"). In such cases the detection occurs through system feedback.
2. *Internal feedback vs. external communication*. Detection through external communication means that the error is detected directly on the basis of on-going communication. Detection through internal feedback is related to previous communication or when a communication-not-made is detected (in neither of the cases cues for the error detection are available in the environment).
3. *System feedback vs. external communication*. When an ATCO discovers an error while discussing plans with another ATCO the detection will be on the basis of external communication (even though they also use the system feedback as a source).

The What-Question – Outcome

The concept of "Undesired state" is in this study expanded to not only include incidents and separation violations. Basically, undesired states are all episodes that could have potential negative safety implications. This was done because no incidents as such occurred in the simulator trials. Some examples of undesired states related to coordination with other sectors and communication with aircraft are provided below:

Coordination with other sectors:

a) Omission of a required coordination before aircraft is handed over (e.g. that aircraft flying at "wrong flight level" or flying via another route than the planned) or wrong information is provided during a coordination (so the other ATCO, for example, expects that the aircraft flies another route than it is actually doing).
b) No hand-over is made before the aircraft leaves the sector.
c) An aircraft diverting to a new airport is not coordinated with the new airport (so they do not know about the aircraft before it arrives at their airport).

Communication with aircraft:

a) No instructions concerning transfer of frequency to another sector is given to the pilot (so the pilot enters another sector where the ATCO cannot get in contact with the aircraft).
b) A wrong frequency or flight level is given to a pilot and implemented.
c) Aircraft are given instructions that bring them in direct conflict with each other.

9.3 Results

The main goal of this study is to test the validity and reliability of classifications based on the error taxonomy.

9.3.1 Reliability

In the following is reported the results from the intra- and inter-rater analysis.

Intra-rater reliability

The same cases were analysed twice by the same observer (i.e. the author) with a one-month interval. The obtained Kappa coefficients and P-values are shown in the table below.

Error		
Error producer		
Kappa	0.90	P-value<0.001
Cognitive domain		
Kappa	0.86	P-value<0.001
Detection and recovery		
Who – detection		
Kappa	0.94	P-value<0.001
Who – correction		
Kappa	0.88	P-value<0.001
When – detection stage		
Kappa	0.89	P-value<0.001
How – detection source		
Kappa	0.84	P-value<0.001
Response and outcome		
What – response		
Kappa	0.94	P-value<0.001
What – outcome		
Kappa	0.74	P-value<0.001

Table 7: Intra-rater kappa coefficients and P-values for each of the main dimensions in the framework (study 2)

As can be seen a very high level of agreement was achieved across all dimensions (both measured on the basis of the Kappa-values and the P-values). A particularly interesting result is that the *When*- and the *How*-question received a very high level of agreement in this study compared with the study focusing on Swedish incident reports (Study 1) and it was not required to collapse any categories. So, even though no interviews with the participants in the experiment were made and no think-aloud protocols were available it was still possible to obtain very robust classifications for these two dimensions.

Inter-rater reliability

Before being able to make a Kappa analysis of the inter-rater reliability it was necessary to give the second rater some training in applying the classification. As described previously, the first 4 out of 10 scenarios in the data material were used for this purpose and disagreements related to lack of familiarity with the taxonomy were resolved.

Below are shown the results of the inter-rater reliability analysis of the remaining 6 scenarios containing a total of 152 error events.

Error		
Error producer		
Kappa	0.95	P-value<0.001
Cognitive domain		
Kappa	0.69	P-value<0.001
Detection and recovery		
Who – detection		
Kappa	0.81	P-value<0.001
Who – correction		
Kappa	0.69	P-value<0.001
When – detection stage		
Kappa	0.60	P-value<0.001
How – detection source		
Kappa	0.68	P-value<0.001
Response and outcome		
What – response		
Kappa	0.80	P-value<0.001
What – outcome		
Kappa	0.50	P-value<0.001

Table 8: Inter-rater kappa coefficients and P-values for each of the main dimensions in the framework (study 2)

The results indicate a high level of agreement across all dimensions applied in this study. The Kappa-values are a bit lower than for the intra-rater reliability analysis. It is not surprising that, in general, a lower level of reliability is obtained for the inter-rater measurements compared with the intra-rater measurements. There are at least two reasons for this:

- Basically it is easier to agree with yourself than others!
- Due to practical limitations one of the observers was given a fairly short training in using the framework and, if had been given more training, would probably have been able to apply the framework even more consistently.

Both for the intra-rater and the inter-rater analysis a somehow lower level of agreement was achieved for the outcome-dimension. A reason for this is that a more broad definition of "undesired state" was applied in this study as a consequence of the fact that no incidents occurred in the observed scenarios. Even though a list of events that constituted undesired states was developed before the classifications there was still some room for interpretation concerning what constituted an undesired state. The problem here is related to getting a robust operationalisation of what is meant by "consequences". Whether specific types of situations should be considered consequential or not would probably be best determined by having a group of domain experts working together on analysing a battery of concrete situations and on this basis come up with a list of "undesired states".

9.3.2 Pattern analysis

A total of 250 errors were identified in the 10 scenarios. As previously described the data-set was analysed twice by the same classifier (to get a measure of intra-rater reliability) and once by a second classifier (to get a measure of inter-rater reliability). In order to get a comprehensive database it was decided to combine these analyses. First, on the basis of the two datasets used for the intra-rater reliability study were compared and all disagreements were resolved by either deciding on one appropriate classification or for the cases where the category could not be resolved the event was classified as "Unknown". Second, the consensus dataset from the intra-rater analysis was compared with the dataset from the second classifier and on this basis a final consensus dataset was generated.

In the following is reported the results from the pattern analysis. First, we will review the *main effects* of the analysis of the individual dimensions. For the statistical analysis of the main effects an exact Goodness-of-Fit test based on a uniform distribution will be used. Secondly, we will examine *interaction effects* between the dimensions within the framework. For the test of independence is used Pearson's Exact Test. In those instances where interaction is found we will explore which specific cells that contribute to this effect.

Main distributions

In the figure below is shown the distribution of the error producer. As can be seen a large majority of the errors observed were made by the Radar Controller (57.6%) and a smaller amount was committed by the Planner Controller (31.2%). The difference in error producer is significant, $X^2(3, N=250)=189.26$, $P<0.001$.

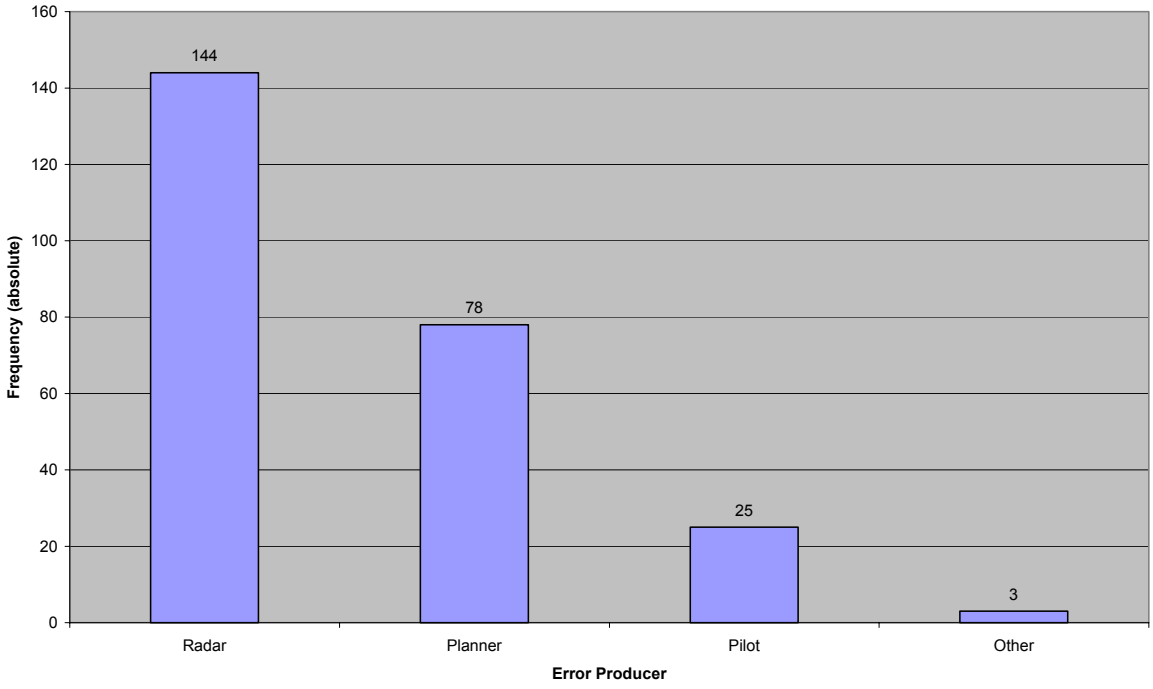


Figure 25: Distribution of errors for each error producer.

In the figure below is shown the distribution of cognitive domains. The differences in cognitive domain are significant, $X^2(4, N=233)=11.14$, $P=0.025$. One thing that is particularly noteworthy is the amount of Long-Term Memory (LTM) errors (27,5%). This high amount of LTM-errors is directly related to the fact that the study was focused on training scenarios and it can therefore be expected that there will be a significant amount of novice-errors.

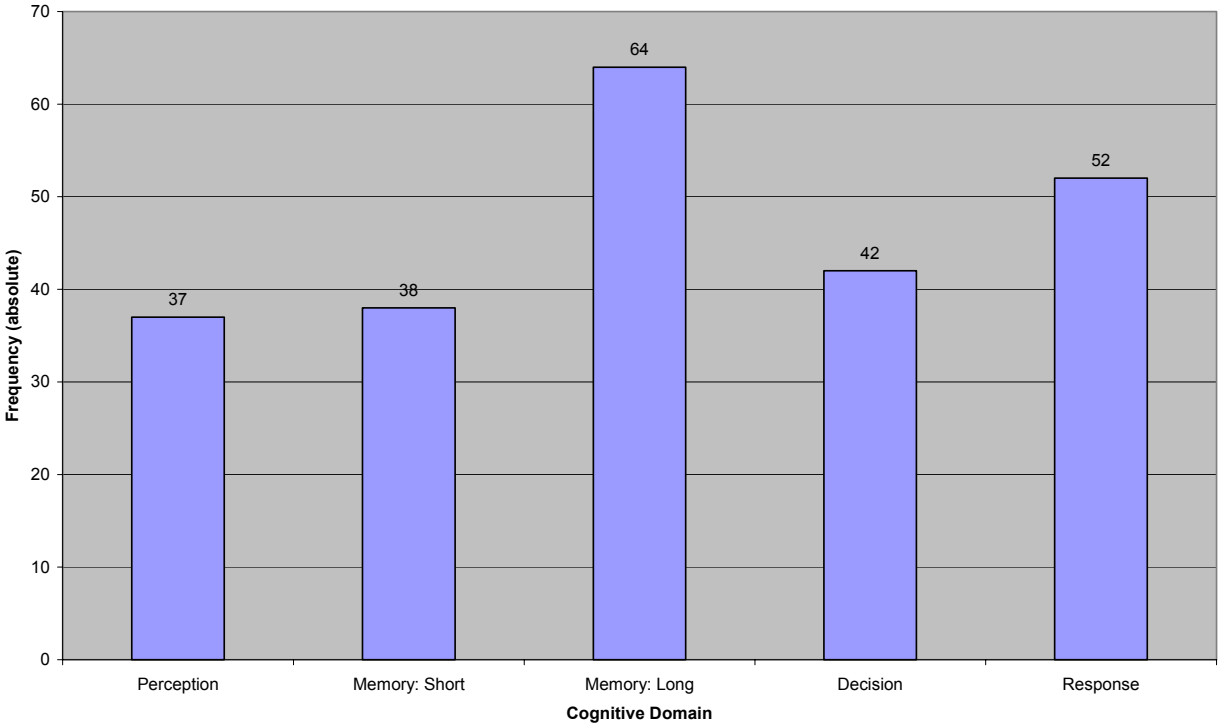


Figure 26: Distribution of cognitive domains.

In the figure below is shown the distribution of actors involved in the detection and correction. The differences in frequencies are statistically significant (Detector: $X^2(5, N=243)=125.05$, $P<0.001$; Corrector: $X^2(5, N=248)=146.40$, $P<0.001$). As can be seen, the detector and corrector is most often the producer of the error – a result that is in good concordance with the previous study. However, ATCOs different from the error producer and the instructor were also involved in a significant part of the error detection and recovery. In a large majority of the cases the error detector was also the error corrector (72%). The results also correspond well with the previously described notion about errors having a larger chance of being discovered by other people if these share a large part of the context (i.e. another ATCO or instructor). It might seem surprising that the instructor does not detect and correct a larger part of the errors than is the case. There are two explanations for this: (1) the instructor might consider some errors more important to correct than others; (2) the instructors were encouraged to limit their intervention in the experimental scenarios.

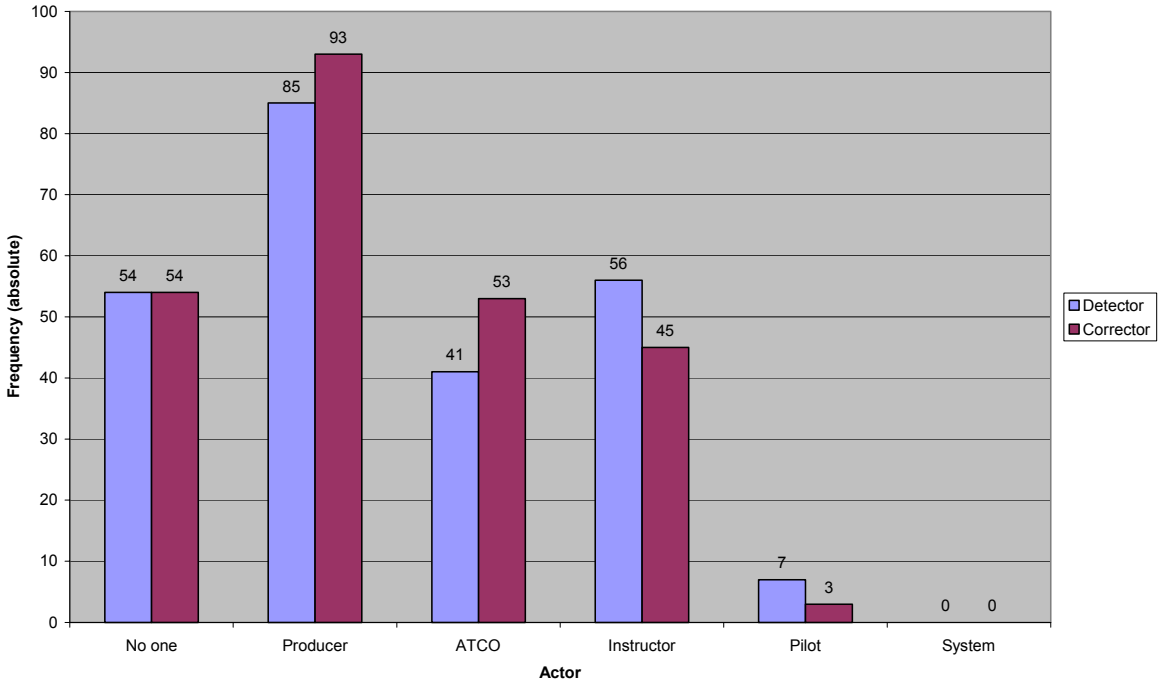


Figure 27: Distribution of error detector and corrector

In the figure below is the cognitive stage at which the errors were detected. The differences in frequencies are highly significant, $X^2(2, N=192)=17.09$, $P<0.001$. As can be seen most errors were detected before the outcome stage – namely at the planning stage (26.6%) or execution stage (47.4%). This is in concordance with the notion that many errors are detected at an early stage of the error evolution. It is also interesting to note that in comparison with the incident study the distribution of detection stages is much more varied. This indicates that the distinction between different kinds of detection stages might have a larger analytical power when applied to a normal everyday setting compared with incidents.

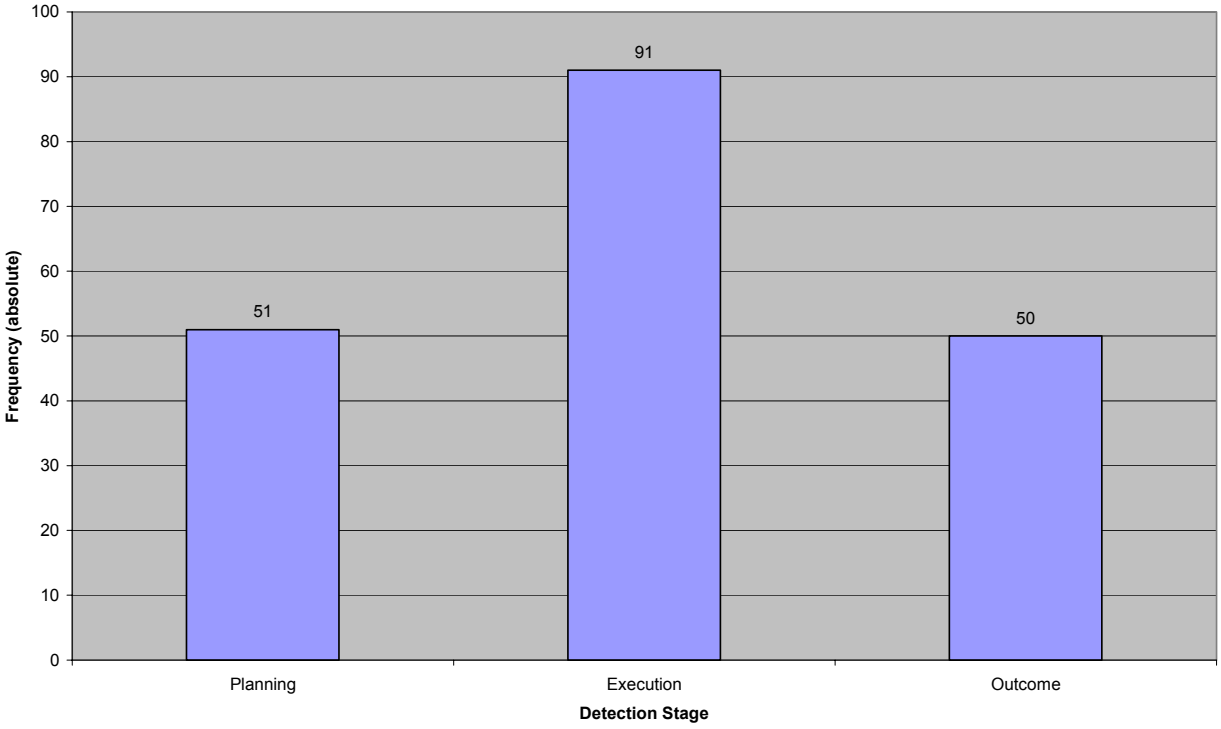


Figure 28: Distribution of detection stages.

In similar vein, the results from the analysis of the detection source reveal a more varied pattern than the study of incident reports. Actually, system feedback is the least prevalent source of feedback (20.6%) whereas internal feedback (46.1%) and external communication (33.3%) account for a large majority of the cases. The differences in frequencies are highly significant, $X^2(2, N=180)=17.63$, $P<0.001$.

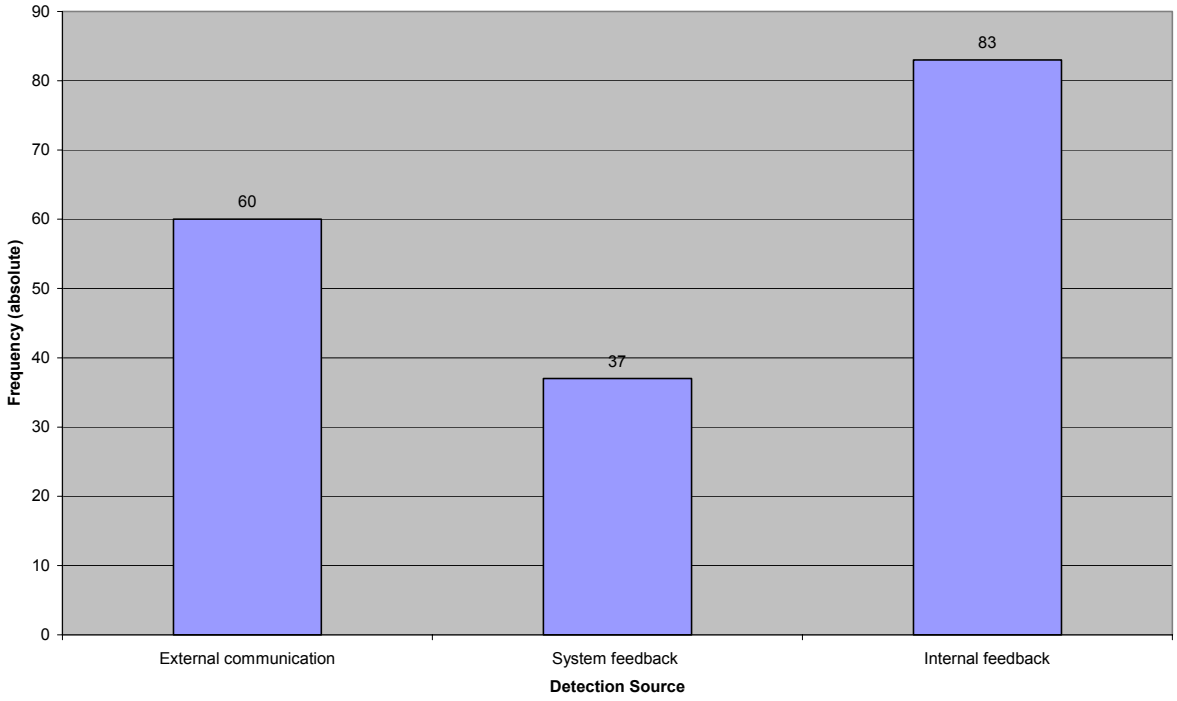


Figure 29: Distribution of detection sources.

As can be seen in the figure below the most frequent response was to trap/mitigate the error or its consequences (71.6%). Conversely, the data revealed that a minority of the error events were not responded to (28.0%). Only in one case was the response an exacerbation of the situation. The differences in the distribution of response were significant, $X^2(2, N=250)=193.30$, $P<0.001$.

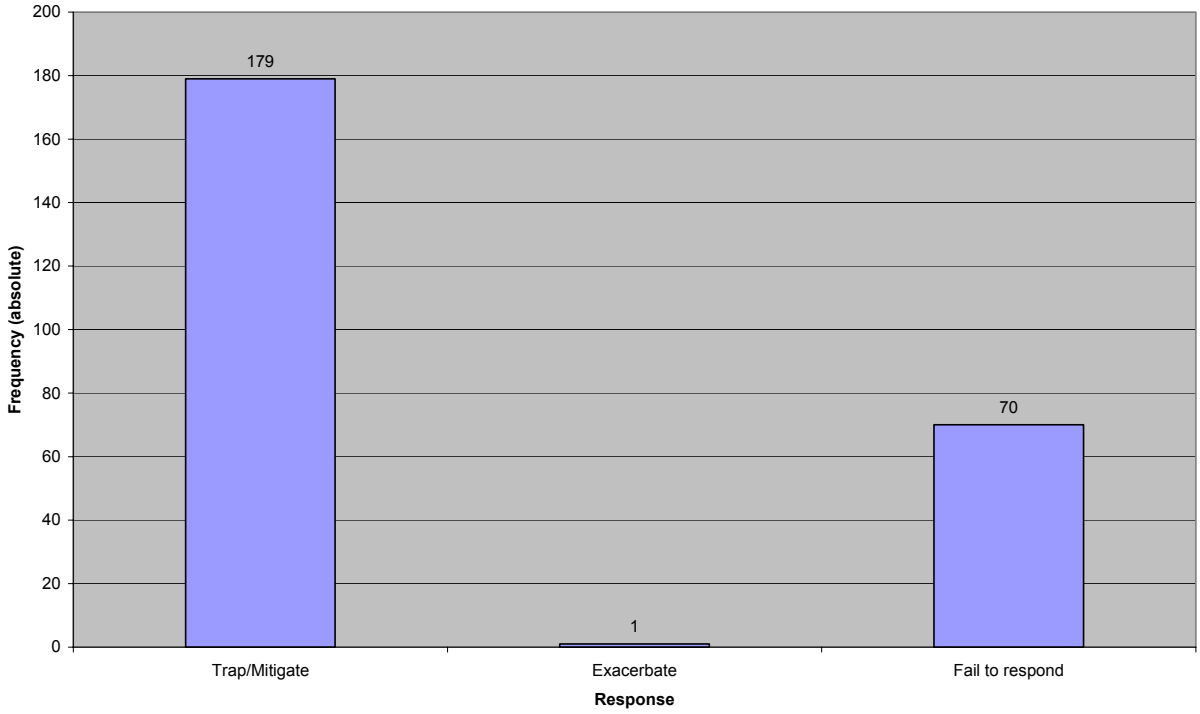


Figure 30: Distribution of response types.

In the final figure is shown the outcome of the individual error and error management events, $X^2(2, N=248)= 323.19$, $P<0.001$. As can be seen a large majority of the errors were inconsequential (87.1%). This is a reverse pattern compared with the incident reports which corresponds with the notion that only a small amount of the errors committed will end up in incident reports – namely the ones that lead to serious negative consequences. It is also interesting to note that the results showing that 87 per cent of the errors were inconsequential is in good concordance with another study by Klinect et al. (1999) based on observations of pilots during normal operations which revealed that about 85% of the observed errors were inconsequential.

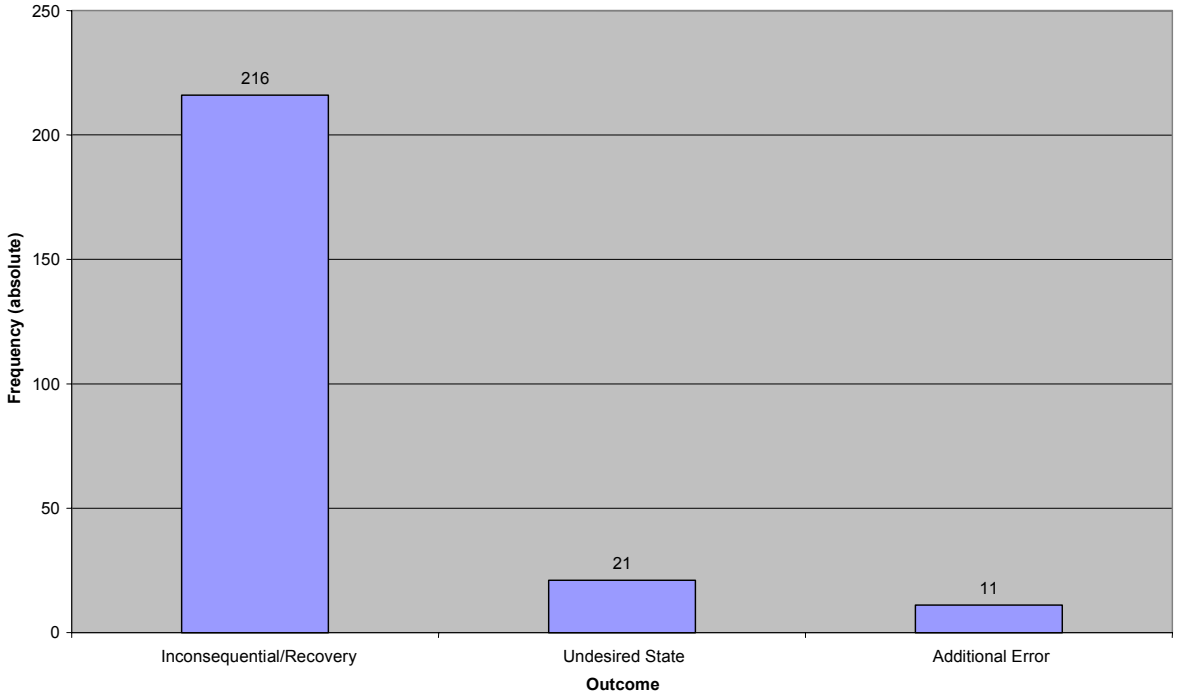


Figure 31: Distribution of outcome types.

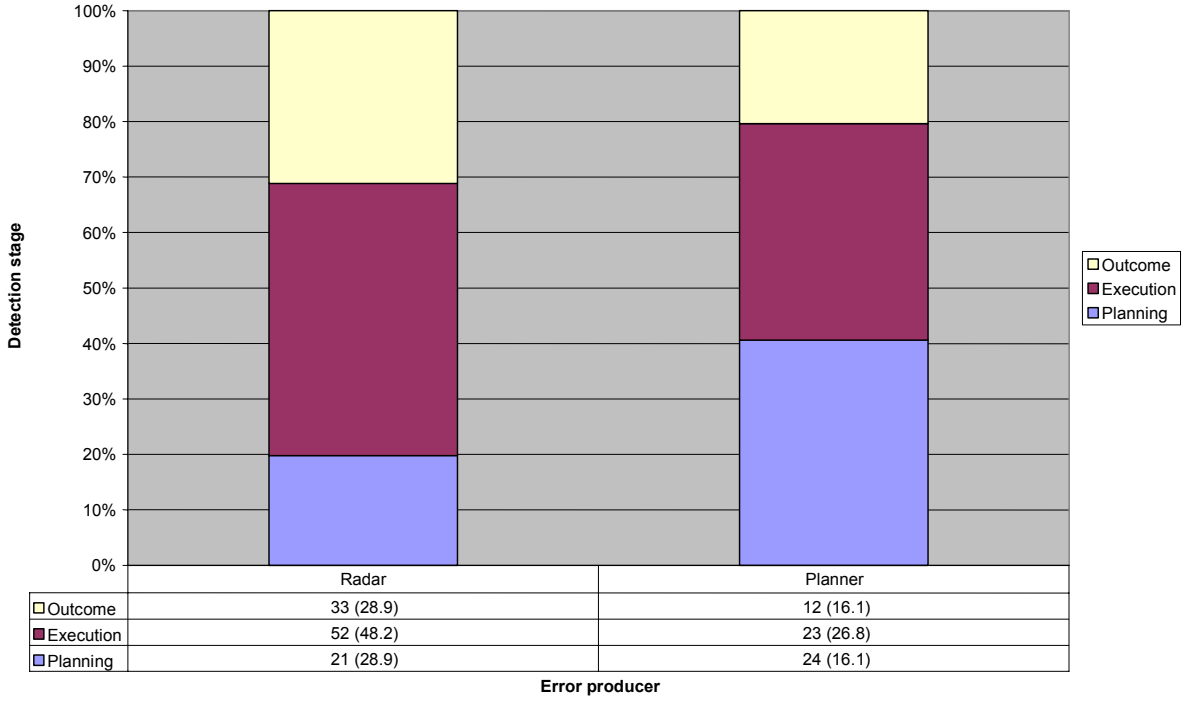
Interactions

In the following we will examine the interaction between different dimensions within the framework[15]. These interactions can be of interest insofar as they are less vulnerable to biases and distortions compared with the data presented above. It should be noted that for the analysis of interaction only data related to errors committed by either the Radar controller or the Planner controller will be included. For the analysis Pearson's Exact Test will be used to determine whether different dimensions are independent or not. In addition, adjusted residuals (AR) will be used to determine which cells contribute most toward the effect[16].

[15] A note should be made concerning how "missing values" (i.e. instances where no categories are determined) are dealt with in this project. Several approaches are available. One solution was to exclude all observations where a category was not determined within one or several of the dimensions. Another solution was to include all observations irrespective of whether some values might be missing within some of the dimensions. In the current context the latter solution was chosen – i.e. to maintain as much data as possible in relation to the analysis of main effects and interaction effects. Consequently, a summation of instances within a single dimension might not be exactly the same for the main and the interaction analyses. In principle, it could have been possible to exclude all instances where one or several values were missing, but this was not decided insofar as this would reduce the data material. Furthermore, it was expected that these missing values constituted random "holes" in the data material and a reduction of the data material was therefore deemed unnecessary.

[16] According to Agresti (1996), the standardised residual (also called the 'Pearson' residual) is calculated by dividing the raw residual (i.e. the difference between the observed and expected counts) by the estimated standard deviation. These residuals follow an approximately normal distribution when the sample size is large. The Pearson residual is divided by its standard error to get the adjusted Pearson residual (a.k.a. the 'standardised Pearson residual') which follows a standard normal distribution (which is similar to the z distribution). Agresti (1996) states: "Adjusted residuals larger than about 2 in absolute value are worthy of attention, though one expects some values of this size by chance alone when the number of categories is large" (p. 91). More precisely, the cut-off point should be at values below –1.96 or values above 1.96 in the case of a two-tailed test at the 0.05-level.

In the chart below is shown the relationship between Error Producer and the Detection Stage (in the table below the chart is shown the observed counts and, in parenthesis, the expected counts). The dependence between the two dimensions is significant, $X^2(2, N=165)=8.52$, $P=0.016$. The main contribution to the dependence is that Planner has a relatively higher amount of errors that are detected at the planning stage (AR=2.88). Conversely, the Radar detects a less than expected amount of errors at the planning stage (AR=-2.88). This can be related to the fact that the Planner normally has a longer time frame to carry out his/hers tasks. That is, the Planner works on a more strategical level whereas the Radar is working on a more tactical level.

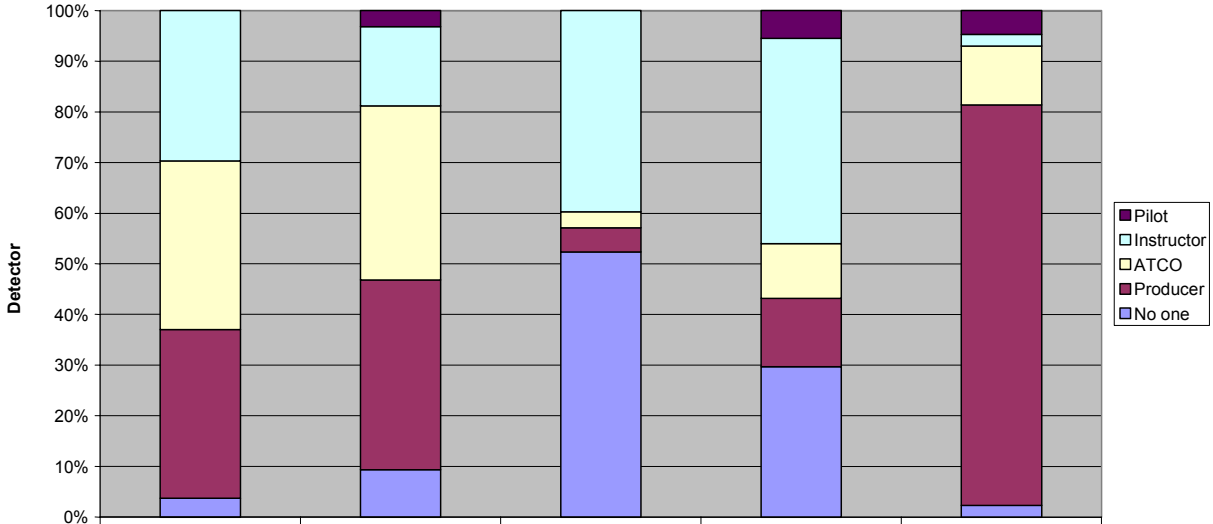


	Radar	Planner
Outcome	33 (28.9)	12 (16.1)
Execution	52 (48.2)	23 (26.8)
Planning	21 (28.9)	24 (16.1)

Error producer

Figure 32: Interaction between error producer and detection stage.

In the chart below is shown the interaction between the Cognitive Domain and the Error Detector. There is a high level of dependence between the two dimensions, $X^2(16, N=202)=129.35$, $P<0.001$.



	Perception	Short-term memory	Long-term memory	Decision-making	Response execution
Pilot	0 (0.7)	1 (0.8)	0 (1.6)	2 (0.9)	2 (1.1)
Instructor	8 (7.2)	5 (8.6)	25 (16.8)	15 (9.9)	1 (11.5)
ATCO	9 (4.1)	11 (4.9)	2 (9.7)	4 (5.7)	5 (6.6)
Producer	9 (8.4)	12 (10.0)	3 (19.7)	5 (11.5)	34 (13.4)
No one	1 (6.6)	3 (7.8)	33 (15.3)	11 (9.0)	1 (10.4)

Cognitive Domain

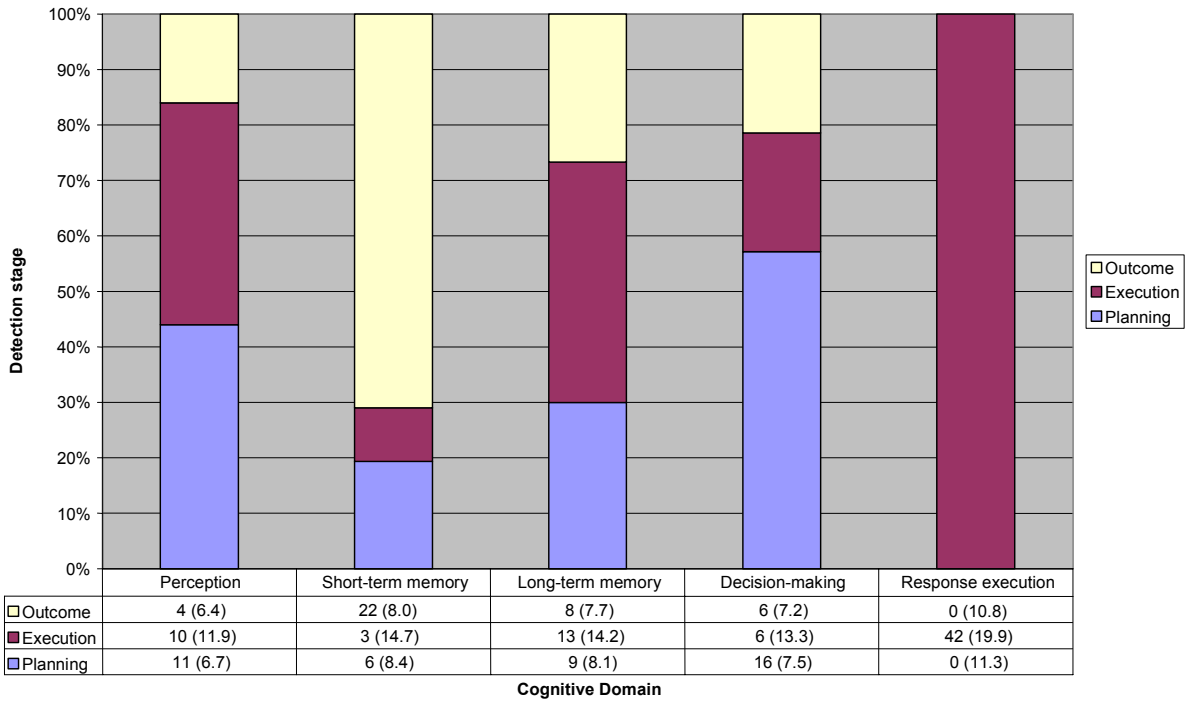
Figure 33: Interaction between cognitive domain and error detector.

Some of the interesting relationships are:

1. Perception errors are rarely detected by "No one" (AR=-2.68) and are often detected by the ATCO colleague (AR=2.79).
2. Short-term memory errors are also rarely detected by "No one" (AR=-2.14) and frequently detected by the ATCO colleague (AR=3.26).
3. Long-term memory errors are most frequently detected by "No one" (AR=6.28) or the Instructor (AR=2.80). On the other hand, they are rarely detected by the producer (AR=-5.46) or the ATCO colleague (AR=-3.23). The actor most frequently involved in the detection was the instructor (AR=2.80). These results make sense insofar as the producer has little chances of detecting errors that have occurred due to incomplete experience or knowledge. On the other hand, an important part of instructor's task is to monitor such errors and correct them.
4. Decision-making errors are rarely detected by the error producer (AR=-2.57), but are frequently detected by the instructor (AR=2.10). In this manner the decision-making errors are very similar to long-term memory errors.
5. Response execution errors are frequently detected by the error producer (AR=7.64) and rarely by "No one" (AR=-3.78). That is, the error producer has a larger tendency to detect his/hers response execution errors. On the other hand, these are rarely detected by the instructor (AR=-4.08) (at least there are no

indications that they are detected by instructor). This corresponds with the previously described hypothesis about response execution errors would frequently be self-detected.

In the chart below is shown the distribution of Detection Stage as a function of Cognitive Domain (it should be noted that only errors that were detected while it was still possible to do something about them are included in this analysis). The interaction is statistically significant, $X^2(8, N=156)=99.17$, $P<0.001$.



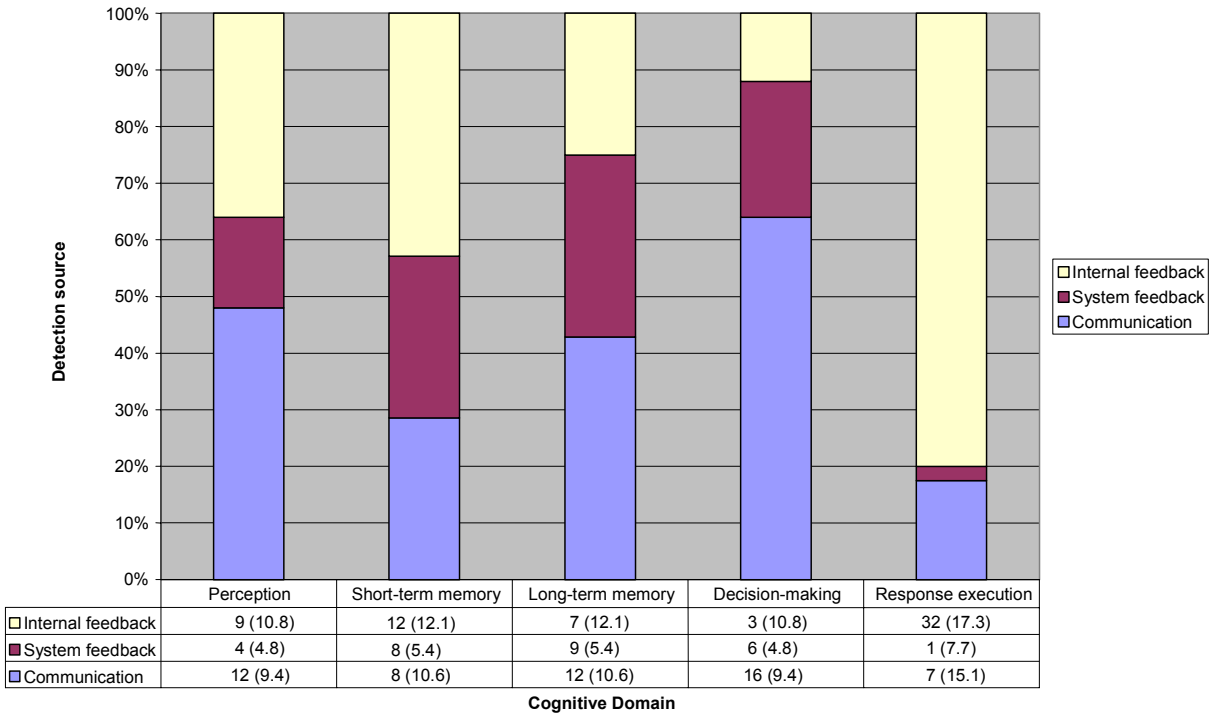
	Perception	Short-term memory	Long-term memory	Decision-making	Response execution
Outcome	4 (6.4)	22 (8.0)	8 (7.7)	6 (7.2)	0 (10.8)
Execution	10 (11.9)	3 (14.7)	13 (14.2)	6 (13.3)	42 (19.9)
Planning	11 (6.7)	6 (8.4)	9 (8.1)	16 (7.5)	0 (11.3)

Figure 34: Interaction between cognitive domain and detection stage.

Some interesting contributions to the interaction are the following:

1. Perception errors are frequently detected at the planning stage (AR=2.10). In other words, they are frequently caught very early and will have limited chances of affecting system safety.
2. Short-term memory errors are more frequently detected at the outcome stage (AR=6.46) and rarely at the execution stage (AR=-4.70).
3. Decision-making errors get detected more frequently at the planning stage (AR=3.98) and rarely at the execution stage (AR=-3.04). The detection tends to happen when the ATCOs are discussing plans for the air traffic and before they are actually implemented.
4. The largest contribution to the dependence is that response execution errors are most frequently detected at the execution stage (AR=7.98).

The relationship between Cognitive Domain and Detection Source is shown in the figure below (again, the presented data is restricted to the instances where an error detection happened while it was still possible to do something about it). The dependence between the dimensions is statistically significant, $X^2(8, N=146)=41.07$, $P<0.001$.



	Perception	Short-term memory	Long-term memory	Decision-making	Response execution
Internal feedback	9 (10.8)	12 (12.1)	7 (12.1)	3 (10.8)	32 (17.3)
System feedback	4 (4.8)	8 (5.4)	9 (5.4)	6 (4.8)	1 (7.7)
Communication	12 (9.4)	8 (10.6)	12 (10.6)	16 (9.4)	7 (15.1)

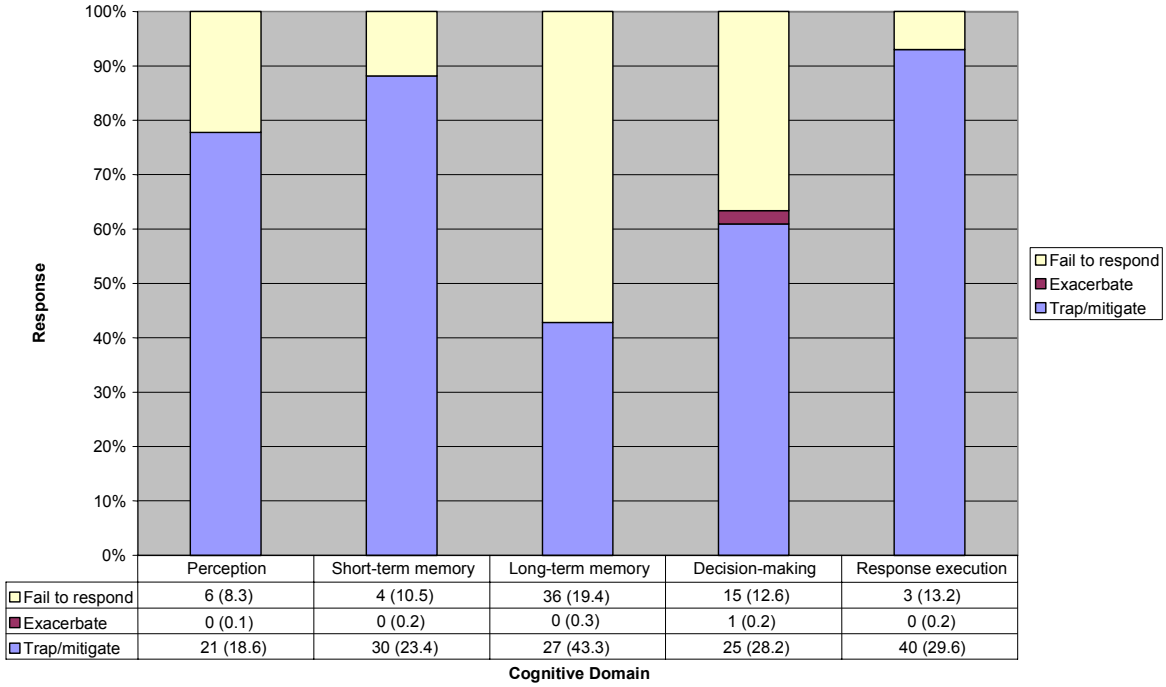
Cognitive Domain

Figure 35: Interaction between cognitive domain and detection source.

The important contributions to the interaction are:

1. Long-term memory errors are rarely detected through internal feedback (AR=-2.16).
2. Another important contribution is that decision-making errors are more frequently detected through communication (AR=2.98) and rarely through internal feedback (AR=-3.45). In other words, external communication plays a vital role in relation to containing decision-making errors.
3. The main contribution to the dependence is that response execution errors are most frequently detected through internal feedback (AR=5.52) and rarely through either communication (AR=-3.09) or system feedback (AR=-3.14). Most of the response execution errors were slip-of-the-tongue and, consequently, only the perpetrator's knowledge of the correct response could be used for the detection.

The relationship between Cognitive Domain and Response is shown in the table below, $X^2(8, N=208)=43.70, P<0.001$.



	Perception	Short-term memory	Long-term memory	Decision-making	Response execution
Fail to respond	6 (8.3)	4 (10.5)	36 (19.4)	15 (12.6)	3 (13.2)
Exacerbate	0 (0.1)	0 (0.2)	0 (0.3)	1 (0.2)	0 (0.2)
Trap/mitigate	21 (18.6)	30 (23.4)	27 (43.3)	25 (28.2)	40 (29.6)

Cognitive Domain

Figure 36: Interaction between cognitive domain and response.

The following interactions were found:

1. Short-term memory are rarely not responded to (AR=-2.63) and frequently trapped/mitigated (AR=2.68).
2. On the other hand long-term memory failures are often not responded to (AR=5.43) and rarely trapped/mitigated (AR=-5.31).
3. An important contribution to the dependence is that the ATCOs rarely fail to respond to response execution errors (AR=-3.80) and that these have a tendency to be trapped or mitigated (AR=3.86).

In sum, the interaction analysis provided some interesting insights concerning the error management patterns associated with the different kinds of cognitive errors:

- Perception errors are frequently detected and the detection frequently happens by the aid of a colleague. When they get detected it is frequently at the planning stage. Basically, these errors have a high chance of getting discovered at an early point in time and by the assistance of others.

- Short-term memory errors are frequently detected and a colleague frequently initiates the detections. The detection normally occurs at the outcome stage and the errors are frequently trapped/mitigated. Consequently, short-term memory

errors are frequently discovered and other people play a significant role in relation to controlling the effects of this kind of error.

- Long-term memory errors are frequently either not detected or detected by the instructor. The fact that they are rarely detected also means that they are rarely trapped/mitigated. This kind of error is in general difficult to discover by the error producer insofar as people with this type of error do not have a standard for comparing their own behaviour with (even though they might have meta-knowledge concerning the limits of their current knowledge and experience).

- Decision-making errors are rarely detected by the error producer, but frequently detected by the instructor. When they get detected it is frequently at the planning stage and this happens frequently through external communication. This result is highly interesting because it is an integrated part of many TRM and CRM courses to encourage the participants to clearly state their plans and intentions so it is possible for other people to be in the loop and be able to criticise potentially flawed plans. The results from this study provide support to the importance of this philosophy in relation to containing decision-making errors.

- The error producer frequently detects response execution errors. The detection normally occurs at the execution stage and through internal feedback. The errors are frequently trapped/mitigated. These results clearly show that people contain robust internal mechanisms to control this type of error.

9.3.3 Validity

The issue of validity of the framework will be explored in detail in chapter 12. In the current context the focus will be on the content validity or the comprehensiveness of the framework based on the results from the simulator study. A way to get an indication of this is by looking at the amount of "Unknown" classifications within the individual dimensions of the framework. The amount of "Unknown" classifications are shown in the table below:

Error	
Error producer	
0 (250)	0,0 %
Cognitive domain	
11 (250)	4,4 %
Detection and recovery	
Who – detection	
1 (250)	0,4 %
Who – correction	
1 (250)	0,4 %
When – detection stage	
2 (195)	1,0 %
How – detection source	
7 (195)	3,6 %
Response and outcome	
What – response	
0 (250)	0,0 %
What – outcome	
0 (250)	0,0 %

Table 9: Amount of "Unknown" classifications for each dimension (study 2)

As can be seen from the table only a very small percentage of the classifications could not be determined. The dimensions that had the highest rate of unclassifiable events were Cognitive Domain (4,4%) and Detection Source (3,6%)[17]. Hence, the application-rate of the categories within the individual dimensions of the taxonomy seems to support that the framework covered the observed events in a comprehensive manner.

Another way to get a picture of the content validity is the extent to which the variety of categories within the framework was used. In the current context it is of particular interest to note that the classifications associated with the detection processes (i.e. the when- and the how-question) resulted in a much more varied pattern than in the previous study based on incident reports and, consequently, all of the categories within these dimensions revealed their relevance.

[17] It should be emphasised that the "Unknown" categories can be split into two different kinds of groups. The first group concerns cases where the information in the data material is in itself incomplete and it is therefore not possible to determine the underlying causes of a given phenomenon (e.g. the cognitive foundation of an error). The second group is related to situations where, even with all the information available, it would not be possible to determine one specific category. For example, an ATCO may decide that he will follow up on a potential conflict between two aircraft later on, but fails to do so with the result that a separation violation occurs. In this case the error might be attributed to forgetting to carry out the intended action. However, it could, in principle, also be a perception error insofar as the ATCO does not discover the emerging conflict when it gradually becomes more and more salient. In the current context there are not made any attempts at distinguishing between these two sources of indeterminacy.

A dimension that did not display much variation in this study is the what-dimension. Within the response-dimension the "exacerbate"-category was only used once and the relevance of this category could therefore be questioned. In response to this it should, first of all, be stated that it is fortunate that the response following an error is rarely leading to an even worse situation. Secondly, the category seems relevant to maintain within the framework to cover all possible types of responses that can be produced and it might be useful to give particular focus to errors that become exacerbated insofar as they might constitute an especial safety risk that could require intervention (i.e. they contain a higher probability of incident or accident). The other part of the what-dimension, namely the outcome, also resulted in a limited variation. However, this is a direct consequence of the type of error material used.

9.4 Conclusion

In this chapter error events in ATC simulator training scenarios were analysed. In the study a high degree of reliability was found across all of the examined dimensions within the framework. This was both for the intra- and the inter-rater reliability analysis. As expected the intra-rater analysis yielded higher results than the inter-rater analysis. Nonetheless, across all dimensions of the framework robust analyses could be obtained. These results are promising – in particular, because it was in this study not necessary to collapse any of the categories within the dimensions of the framework.

The only dimension where a somehow lower (but still acceptable) agreement was achieved was the outcome-dimension. This result was due to the fact that in this study it was chosen to apply a broader definition of "undesired state" compared with the study with incident reports in study 1. This was necessary because no cases occurred where the separation standards were violated and therefore a list of less serious "undesired states" was developed. The list of undesired scenarios used in this study was in no way complete and to get more comprehensive list of generic types of undesired states it would be useful to include subject matter experts.

Since this study focused on normal everyday errors it was expected to see some deviations from the error and error management patterns found in the incident reports. This expectation was also confirmed. For example, the simulator study showed a much more equal distribution of errors (with a slight majority of long-term errors due to the inexperience of the ATCOs) compared with the incident reports where a large majority of errors were decision-making error. In similar vein, a much more varied distribution of detection stage and detection source was found in the data from the simulator scenarios.

Also many interesting types of interaction were found between categories from the different dimensions. It was, for example, shown that different kinds of cognitive errors had a tendency to become detected by different actors (if any) and, furthermore, these errors were detected through different kinds of detection sources and at different kinds of detection stages. The results were in good concordance with the a priori established hypotheses (which will be elaborated later on) and at the same time new insights were

produced. Of new insights can be mentioned that decision-making errors have a larger tendency to be detected through communication than other errors. It was also demonstrated that both perception and short-term memory errors have a tendency to be more frequently detected than other errors and the detection frequently happen by the assistance of the colleague.

A limitation in the study that should be mentioned is related to the classification of errors. One of the two classifiers was also the one who originally made the descriptions under the heading of "contextual information" (see section 9.2.2). Ideally, it would have been more optimal if the one making the descriptions and the people classifying the error events were completely independent. This was, however, not possible due to pragmatic reasons (i.e. resource limitations).

Another limitation in the current study is that no information was ever elicited from the trainees involved in the training scenarios. All the information elicitation was done on the basis of the video recordings, the communication and comments from the instructors. Therefore, there were some inherent limitations in the amount and quality of information that could be obtained to understand the underlying processes of the behaviour and it was necessary to make assumption concerning the knowledge and intentions of the ATCOs. Furthermore, errors that did not result in observable behaviours or verbalisations could not be identified and analysed. If it had been possible to use think-aloud protocols (retrospective) it would have been possible to gain more detailed insight into the cognitive processes underlying the behaviour of the trainees and, most likely, it would have been possible to get a better foundation to base the classifications on.

Even though it was not possible to get information from subjects involved in the scenarios subject matter experts had a very important contribution to the study in the identification and explanation of many error events. In particular, a large part of the long-term and decision-making errors could not have been identified and understood without the help of these subject matter experts. This fact clearly illustrates that for any study that tries to tap into the underlying cognitive processes of human behaviour in a complex domain such as ATC it is important to include domain experts in the analysis because they have a much better understanding of the context.

10 Study 3 - Expert evaluation

10.1 Introduction

The purpose of this study was to get an expert evaluation of the framework. This was considered important feedback insofar as it could be used to get some indication of the strength and weaknesses of the framework seen from both a theoretical and practical point of view. The focus of this study is on obtaining input concerning the face/content validity of the framework from human factors experts who have experience with developing and/or applying conceptual human factors frameworks. Furthermore, these researchers bring along experiences from many different domains which is useful when considering the more general usefulness of the framework (and which modifications should be considered if the framework should be used in other domains).

10.2 Method

10.2.1 Subjects

For the study a series of relevant participants were selected on the basis of the fact that they had been involved in research that was highly relevant in relation to this project. This included other conceptual and empirical work related to human error and error management. Also researchers who had been involved in development of comprehensive conceptual human factors frameworks were considered highly relevant for the current project. A total of 21 researchers were identified and 11 of these responded to the questionnaire (response rate: 52.3%). A few additional responses were received in the form of informal comments.

10.2.2 Questionnaire

Each of the participants in the survey received two documents. The first contained a short description of the framework and its main components. The second contained a questionnaire where they were asked to give their opinion of both the components and the overall structure of the framework. The main focus was on the relevance of the individual items within the framework. This was done, first of all, on the basis of a rating scale from one to four (1=Irrelevant, 2=Somewhat irrelevant, 3=Somewhat relevant and 4=Highly relevant). In addition, the subjects were encouraged to give free-text comments to the individual items. The questionnaire could be filled out electronically and e-mailed back to me.

10.3 Results

In the questionnaire the participants were asked to give both quantitative and qualitative feedback concerning the framework.

10.3.1 Quantitative results

In advance it was decided that average ratings below 3 would be considered critical for the relevant item and, consequently, the item might have to be dropped from the framework. In the figure below is shown the average rating of the main dimensions within the framework on the basis of the questions in the questionnaire. As can be seen in the chart all of the average ratings were between "somewhat relevant" and "highly relevant". In other words, the experts found all of the items relevant. The dimension that received the lowest rating was "detection stage" (the when-question).

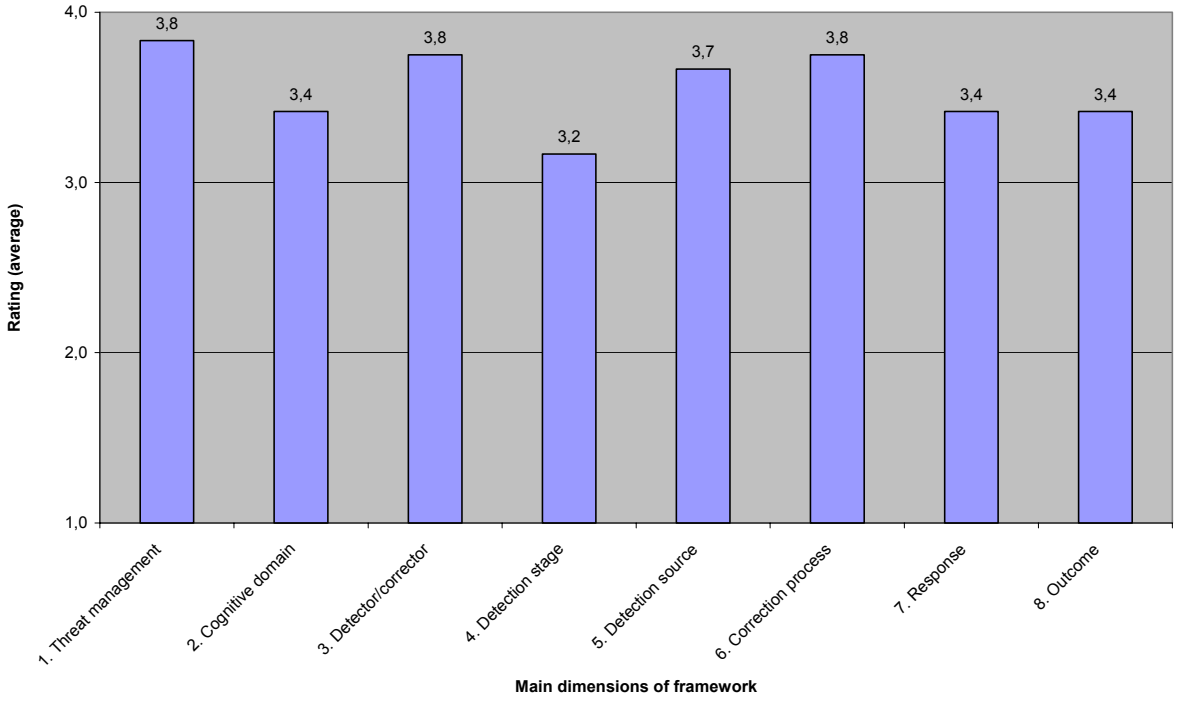


Figure 37: Ratings for core components of the framework

In the next chart is shown the ratings of the main groups of PSFs. Also in this case the dimensions received a high average relevance rating. The two dimensions receiving the lowest rating were “ambient environment” and “person related factors”.

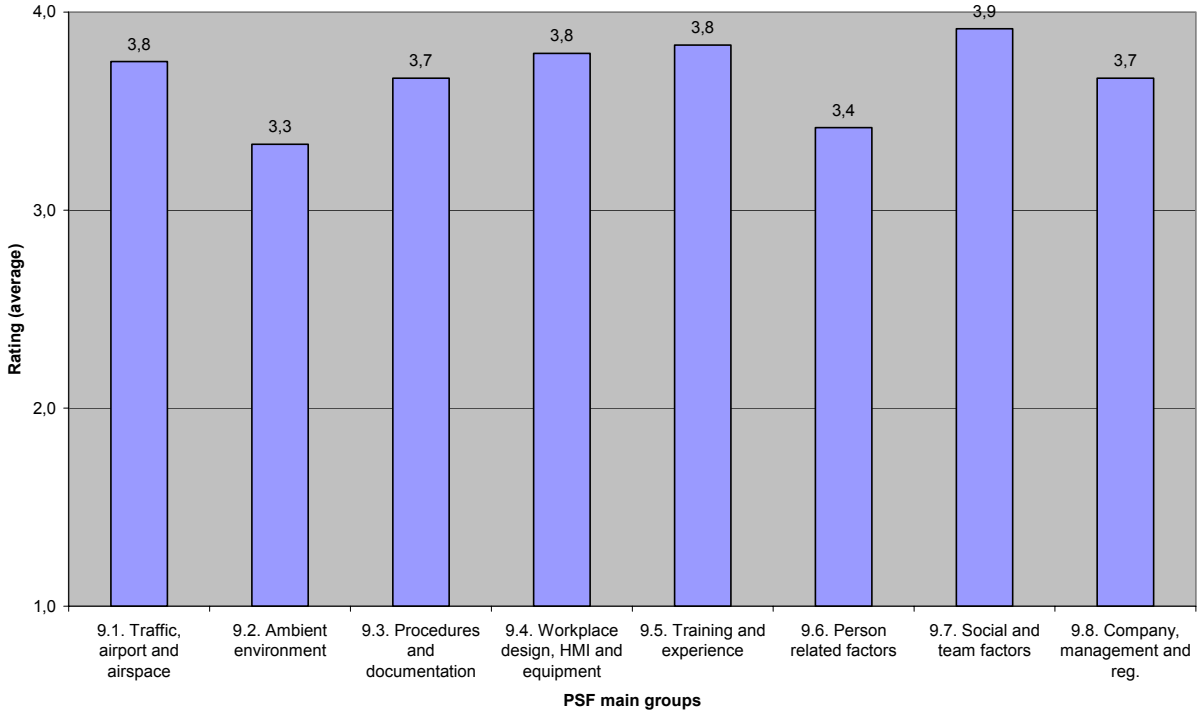


Figure 38: Ratings for PSF-components of the framework

Finally, the overall ratings of the PSFs and the framework in general are shown below. In concordance with the previous ratings the experts rated both the PSFs and the overall framework high on relevance.

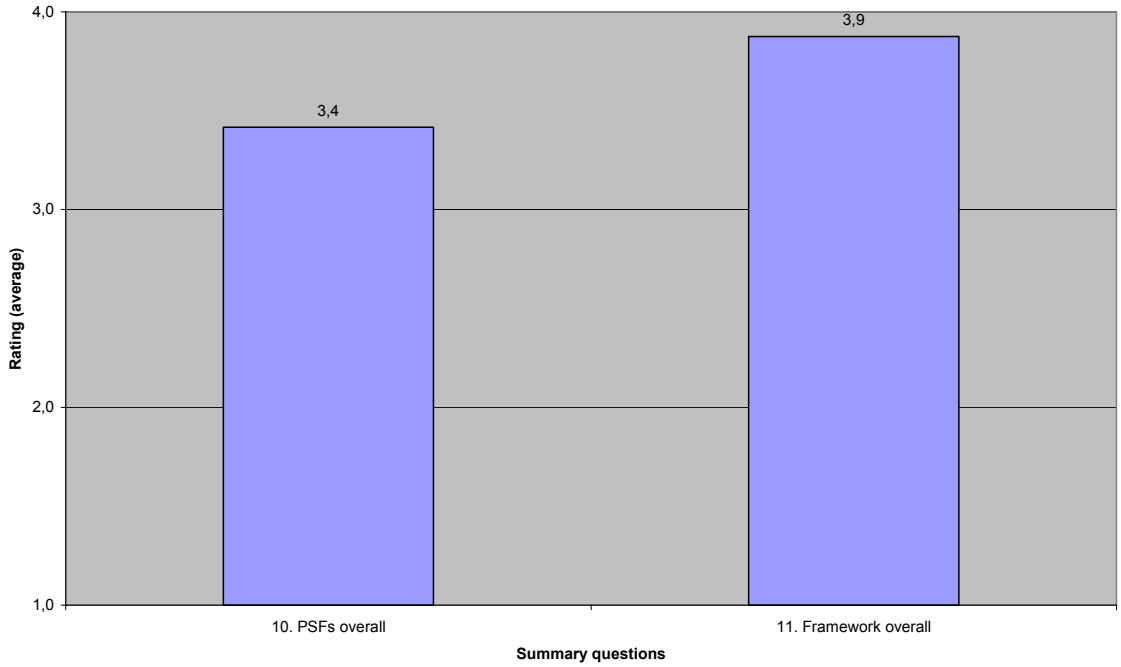


Figure 39: Summary ratings of the framework

In conclusion, it can be stated that the experts gave high ratings on all of the dimensions within the framework and its overall structure. This has a positive and a negative aspect. The positive side is that the framework has a high degree of expert acceptance. In particular, it is encouraging that all of the dimensions, which have been given only limited attention in the existing literature, actually were well accepted by the subjects. The negative side is that the responses contained a limited variability and, consequently, the diagnosticity of the ratings is also reduced. To gain a more detailed understanding about strengths and weaknesses we will now turn to the qualitative comments.

10.3.2 Qualitative results

In the following is reviewed some of the qualitative results that were elicited through the questionnaire. It should be noted that different researchers might have different focus of interest and different experiences. It would therefore be impossible to completely accommodate all their comments into the framework. Nonetheless, many very useful comments were made. If several researchers displayed a similar view on a certain subject only one of these will be included. The comments presented in the following are structured according to the individual items on the questionnaire (and, hence, the individual dimensions of the framework).

Threat Management

> *"The difficulty lies in knowing whether or not people really have anticipation. Hindsight may not be accurate."*

In relation to retrospective analyses it is correct that some degree of hindsight bias may distort the results that can be obtained from this dimension. That is, people would like to present themselves in a favourable light and therefore might tend to exaggerate the extent to which they were aware of task relevant properties (i.e. threats). This problem is less likely to be present in studies based on real-time analyses. For example, studies at University of Texas have been successful in conducting observational studies of pilot crews' threat management behaviours (Helmreich et al., 1999).

> *"Preparedness (lack of) is important in the causal development of maritime accidents, but it is difficult to examine empirically in the everyday routine work since it is not always reflected in the behaviour or communication."*

Even though it is not easy to study details related to threat and error management without insight into the processes underlying the overt behaviour it does not mean that it is not possible. Again, studies at University of Texas have demonstrated that it is possible to train observers in analysing threat management of flight crews.

> *"Related to training, task familiarisation - quite relevant"*

Threat management is definitely an issue that should be addressed in training courses. For example, in NavCanada they have already initiated ATC training courses with the goal that "the participants will develop strategies to manage threats and errors in the operational environment" (Down, 2001).

> *"I believe that threat management is highly relevant to the whole process of error management especially where anticipation is involved. This is part of the actors being aware of all the threats within their enviroment that can lead to errors."*

> *"A majority of maritime accidents is actually caused by lack of anticipation, confirmation bias or lack of situational awareness at this very early stage or level. However, I think this problem is also a matter of perception and attention."*

The observation is correct that threat management is closely related to other theoretical concepts such as situation awareness (Endsley, 1994). Both of these concepts are related to having a appropriate understanding of the task-relevant elements of the environment and being able to make anticipations about the future state of the environment. The concept of threat management is a bit more comprehensive since it also covers successful

and unsuccessful responses associated with the comprehension of the task-relevant elements and their future state (see the definition of threat management in section 3.3.2).

> *"You could also treat this as a performance shaping factor, related to training and also to experience etc. I have actually also singled out a few performance shaping factors about which I specifically record information for the incidents I have analysed. I think it is hard to decide which factors are the most important, especially since they may vary from case to case."*

It is true that threat anticipation, in principle, could be analysed as a PSF. Nonetheless, the reason for not choosing this solution is that effective threat management may result in no error occurring at all. Since important lessons can be learned from analysing how threats are effectively or ineffectively managed, it seems reasonable to include the threat part as an integrated part of the core of the framework.

> *"It is important whether anticipation results in prevention or in contingency planning, as you have commented in your full text."*

This point is relevant because even though errors and associated problems might be anticipated to some extent it is not necessarily the case that the controller will respond effectively in advance by either avoiding these problems from occurring or being prepared with solutions if problems should occur (i.e. contingency planning). Consequently, it is relevant to consider whether the anticipation-category should be subdivided in such a way that it is possible to distinguish between anticipation associated with precautionary initiatives and anticipation not associated with any precautionary initiatives. By including the potential response into the distinctions the categories become in better concordance with the definition of threat management (see section 3.3.2).

> *"...One aspect I was not sure of though, and it is obviously central to the framework you're proposing, is the 'threat assessment'. I'm not too much aware of this concept and its rationale, and was not sure how this bit gets used. Now I know that Bob Helmreich has been doing this for a while, but I've not properly understood its application. As an example, would you consider 'wrong threat identified?' and 'partial threat identified', and of course 'threat misunderstood' - we get into error analysis of error analysis here, but that could be useful."*

The problem with this suggestion is that if we start analysing different ways in which the threat might be incompletely or inappropriately managed this will introduce some redundancy into the classification system insofar as the next stage in the framework is concerned with the analysis of the error.

Error – Cognitive domain

> *"The devil is in the detail. At this level of description, I'm not sure what can be usefully deduced. In other cases, however, this is all you can deduce from incident reports etc. These domains are the same as TRACEr Lite [i.e. the precursor to HERA], which is being used in Manchester Area Control Centre for incident investigation. However, they are broken down into several error modes and error mechanisms - about half as many as HERA and TRACEr."*
>
> *"Why is 'interpretation' missing ?"*

The above comments are related to the level of detail and the comprehensiveness of the cognitive error analysis. Concerning the level of detail it is clear that the different cognitive domains can be broken down in much more detail (as in the case of the HERA framework). Whether this is a good idea or not is dependent on a several factors. First of all, the level of detail should be dependent on how much information that can be obtained from the data material and the extent to which it is possible to make reliable distinctions between subcategories. Secondly, the practical use of breaking down of a highly detailed analysis of the cognitive mechanisms underlying the cognitive domains should be weighted against the extent to which an adquate amount of examples of subcategories can be obtained so that it would possible to make meaningful statistical analysis.

> *"I think we talked about this before: you only study recovery from errors (that is, human errors) and not just any kind of failure."*

Even though the focus in this project has been on the management of errors it should, in principle, not be a problem to adapt the framework to management of faults as such (actually, faults can be considered as a specific subgroup of threats).

> *"This is a psychological definition of human errors. You can also use an operational definition of human errors along the lines of J. Rasmussen. In the latter case you have: observation errors, identification errors, interprepation errors, decision errors, planning errors and response errors. Because, your framework is big, you should better use one or the other classification."*

The choice between different types of human error taxonomies has been discussed in the literature review section and will therefore not be repeated here.

***Who* was involved in the detection and recovery of the error and/or its consequences?**

> *"Obviously, may be several of above. Word 'context' will not be familiar to investigators though."*

It is correct that sometimes several of the possible actors might apply to the same situation insofar as the ATC environment is a distributed system where several actors are involved in the process and can, in principle, carry out several independent recovery related activities (actually there are examples where pilots do not respond according to the instructions by the ATCO or sometimes even do the opposite which makes the analysis difficult). However, on the basis of the segmentation and classification principles previously described it should, in most cases, be possible to identify a single actor (please note that the segmentation and classifications principles were not described in the paper sent out to the participants). In relation to the word "context" it was the intention that the distinction between "co-actor in context" and "co-actor outside context" should be adapted to the specific study at hand and not, for example, be left to the investigator's judgement to decide when an co-actor is within or outside the context.

> *"Think this is important - many cognitive errors are due to fixation, and therefore it is important to see who produced the error in relation to who detected/corrected it."*

This is a correct observation insofar as fixation-errors are most frequently detected by a person different from the error producer.

> *"I think, that it is important to collect data about "why" the actor did or did not detect/correct the error or state. Otherwise you could not draw any conclusions from the data collected in (3) [i.e. The Who-Question]. Example: Why did the system not detect/correct the error or state: Not designed for it, out of order, not operated according to instructions or procedures etc."*

It is true that it is of critical importance when analysing error management events to obtain information about the underlying causes of a successful or unsuccessful detection and recovery. However, what is referred to as the "why"-question above is to a large extent covered in the Performance Shaping Factors taxonomy.

> *"I think it is difficult to decide how many different types of actor to distinguish. For example, it may also be relevant to look at the other involved actor's role with regard to the one who committed the error (co-worker, supervisor, trainer, etc.)."*

Different domains may differ with regards to how many types of actors that could be involved in the error management. In the ATC context it seems to be a relatively restricted amount of groups of actors that can be involved in the error management.

"Another comment is that you don't seem to record anything about the stage where the problem is defined and the causes are identified (explanation phase). I realise that this phase does not always occur, but in some cases it may have a very important role, also because of the negative consequences of being wrong during this phase."

The concept of error identification was deliberately chosen not to be included in this framework. This was done because the identification phase is rarely carried out before the correction (and therefore not relevant for the correction). Even more important in the current context is that error identification does not seem to be relevant when it comes to solving problems in ATC. This issue might be different for other domains (e.g. in the medical domain or in the process control industry) and the relevance of the phase should therefore be carefully considered when analysing other domains.

***When* was the error or its consequences detected?**

"I also use these stages in TRACEr (full version) - basically from Sellen [1994], etc. - but in practice it is hard to classify, or else all are 'Outcome'."

This is a relevant comment insofar as the analysis of the incidents has indicated that only a very limited distribution of the categories can be obtained within this dimension (see Figure 21). On the other hand, the analysis of real-time data has indicated a much more varied distribution of data (see Figure 28). Hence, the relevance of this dimension seems to be dependent on the type of material used in the analysis.

"'When' is may be not as interesting as how? When detected should also be seen in relation to when it was recovered."

"...don't you think it is also interesting to look at when a problem is recovered - people may delay their response for a while after detection as other things may be more urgent or the potential consequences won't occur for quite a while or are not so serious?"

In most cases the detection and the correction will occur at the same stage. However, it is true that sometimes they may occur at different stages (e.g. when choosing to postpone the correction). Therefore, it could be of relevance to include both the detection and recovery in the analysis of the "when"-question.

"This is somewhat relevant as timely action can have an effect on outcome."

> *"Just to share some of my findings with you: only for human error it seems to make sense to look at the performance stage in which the error is detected. Other failures are most often only detected in the outcome stage."*

If the framework should be expanded to also include technological faults it is clear that the detection stages would not be applicable to those events. The detection stage taxonomy is based on cognitive stages and is therefore only intended to be analysed for the recovery of human errors. However, it should be emphasised that in the current context the focus is on the management of errors and not faults in general.

***How* was the error and/or its consequences detected?**

> *"Again, there is a link between this part of your framework and the performance shaping factors"*

This comment seems to be suggesting that the detection source should be included under the PSF section instead of a part of the core of the framework. Which aspects to be included as performance shaping factors and which to be included in the core of the framework is a choice that is dependent on the researcher's discretion and goals. In the current context it was chosen to be included as an integrated part of the core because the detection source was considered an issue of general importance for all cases of error recovery.

> *"Does system feedback also include operating procedures? Many errors are detected through systematic work routines/ go through procedures."*

It is very uncommon to look up procedures in ATC so this issue has not been encountered during the ATC studies. However, in many other domains such as process control and aviation the use of checklist procedures is much more common. In cases where errors are detected on the basis a standard checklist that is required being carried out in specific situations it would seem reasonable to classify these as detection on the basis of system feedback. However, it should be noted that system feedback is a fairly broad category so it might make sense to divide it into sub-groups (perhaps dependent on the specific domain being studied).

> *"I think - as far as the maritime domain is concerned - that the "system feedback" should be divided into "automation feedback" and "environment feedback". It is worth knowing if the cues were found in the automation system or by "looking out of the window". It is often said that information from the automation system should be cross-checked visually, and it would be interresting to know if that makes sense at all when it comes to the detection of errors."*

There seems to be two important points in this comment. First, it is suggested to subdivide the system feedback into two categories dependent on whether the information is derived from the system interface or directly from the physical surroundings. In principle, such a distinction could also be applied to the ATC environment insofar as error detection can occur when a tower controller or a pilot looks out of the window. Nonetheless, the distinction would not make any sense for the approach and en route position insofar as they have to depend on mediated sources of information (i.e. they have no direct sighting of the aircraft). In the comment it is also suggested that crosschecking or verification of information should be a separate category. This corresponds to the category previously referred to as "standard check". As it was discussed then the category does not fit very well with the ATC environment, but it is a category that could be of more relevance in many other contexts. This issue is also addressed by some of the following comments from another human factors specialist.

> *"You must add another three mechanisms of error detection:*
> *(1) ROUTINE CHECKS (2) SUSPICIOUS CHECKS and*
> *(3) CONTINGENCY PLANS."*

This comment was further elaborated in a separate response to me. Here it was suggested that there are in total six different ways of detection.

> 1) *Communication*
> 2) *System feedback*
> 3) *Inner feedback*
> 4) *Routine check*
> 5) *Suspicious check*
> 6) *Contingency plan*
>
> *Communication - system feedback - inner feedback are PASSIVE BEHAVIOURS because the Air Traffic Controller (ATCO) does not initiate a check or he is dependent upon the performance of the system or his colleagues. You definitely need to have some ACTIVE BEHAVIOURS in your framework. In my paper in Safety Science, I call them planning behaviours because the ATCO must be vigilant and take initiatives to inspect the products of his own work.*
>
> *Routine check can be a standard check, a revision plan or an external plan comparison. You should better group them all as "routine check". Examples are when the ATCO occasionally makes a standard check of his progress towards his goal (i.e., put all aircraft in a sequence, so that all are landed with 5 minutes difference). ATCO and pilots follow checklists (external plan comparison). I guess that many errors are detected by pilots when they use their checklists! Finally, ATCOs under stress tend to revise their work to see if they have missed something out.*
>
> *How often should ATCOs review their plans? This is an important issue. Expert ATCOs know this so that they can catch any error*

without spending too much time on their plans. So you must have a category ROUTINE CHECK here.
Suspicious check refers to what I call "error suspicion and curiosity". Sarter & Alexander [2000] also use this category. They have found some data for pilots and you may find some data for ATCOs. A suspicious check is different from a routine check because the former does not necessarily mean following checklists.
A contingency plan is similar to a suspicious check but different in some respect. For instance, a contingency plan means that the ATCO has decided in advance WHEN and HOW he is going to review the situation and check for errors. Contingency planning is similar to anticipating errors on his own. In your model, you have THREAT ANTICIPATION as a starting box. You can have ERROR ANTICIPATION as an error mechanism. Error anticipation is contingency planning.
Careful. A suspicious check is different than a contingency plan. The ATCO does not make any planning in a suspicious check. In fact, an ATCO can be suspicious because he has not made any good anticipation or because he has no contingency planning. The fact that an ATCO is suspicious may imply "lack of contingency planning".
When you find some "near miss reports" or any simulator data, it is better to keep your list of detection mechanism open. My guess is that it would be difficult to find how contingency plans contribute to error detection. Only if you have simulator data, you will be in a position to make an in depth analysis of detection mechanism.

These are very detailed and interesting comments about error detection mechanisms. In particular, the distinction between active and passive detection mechanisms seems to make a lot of sense. The main problem with the active detection mechanisms is that it can be associated with some problems applying them directly to the domain of ATC. An example is routine check: In principle, the ATCO is monitoring the situation all the time (it is an integrated part of their work to do a "sweep" around the radar) and it is therefore difficult to say that an error was detected on the basis of standard check. Another example is suspicious check: If, for example, an ATCO sees two aircraft getting in conflict with each other he does not have a suspicion about a problem - he can tell for sure (this is related to a more general issue, namely that the concept of diagnosis - in contrast to many other safety critical domains - is not applicable to ATC). Even though these two types of detection mechanisms are not expected to be frequent in the domain of ATC they have been included under the category of "Internal feedback" to ensure that the framework is comprehensive (see section 6.3.7). In relation to the last type of "active behaviour" - namely contingency planning – it is true, as described in the comment above, that this factor is directly related to the issue of threat anticipation and

management. Therefore, to also have this as a separate factor under detection mechanism would introduce redundancy in the framework.

***How* was the error and/or its consequences corrected?**

> *"Makes sense to me, but not so much to controllers and investigators. Depends on who is classifying. Apply rule, choose option and create solution could be interchangeable though."*
>
> *"I had difficulty distinguishing between the 'apply rule' and the 'choose option' – would suggest a better explanation on these…"*

These comments are related to the reliability and mutual exclusivity of the cognitive categories. Empirical data in the next study will be used to get a measurement of the reliability of these distinctions. Nonetheless, the concepts were chosen because they appeared intuitively understandable and did not seem to require any theoretical understanding to be applied.

> *"This is for me an interesting phase. But: cognitive errors could also be corrected by the automatic system, by a change in the environment, by chance, etc. Have you chosen to leave out all solutions that are not cognitive related in your framework? If so, I don't find anything wrong with it, as long at the persons using the framework are aware of this. I also think that it is important to distinguish between application of actual rules (like procedures, work orders, etc) and mental rules (work models)."*

Concerning the first comment the answer would be that the classifications related to the problem-solving are only relevant in the cases where a human operator is actively involved in process (the rules for transitions between the dimensions in the framework were not explained in the document that was sent to the participants in this study). The distinctions do not make any sense when an automated agent is the active part in the error correction. In relation to the second comment it could be interesting to distinguish, as suggested, between responses based on formalised procedures and automated responses based on a high degree of experience with certain types of situations. However, in the domain of ATC which is, as previously described, less proceduralised this distinction is less relevant.

> *"I like this part a lot, I also think the categories you have chosen capture all the different possibilities."*
>
> *"This is very good and useful!"*

What* was the behavioural response*?

> *"Again important information is missing if the question "why" is not asked. It is for example important to know if the lack of response is caused by (1) error not detected, (2) error detected to late or (3) simply ignored."*

It is true that the dimension in it self does not give an answer to the why-question. Nonetheless, when combining it with other dimensions (e.g. error detector, problem-solving associated with error management and the PSFs) these questions can be answered.

***What* was the outcome?**

> *"There seems to be an overlap between the "what" and the "when" process."*

The main difference between the what- and the when-question is that the when-question is a cognitive classification whereas the what-question is related to the consequences. Nonetheless, the observation that the two dimensions are partially overlapping is correct. Consequently, it seems reasonable to consider whether it is relevant to keep both of these dimensions in the framework.

> *"In case of an undesired state or an additional error the severity should be evaluated. The undesired state or additional error could actually be less severe than the initial state and thereby be an improvement. Or they could be more severe and thereby leading to a situation that is worse than the initial situation."*

There are several aspects in this comment. First of all, as mentioned before it could be relevant to have more exact details concerning the severity instead of just using "undesired state". This could either be on an ordinal scale (e.g. ranging from 1 to 5) or by having a list of general undesired situations (as has been done in the LOSA system at University of Texas). Concerning the second issue in the statement – about being able to obtain information concerning whether the recovery action affected the situation in a positive or negative way – it seems obvious to combine results from the two what-dimensions, namely response and outcome. For example, the combination of trap/mitigate and undesired state means that an error was not prevented until an undesired state did occur, but it was eventually prevented from developing into an even worse situation. In similar vein, the combination of "exacerbate" and "undesired state" means that the response to an error (or its outcome) created an even more sever situation.

> *"In my own research I have recorded more details regarding this question (severity and type of remaining consequences etc.) but this may be a bit 'over the top'. I agree that well-intended*

recovery actions may lead to additional error. But I have not been able to collect information about such cases... have you?"

The issue of remaining consequences and the recovery of these might be of relevance in some domains (in particular, process control). However, it makes less sense in ATC to talk about remaining consequences insofar as things here tend to be more discrete. Either a situation is resolved without any physical consequences (either spontaneously or through some kind of intervention) or a disastrous and unrecoverable situation occurs. In neither case there does not seem to be any "remaining consequences". In relation to the question about "additional error" the relevance of this category has been supported by the two previous studies (see Figure 24 and Figure 31).

Performance Shaping Factors

"I think that they are all relevant when it comes to the identification of causes for human error and their avoidance and recovery. But I also think, that you would find, that they play very different roles, and contribute very different to the overall causality behind these phenomena."

It can be expected that different PSFs might have different effects on error production, detection and recovery (as also suggested by Van der Schaaf & Kanse, 2000). Furthermore, they might vary concerning their tendency to contribute positively or negatively to error events and safety in general. Therefore, it is of importance to include information about both the stage that they affect and the kind of influence they have on the analysed events.

"Too many PSFs? It might be difficult to rate positive/negative contribution for some of them. Also feel that some of them are at different levels. A bit difficult to see a clear connection between this and the rest of the framework. I think that PSFs are important per se, but some PSFs may e.g. be negative in the detection phase and positive in the correction phase - it is not possible to illustrate this as the framework is now."

There are several points in this comment. First of all, whether there are too many PSFs or not is an open question. The list that the participants in this study was presented with was a boiled down version of contextual factors found in other frameworks. Nonetheless, further condensation might be possible (e.g. the "Ambient Environment" dimension might be moved insofar as it contains few categories and received relatively lower expert ratings than the other dimensions). In addition, empirical data from the critical incident study might give an indication of whether it is possible to boil it even further down. Concerning the relationship between the PSFs and the core of the framework it was deliberately determined not to try depicting this in this model insofar as they can affect all stages in the model and to try to depict this would compromise the overview that the model provides.

> *"If the list of PSFs is going to be used in the maritime domain, it should be expanded in certain categories. The category number 6 [i.e. Person Related Factors] should for example be expanded with items covering deprivation from family and domestic issues and concerns."*

It is clear that adaptation will be necessary when applying the PSFs to different domains. The is a direct consequence of the fact that PSFs described major aspects of the operational context and some of these contextual factors will inevitably be unique to the specific domain of interest. However, the example mentioned in the comment is perhaps not a factor that can easily be associated with a specific error and consequently might not be a very diagnostic factor to include in the PSFs.

> *"…distinctions between different levels of flexibility should be made. Some factors can be easily changed or corrected leading to improved performance while other factors are more or less stable and unchangeable or dependent on parameters which can not be manipulated ("outside reach")."*

This is a very valid point. The current framework is mainly focused on identifying factors that contribute positively or negatively to the error episodes. However, if safety-enhancing initiatives should be generated some additional information might be required to make an effective prioritisation of which PSFs that should be targeted. In particular, safety managers might benefit from information about the severity and changeability of the individual factors (as has been suggested by the Human Factors Research Laboratory at University of Texas).

> *"If you wouldn't cluster the long list of factors, it would be very difficult to use, so grouping is important.*
> *Another point I would like to make is that there are direct and indirect influences, some of these factors influence recovery via another factor."*

It is true that several layers of factors might affect each other in a way similar to Reason's Swiss Cheese model. However, as previously described the more distant the factors become the less direct influence they have and consequently the more difficult it becomes to estimate their contribution to individual errors. Furthermore, to establish relationships between several distant layers on the basis of errors observed at the front-line seems to be a daunting task.

> *"There are many ways to "cut the cake" here! It is very difficult to say what is the most useful classification scheme for PSFs. Yours is Okay."*

The overall framework

> *"I expect that the user of the framework will have the table displayed in a way that makes it easy to fill it in. Are you thinking of using the framework on-line in simulator studies/training, or more as an analysing tool?"*

Whether the framework can be used or not to do on-line scorings is difficult to answer. However, the table with the framework is fairly simple to fill-out and comparable frameworks (such as the LOSA) have been successful in conducting on-line analyses of error and error management events. For on-line analyses to be successful it would require that the observer has received extensive training in using the taxonomy and at the same time it is necessary to have a high-level of domain knowledge.

> *"Is it supposed to be a purely classification system, or do you plan to develop it into an HRA methodology, assigning numbers to the different boxes?"*

Currently, it is just intended to be a descriptive classification system. Whether it can be adapted to a HRA methodology is an interesting issue, but outside the scope of this project.

> *"Of course I think the framework is very relevant for the analysis of recovery processes…I can definitely apply this to the chemical process industry domain. The main things I miss are the explanation phase, and timing of the recovery steps after detection. Also, what do you do if multiple corrective steps are involved in recovery from one error, involving different actors at different times?"*

Some of these issues have been covered in previous responses. In relation to the issue of "multiple corrective steps" the answer would be that it is correct that several corrective steps might be carried out by different actors and with different effects on the situation. However, as previously described, to enhance the chances of obtaining useable results from the error management analysis some segmentation principles have been introduced (see e.g. section 8.2.2). The main purpose behind these principles is to provide a consistent and logical way to structure the segmentation so that the event description can be adapted to a traditional "flat" table. If the recursive nature of these recoveries should be maintained this would seriously compromise the chances of carrying out statistical analyses on an error management database and, consequently, the practical utility of the framework would be reduced.

> *"All [items] are relevant and valid in my opinion, but not all so useful in practice."*

"Generally I think the framework looks promising. You have put together work from different researchers in a neat way, and made it applicable. I absolutely think that the framework, at least the general part of it, is applicable to many different domains."

"I find that this work you are doing is very relevant to many domains and I can see it being applied to the maritime domain with a few modifications."

"I find your approach very relevant and interesting. It is both thorough and wide, and takes into account the different aspects of error management."

"I think your work looks interesting, and it is very good to know that there are other people working in this same area. Personally, I find the field of recovery/ error management very interesting, and the work you are doing here I think a lot of people will find interesting and useful."

10.4 Lessons learned

Many interesting comments were provided on the basis of this study. In particular, many suggestions were provided in relation to issues that should be taken into consideration if trying to apply the error management framework to other contexts. In relation to the current project and the ATC context some useful comments were made in relation to refining the framework:

- In relation to threat management it was suggested that the taxonomy should not only reflect the anticipation, but also the potential response associated with managing the threat. Therefore, it was decided to include analysis of whether the anticipation was followed by an avoidance response or not.

- Some of the respondents were a bit critical concerning the usefulness and relevance of the detection-stage dimension. This was reflected both in quantitative and qualitative results. Consequently, this is a dimension which relevance should be given some consideration.

- In relation to the PSFs it was suggested that the list was too long and, consequently, some effort to condense the framework should be considered. "Ambient environment" was the main group that received the lowest rating and was also a group that contained a very limited range of categories. Hence, it might not be reasonable to maintain this as a main category.

- Also, it was suggested that not only the influence (positive or negative) of a PSF, but also the performance stage (error, detection and recovery) should be identified

when conducting analyses of the PSFs. These issues will be explored in the next study.

10.5 Conclusion

The purpose of the study reported in this chapter was to get expert input to the framework. In particular, the goal was to get a quantitative evaluation of the relevance of both the components of the framework and its overall structure. Furthermore, more qualitative impressions were elicited insofar as the participants could comment in free-text on strengths and weaknesses of the framework.

An acceptable response rate was obtained. The results on the quantitative part of the questionnaire revealed a high average expert acceptance on all of the main dimensions of the framework (all average ratings were somewhere between 3 and 4 which corresponds to somewhere between relevant and highly relevant). In this manner an overall high degree of face/content validity was obtained.

The qualitative comments from the participants were very interesting in relation to highlighting the extent to which the framework could be applied to other domains – such as the maritime domain and process control - and which issues that should be taken into consideration if trying to adapt the framework (i.e. external validity). For example, the extent to which checklist procedures are an integrated part of a domain's problem solving and recovery should be taken into consideration when trying to apply the framework to other domains.

A criticism that could be raised in relation to this study is that no operational people were involved in the survey. That is, the focus was on human factors experts and not domain experts as such. The reason for this choice is that many of the concepts within the framework are very theoretical and would require some familiarisation before they are properly understood. Furthermore, by using human factors experts from many domains it became possible to get some input in relation to the relevance and applicability of the framework to a number of other safety critical domains.

Another potential criticism is that the respondents in the survey did not try to apply the taxonomy themselves to error events and without this hands-on experience they might not have had a good chance of evaluating the framework. Ideally, it would have been nice if the respondents also had had the chance to get practical experience with the framework before providing feedback. However, due to practical constraints this was not feasible. Nonetheless, on the basis of the constructive and highly relevant comments made by the respondents it seems not to have been a huge obstacle for them. That is, they seemed to be able to provide useful feedback even though they had not tried the framework on real cases. This may be a result of the fact that all of the participants had had previous experience with developing and applying conceptual frameworks related to the area of human error and/or error management.

A more comprehensive version of this study could, in principle, have been carried out where a series of additional questions could have been posed to the participants. In particular, it would have been relevant for each of the dimensions to obtain a rating of its usability, comprehensiveness and diagnosticity. The approach was not chosen in the current context simply because an excessive amount of questions might negatively affect the response rate (and the time required to read through the description of the framework and to answer the questionnaire was already a concern). Furthermore, it was judged that the question of relevance would encompass several of these issues and the free-text answers would give some insight into these issues. As the results also showed this expectation was to a large extent fulfilled.

In sum, the approach taken to obtain input from human factors experts seemed worthwhile and productive. In particular, when having been involved in the development with all of the details of the framework it can be difficult to be effective in evaluating it and by having some fresh eyes to look at it can provide very useful feedback.

11 Study 4 – The critical incident technique

In the previous chapters the error management framework was applied to the analysis of error events in incident reports and in a simulator study. However, it was not a comprehensive version of the framework that was used in these previous studies and there were some potentially important aspects of the error recovery process which constitute an integrated part of the framework that were not examined. This includes the following three issues.

1. One of these issues is the area of threat management. Traditionally, studies of error and error management neglect this aspect of human performance, namely the fact that people can be aware of some of the threats that may lie ahead and might even have some strategies for coping with these threats either before or after they have led to an error.

2. A second issue is the problem-solving or decision-making process underlying error recovery. Just as important as it is to understand the processes lying behind the discovery of a problem is it to know how the chosen resolution came about. This is an important issue insofar as even a timely detection does not guarantee an optimal correction.

3. A final issue is to develop a list of Performance Shaping Factors (PSFs) that can positively or negatively affect the error and error management process. There has been a tendency to view PSFs from a negative point of view (i.e. how they increase the risk of human errors) and to a lesser extent how they also can contribute in a positive way to human performance (i.e. how they can support error recovery).

To be able to evaluate the comprehensive version error management taxonomy it was necessary to obtain some descriptions of authentic episodes where the issue of error management was important. For this purpose it was chosen to use the critical incident technique. Here it is possible to obtain descriptions of potential critical episodes and to elicit information about recovery related aspects of these situations that are rarely described in the incident reports. The limitation of this method is that it will only be possible to obtain direct information from the people being interviewed. So, it is not possible to get the story seen from the perspective of the other ATCOs or pilots involved in the incident.

11.1 The framework

In this study a comprehensive version of the framework will be used. The components of the core of the framework are presented below.

DESCRIPTION OF ERROR AND RECOVERY #1								
"When SAS asks for clearance to flight level 310, R1 could have discovered the risk for a conflict insofar as the strips for SAS and SCW were available. The conflict was detected by a relieving ATCO when the separation standards between the two aircraft were violated."								
THREAT MANAGEMENT								
Threat	Any threats	Yes				No		
	If "Yes":							
	Types(s)	Internal				External		
	Anticipation	Anticipation				No anticipation		
	If "Anticipation":							
	Management	Yes				No		
ERROR								
Error	Producer	ATCO				Pilot		
	Cognitive Domain	Perception	Short-term memory	Long-term memory	Decision	Response		
	If "Decision-making":							
	Procedural violation	Yes				No		
DETECTION & RECOVERY								
Who:	Error/state detector	No one	Producer	ATCO	Instructor	Pilot	System	
	Error/state corrector	No one	Producer	ATCO	Instructor	Pilot	System	
When:	*If detector <> "No one" or "System":*							
	Time of detection	Planning		Execution		Outcome		
How:	Detection cue(s)	External communication		System feedback		Internal feedback		
	If corrector <> "No one" or "System":							
	Decision-making	Ignore	Apply rule	Choose option	Create solution			
RESPONSE & OUTCOME								
What:	Error/State Response	Trap/mitigate		Exacerbate		Fail to respond		
	Error/State Outcomes	Inconsequential/recovery		Undesired state		Additional error		

Table 10: The analysis framework (study 4)

Please notice that a few additions have been made to the original framework in relation to threat management. The threat management types have been expanded so there is a distinction between anticipation of threat without any associated initiatives to counteract the threat and management where a threat is both anticipated and initiatives are made to counteract it (without necessarily being effective). Another addition is the threat types where we distinguish between internal and external threats. Internal threats are threats that originate within the ATC environment (e.g. an inexperienced ATCO) and external threats originate outside the ATC environment (e.g. a pilot that does not comply with an instruction or weather problems). This dimension was added as an exploratory component of the framework.

Below is shown the PSF-taxonomy:

Performance Shaping Factors		
What was the influence of these factors (positive, negative, both or none)?		
1. Traffic, airport and airspace	*Pos.*	*Neg.*
a. Traffic load/ traffic mix/ R/T workload		
b. Time available and degree of urgency		
c. Call sign similarity		

Performance Shaping Factors		
d. Air space and airport design characteristics		
e. Temporary sector activities – military, parachuting, student pilot		
f. Weather - clear weather, snow/ice/slush, fog/low cloud, thunderstorm, windshear		
g. Other traffic, airport and airspace factors		
2. Procedures and Documentation	*Pos.*	*Neg.*
a. Procedures (availability, compatibility, quality and usability)		
b. Operational materials – checklists/advisory manuals/charts/notices		
c. Regulations and standards		
d. Other procedure and documentation factors		
3. Workplace design, HMI and equipment factors	*Pos.*	*Neg.*
a. Interface properties - radar display		
b. Radar and transponder factors		
c. FPS factors		
d. Communication equipment		
e. Warnings, alarms and automation		
f. Other workplace design, HMI and equipment factors		
4. Training and Experience	*Pos.*	*Neg.*
a. Knowledge/experience		
b. Quality of training		
c. Time since last (re)training in task		
d. Informal work practice		
e. Other training and experience factors		
5. Person Related Factors	*Pos.*	*Neg.*
a. Vigilance (fatigue, boredom, alertness)		
b. Strategies: Risk-assessment/short-cuts		
c. Confidence and trust in self/others		
d. Confidence in equipment and automation		
e. Emotional state (calm, chock, panic, stress)		
f. Other personal factors		
6. Social and Team Factors	*Pos.*	*Neg.*
a. Quality of hand over /take over		
b. Language/phraseology/culture issues		
c. Brevity, timing, accuracy and clarity of communication		
d. Sterility of environment (noise, distraction - supervisors, colleagues, visitors)		
e. Team climate and authority gradient		
f. Monitoring/cross-checking		
g. Verbal statements of plans/challenging plans		
h. Review status/modification of plans		
i. Other social and team factors		
7. Company, Management and Regulatory Factors	*Pos.*	*Neg.*
a. Company/commercial pressure - unsafe ops, failure to correct problems		
b. Regulatory – planning, decision making, feedback		
c. Management/Organisation – planning, decision making, feedback		
d. Organisation of work and responsibilities		
e. Training plan		
f. Personnel selection plan		
g. Supervision		
h. Management attitudes towards human error and safety issues in general		
i. Other organisational factors		

Table 11: Performance Shaping Factors (study 4)

Some minor adjustments have been made in comparison with the original PSF-taxonomy. Some of the comments in study 3 indicated that even though the taxonomy was condensed on the basis of the reviewed taxonomies it might still be too comprehensive. Therefore, efforts were made to bring the amount of categories even further down. In particular, the group category entitled "Ambient environment" – which received the lowest expert rating - has been removed and the "Sterility of environment"-category was moved to "Social and Team factors". Another change made on the basis of comments from study 3 is that for each of the PSF classifications the classifier should determine the performance phase that it affected – that is, the error, the detection or the correction phase (this is not shown in the table above).

11.2 Method

11.2.1 The critical incident technique

The critical incident technique was originally developed by Flanagan (1954) as a systematic effort to gather incidents of effective and ineffective behaviour with respect to a designated activity. Initially the focus was on developing countermeasures to the increasing numbers of aircraft crashes during and after the Second World War, but the technique was soon adapted to many other areas. The data were normally collected by asking people about any critical incident that they had been involved in (or observed) and the circumstances of the incident. On the basis of descriptions of situations in which good or poor performance was revealed it was possible to subtract useful information in relation to both solving practical problems and developing broad psychological principles.

A summary of Flanagan's Critical Incident Technique is:

> *"A set of procedures for collecting direct observations of human behaviour which have special significance and meet systematically defined criteria. An incident is an observable type of human activity which is sufficiently complete in itself to permit inferences and predictions to be made about the person performing the act. To be critical it must be performed in a situation where the purpose or the intent of the act seems fairly clear to the observer and its consequences are sufficiently definite so there is little doubt concerning its effects."*
>
> *(Flanagan, 1954, p. 327)*

The assumptions behind the technique are that the descriptions of performance in specific situations should allow inferences about the competence involved and that the behaviour observed should have a significant contribution, positively and/or negatively, to the general aim of the activity. In other words, the critical incident technique can be seen as a

standard qualitative research methodology to obtain descriptors of human behaviours that make a significant difference (that is, are critical) to outcomes within a defined area of interest.

There are some limitations in Flanagan's definition of the critical incident technique. First of all, in the definition it is stated that method relies on "direct observation" which does not have to be the case. Actually, many critical incident studies rely on retrospective accounts as the main source of information. Another limitation of Flanagan's definition of the critical incident technique is the emphasis upon "behaviours". This is based on the assumption that information contained in critical incident reports should be objectively anchored and potentially verifiable. This view is limited because the underlying mechanisms behind the behaviour may be equally or more important to understand and by focusing exclusively on "behaviours" puts unnecessary constraints on the information that can be elicited from such studies. Instead it is useful to let the respondents reflect on their own strategies and the cognitive bases of their judgement and decision-making.

The critical incident technique has been found very useful in relation to the obtaining information about both self-made and observed errors (see e.g. Jensen, 1997). It is therefore reasonable to try and apply the technique to the area of error management (Van der Schaaf, 1988). A potential advantage of doing this is that, in comparison with e.g. incident reports, the critical incident technique makes it possible to make more detailed inquiries about the specific error management events.

11.2.2 Subjects

In relation to the subjects it seems appropriate to select experienced controllers because they are most likely to have encountered one or several critical situations during their professional career. No specific preferences concerning the operational positions were made. 25 ATCOs were interviewed and all the participants were from Scandinavian countries.

11.2.3 The data material

The 25 ATCOs were interviewed and their narratives were recorded on tape. From these people 43 episodes were elicited. Each of the episodes was after the interview carefully converted into a coherent written description of the event and only information exclusively obtained from the interviewee was included. In some cases the interviewees were contacted again concerning information that was either missing or was unclear.

The focus of the interviews was on so-called critical incidents. These incidents were characterised by the following common features (the first three are associated with the content of the descriptions and the last two the quality of the anecdotes).

- *Potentially undesirable outcome.* A critical incident was defined as an occurrence that might have led (if not discovered in time) or did lead, to an undesirable outcome.
- *Error induced.* The incident must have been caused by an error made by a controller and it had to be clearly preventable
- *Extreme behaviours.* In relation to selecting the situations of relevance an advise from Flanagan was followed to focus on "extreme behaviours", either outstanding effective or ineffective (or a combination) with regards to attaining the general aim of the activity. Such incidents have the advantage of being most memorable.
- *Accuracy.* It had to be described in detail by a person involved in or who had discovered the incident. To avoid problems related to the accuracy of the reporting it was chosen on an a priori basis to only consider reports as accurate if full and precise details could be provided.
- *Recent events.* It is also recommended that the focus should, as far as possible, be on recent incidents that are fresh in mind. This recommendation may not necessarily be compatible with the "extreme behaviour"-criterion (i.e. the most recent incidents may not necessarily be good examples of "extreme behaviour").

11.2.4 Procedure

In the following is described the procedure in this study. As in the previous classification studies there are three main stages of the procedure: Information elicitation, segmentation and classification.

Information elicitation

An interview guide was written with the most essential questions for the ATCOs (see Appendix C). Each interview began with a brief introduction to the interview. In this introduction it was emphasised that we were particularly interested in episodes (1) where the interviewee had played a significant role in relation to the detection and/or recovery of an error and (2) that had occurred within a time window of about one year. The reason for the first desideratum was that the focus of the study was to obtain knowledge about the error management process and it was therefore considered most productive to focus on episodes where the interviewee played a significant role in the recovery so that he/she could elaborate on these aspects. The second desideratum was made because it was expected that the details of recent episodes would be more readily available in memory (and thereby be less vulnerable to distorting and/or forgetting details). It should been emphasised that these were only desiderata and were therefore not rigidly adhered to in all the interviews. The reason for this was that people seemingly varied significantly concerning their ability to recall specific episodes in detail. Consequently, this put some constraints on the requirements that could be made to the episodes reported.

Segmentation

In the current study explicit attempts were made at distinguishing between the segmentation and the classification. Therefore, all of the cases were broken down to a number of error events and for each event narrative information from the incidents was added to the following headings: “Threat & Error”, “Detection”, “Recovery”, “Outcome” (all descriptions were taken verbatim from the case description – for a concrete example see section 6.4). For the segmentation of error events it was often necessary to make a graphical representation of the relationship between error events and recoveries. In addition to the free text descriptions related to the analysis of the error events free text associated with identified PSFs was included separately.

Classification

The procedure for classification of the error events was similar to the one in study 2 and consisted of three sub-phases: (a) *Initial classification* - First all the error events (96) and PSFs (106) were classified twice by the author with one month’s interval and a consensus classification was produced; (b) *Calibration phase* - The second classifier was trained in the use of the taxonomy. In order to achieve reliable and valid categorisations the classifier was given feedback concerning the "correct" classification as a part of the training (based on the author’s classifications) and potential misunderstandings in the use of the taxonomy were calibrated. The first eight scenarios containing 17 error events and 24 PSFs were used for this purpose; (c) *Test phase* - On the basis of the remaining incidents the trained observer was asked to classify the errors and error recoveries (79) as well as the PSFs (82).

A couple of unique issues related to the classification are described below.

- *Procedural violations:* Intentional procedural violations are unique because, in principle, the error perpetrator will already know at the planning stage that an error has been committed. In the current context, it was chosen not to analyse the detection stage and detection cue in such situations because it seems a bit artificial to talk about when, how and by whom the intentional violation was detected.

- *PSFs:* Since the PSFs are not mutually exclusive some decisions need to be made concerning how the classification should be made – in particular, because Kappa analysis is intended to be used on data material with mutually exclusive categories. For this reason it was decided that the classifiers should, as far as possible, try to determine only one category and if several categories seemed applicable then the choices should be ranked in the order that they seemed appropriate. By trying to restrict the choice of PSF to a single option the analysis is “pushed to the limit”, but nonetheless, as previously described, it is desirable if the analysis of PSFs can also be carried out by choosing just one category.

11.3 Results

11.3.1 Reliability analysis

To be able to ascertain whether the proposed conceptual tool is useful in the analysis of error and error management situations it is necessary to test whether it can support reliable and robust analyses. In the following is described the results from the intra- and inter-rater analysis (please note that all "unknown" classifications will not be used in the reliability analysis).

Intra-rater reliability

In the table below are given the Kappa results from the intra-rater analysis. The results are based on 96 identified error events and 106 PSFs.

Threat		
Threat type		
Kappa	0.77	P<0.001
Threat anticipation		
Kappa	1.00	P<0.001
Threat management		
Kappa	0.79	P=0.001
Error		
Error producer		
Kappa	0.94	P<0.001
Cognitive domain		
Kappa	1.00	P<0.001
Procedural violation		
Kappa	0.81	P<0.001
Detection and recovery		
Who – detection		
Kappa	0.90	P<0.001
Who – correction		
Kappa	0.85	P<0.001
When – detection		
Kappa	0.66	P<0.001
How – detection		
Kappa	0.87	P<0.001
How – correction		
Kappa	0.97	P<0.001
Response and outcome		
What – response		
Kappa	0.96	P<0.001
What – outcome		
Kappa	0.79	P<0.001
PSF		
PSF – groups		
Kappa	0.97	P<0.001
PSF – individual		
Kappa	0.88	P<0.001
PSF – influence		
Kappa	0.97	P<0.001
PSF – stage		
Kappa	0.68	P<0.001

Table 12: Intra-rater kappa coefficients and P-values for each of the main dimensions in the framework (study 4)

As can be seen an overall very high level of agreement was achieved across all dimensions. All results are above 0.60 (i.e. a "substantial" level of agreement) and are highly significant. It is, in particular, interesting that a high reliability could be obtained for the threat management categories insofar as these have not be tested for reliability before. Also the high reliability for the PSFs is very promising. This included the main groups of PSFs (PSF-groups), the individual categories at the detailed level (PSF-individual), the positive or negative influence of the PSFs (PSF-influence) and the performance stage that the PSF influenced (PSF-stage). The latter refers to whether the PSF affected the original error or the detection and recovery of the error.

The intra-rater results are very promising for the framework, but should however be taken with some modifications. This is related to the fact the classifier was also the one who originally carried out the interview and the later transcription of the incidents (that is, the author). Consequently, the high degree of familiarity with the stories combined with the high degree of familiarity with the framework might have made it easier to be consistent in the analysis.

Inter-rater reliability

A second rater was given training in using the system. For this purpose the first eight incidents were used (consisting of 17 error events and 24 PSFs) and the remainder of the incidents were used for the inter-rater analysis. In the table below is shown the results of the inter-rater analysis of the critical incidents.

Threat		
Threat type		
Kappa	0.53	P=0,011
Threat anticipation		
Kappa	0.80	P<0.001
Threat management		
Kappa	0.75	P=0.005
Error		
Error producer		
Kappa	0.95	P<0.001
Cognitive domain		
Kappa	0.52	P<0.001
Procedural violation		
Kappa	0.72	P=0.004
Detection and recovery		
Who – detection		
Kappa	0.89	P<0.001
Who – correction		
Kappa	0.54	P<0.001
When – detection		
Kappa	0.27	P=0.119
How – detection		
Kappa	0.71	P<0.001
How – correction		
Kappa	0.60	P<0.001
Response and outcome		
What – response		
Kappa	0.76	P<0.001
What – outcome		
Kappa	0.69	P<0.001
PSF		
PSF – groups		
Kappa	0.64	P<0.001
PSF – individual		
Kappa	0.61	P<0.001
PSF – influence		
Kappa	0.92	P<0.001
PSF – stage		
Kappa	0.67	P<0.001

Table 13: Inter-rater kappa coefficients and P-values for each of the main dimensions in the framework (study 4)

As can be seen all except one of the dimensions achieved an acceptable level of agreement and consequently it can be concluded that the framework allowed relative robust analyses. The results are encouraging insofar as they substantiate that other people can be trained in the system and achieve highly robust results.

The only dimension that produced a Kappa coefficient below 0.40 was the when-dimension associated with the detection stage. One of the reasons for this is that there was in general very little variability within this dimension (almost all detections happened at the outcome stage) and within the few variations the classifications did not correspond very well. Such dimensions are especially troublesome for the Kappa analysis insofar as even minor disagreements on the infrequent categories can have drastical effects on the Kappa-values.

11.3.2 Pattern analysis

The pattern analysis will first focus on the *framework core* and then the *performance shaping factors*. For each of these two parts of the framework both *main effects* and *interactions* will be explored. The same statistical tests that were used in study 1 and 2 will be used in this study.

Framework core

In the chart below is shown the distribution of threat management types for the error events that were preceded with one or several threats, $X^2(2, N=61)=11.25$, $P=0.004$. Please notice that the anticipation category has been divided in two – one is anticipation without an active management (entitled "anticipation") and the other is anticipation with active management (entitled "management"). As can be seen in the chart, many threats were either not anticipated at all or were anticipated but not effectively dealt with. In spite of this it might seem surprising that a significant part of the threats were effectively managed when considering that all the events analysed were "critical" or "potentially critical". The explanation for this is that effective management covers situations where an ATCO was aware of a threat before it developed into an error and that the ATCO did everything possible to counteract the threat or to be ready to respond if an error should occur.

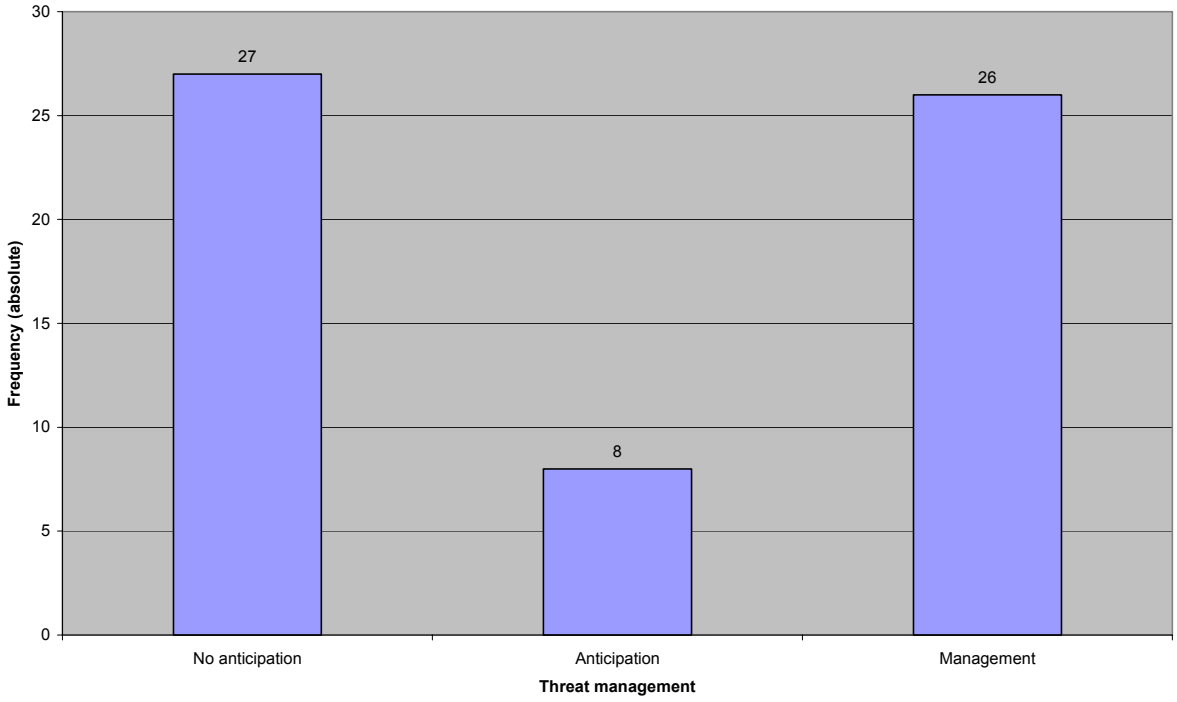


Figure 40: Distribution of threat management types.

In the figure below is shown the distribution of threat types, $X^2(1, N=54)=24.00$, $P<0.001$. As can be seen there is a large majority of external threats compared with internal threats. The internal threats include fatigue, distractions, inexperienced or unskilled ATCO (that e.g. might issue untoward instructions), interface difficulties (e.g. difficulties with distinguishing between fixpoints and other representations on the radar screen). The external threats include problem pilot (that e.g. does not comply with the instructions), traffic congestion, incomplete radar coverage (with the consequence that aircraft under control might not appear on the radar screen), weather (e.g. wind might make it difficult to predict the trajectories of aircraft and bad weather might require aircraft being rerouted), potential conflict between two aircraft (which might be underestimated), VFR pilots (that are not under the control of ATC and might behave unpredicted). Whether the difference in the amount of threat types is a reflection of a reporting bias (i.e. ATCOs being more focused on external problem factors rather than internal) or a reflection of the real distribution of threat types cannot be determined on the basis of the available data.

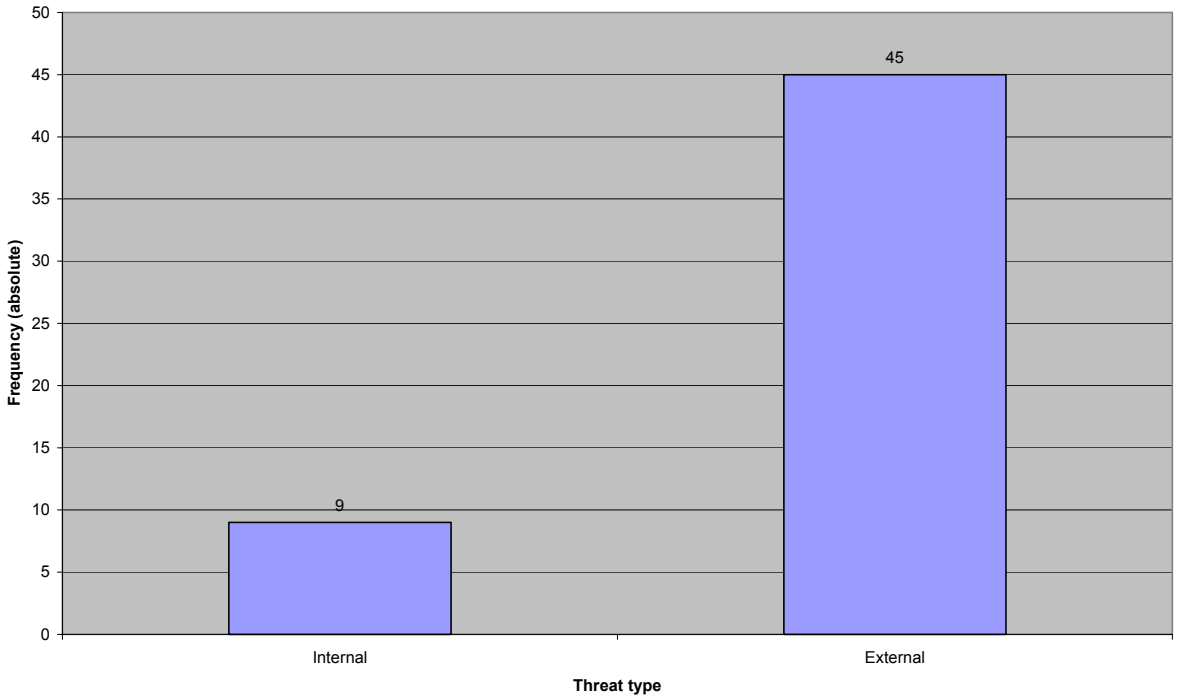


Figure 41: Distribution of threat types

The distribution of error producers is shown in the figure below. The difference between the two groups of producers is not significant, $X^2(1, N=95)=1.78$, $P=0.220$. In other words, there is a comparable amount of pilots and ATCOs who commit errors in the data material.

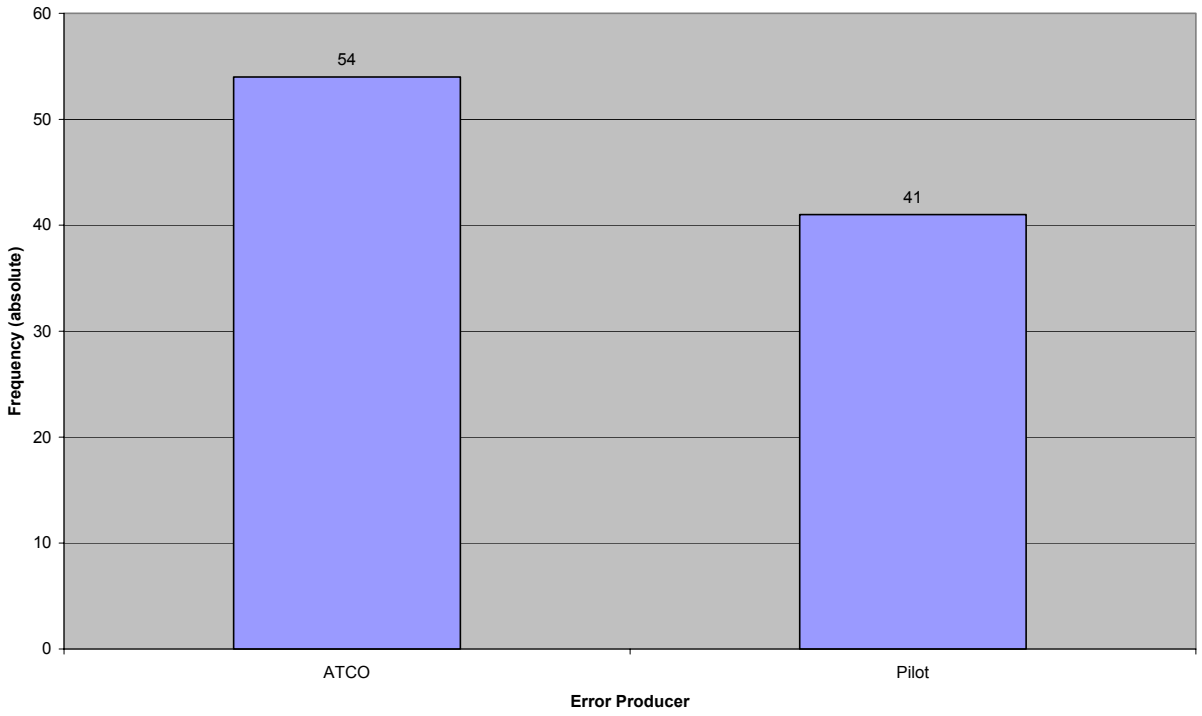


Figure 42: Distribution of error producer

In the chart below is shown the distribution of error types based on the cognitive domain $X^2(4, N=79)=83.34$, $P<0.001$. As can be seen a very large majority of the errors in the critical incidents were decision-making errors. This is in good concordance with other studies that have indicated that decision-making errors are particularly troublesome and are frequently associated with critical situations (e.g. Wiegmann & Shappell, 1997; Klinect et al., 1999). The decision-making errors include procedural violations. 31.3% (i.e. 15 out of 48) of the decision-making errors were procedural violations and 47.9% (i.e. 23 out of 48) were not procedural violations (the remaining 20.8% were unknown).

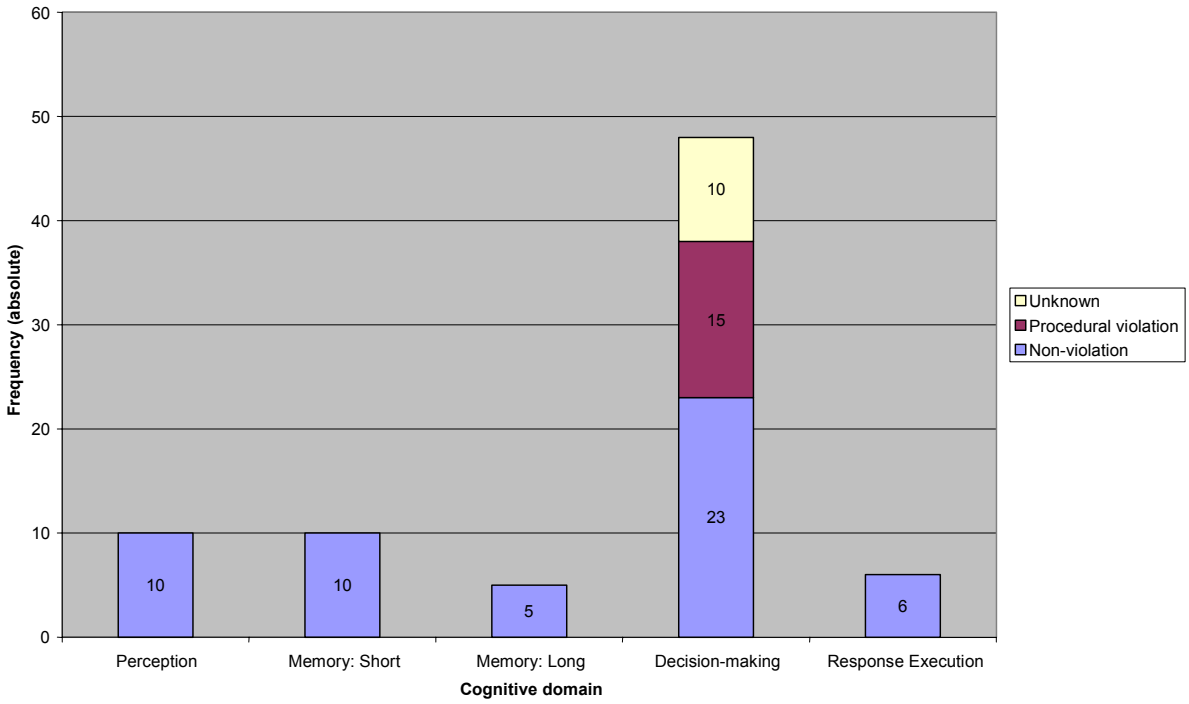


Figure 43: Distribution of cognitive domains.

In the chart below is shown the error detector and corrector (Detector: $X^2(4, N=95)=99.79$, $P<0.001$; Corrector: $X^2(4, N=95)=42.29$, $P<0.001$). Similar to the other studies the results show that in the most cases the error detector is also the corrector. Interestingly, the most frequent detector and corrector is an ATCO different from the error producer. This pattern is different from the one observed in study 1 and 2 where the most frequent detector and corrector was the error producer. The difference in results from these two studies indicates that people prefer to report events where they detected errors committed by other people and is therefore a reflection of reporting bias associated with self-reports (please notice that even though the participants in this study were instructed to report events where they played a significant role in the error detection and/or recovery no instructions were given concerning whether or not they played a role in the error production).

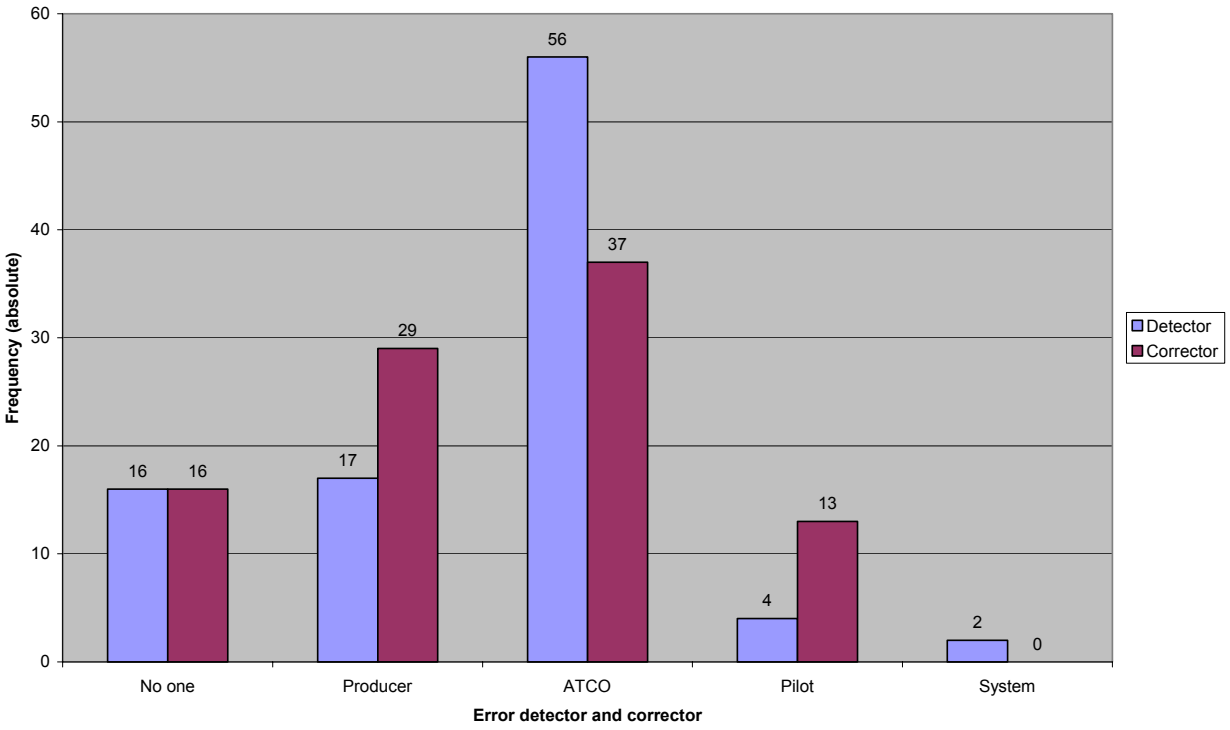


Figure 44: Distribution of error detector and corrector.

In the chart below is shown the distribution of detection source, $X^2(2, N=76)=69.50$, $P<0.001$. Clearly system feedback is the most prevalent source of detection even though a number of errors were detected on the basis of external communication.

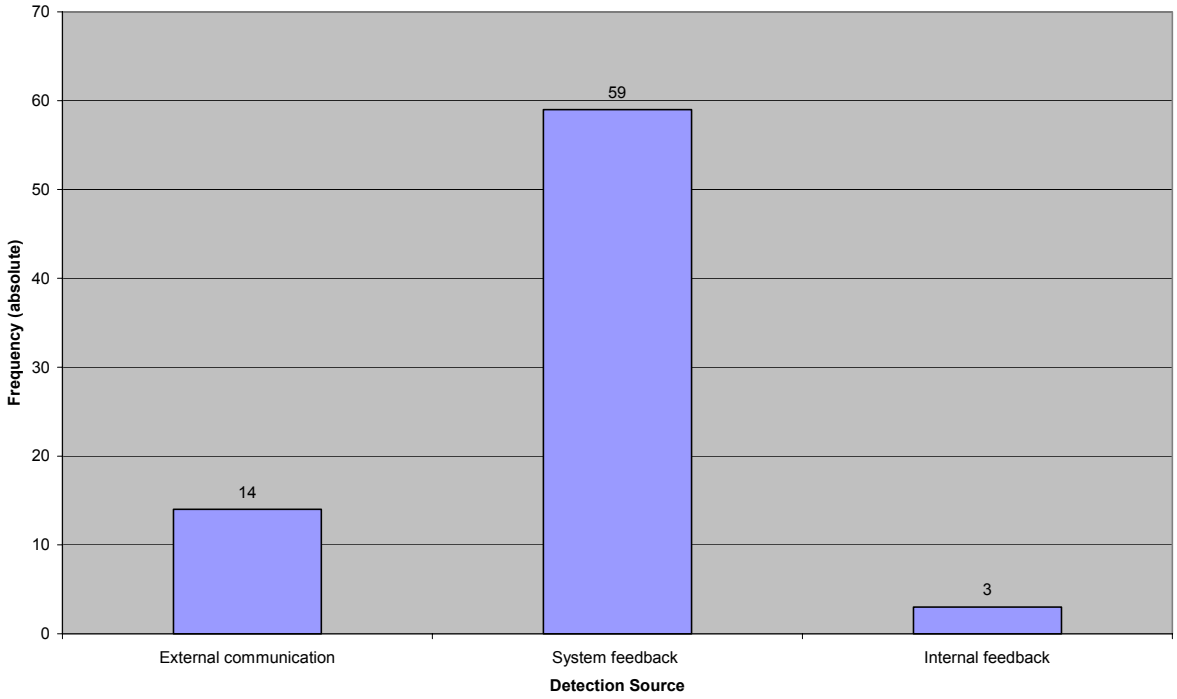


Figure 45: Distribution of detection source

In the chart below is shown the distribution of problem-solving process underlying the correction, $X^2(3, N=63)=55.41$, $P<0.001$. The most frequent type of recovery process is "apply rule". As would be expected, "choose option" and "create solution" are very infrequent (that is, they are associated with more novel and unanticipated situations and should per se be infrequent for experienced controllers).

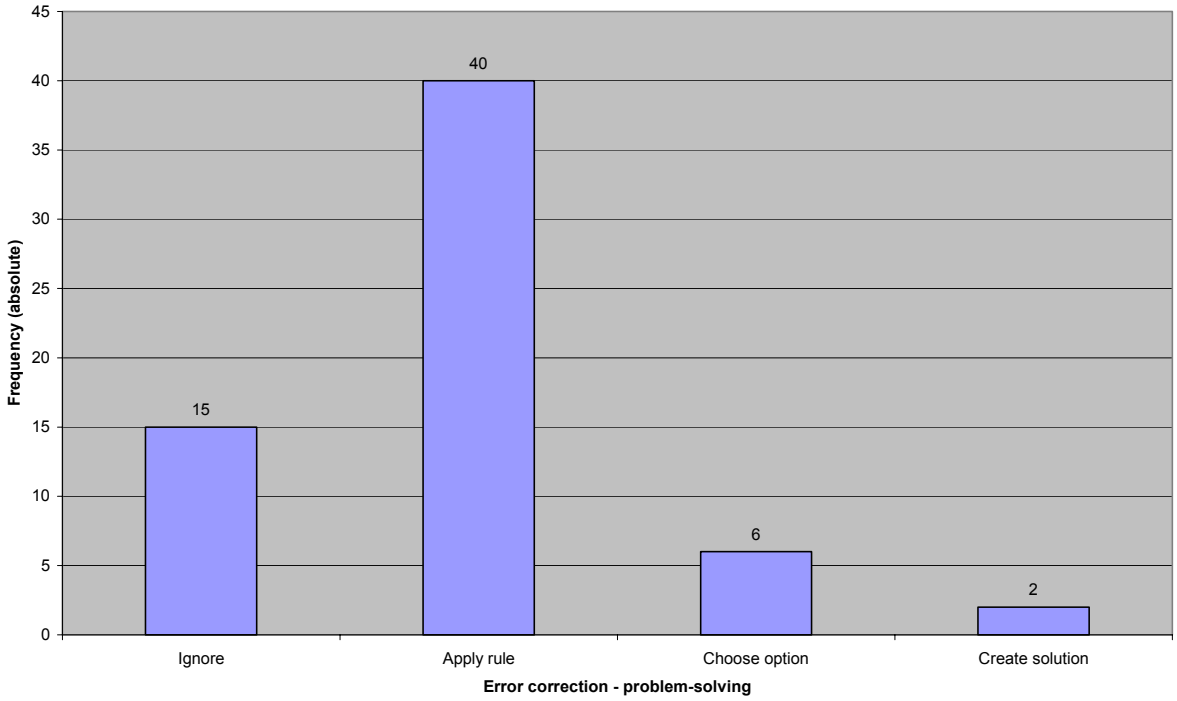


Figure 46: Distribution of error correction – problem-solving.

In the chart below is shown the distribution of responses associated with the correction, $X^2(2, N=95)=31.98$, $P<0.001$. As can be seen a majority of the reported errors are "trapped/mitigated". It is also interesting to note that a larger amount of exacerbations were registered in this study in comparison with the previously described studies.

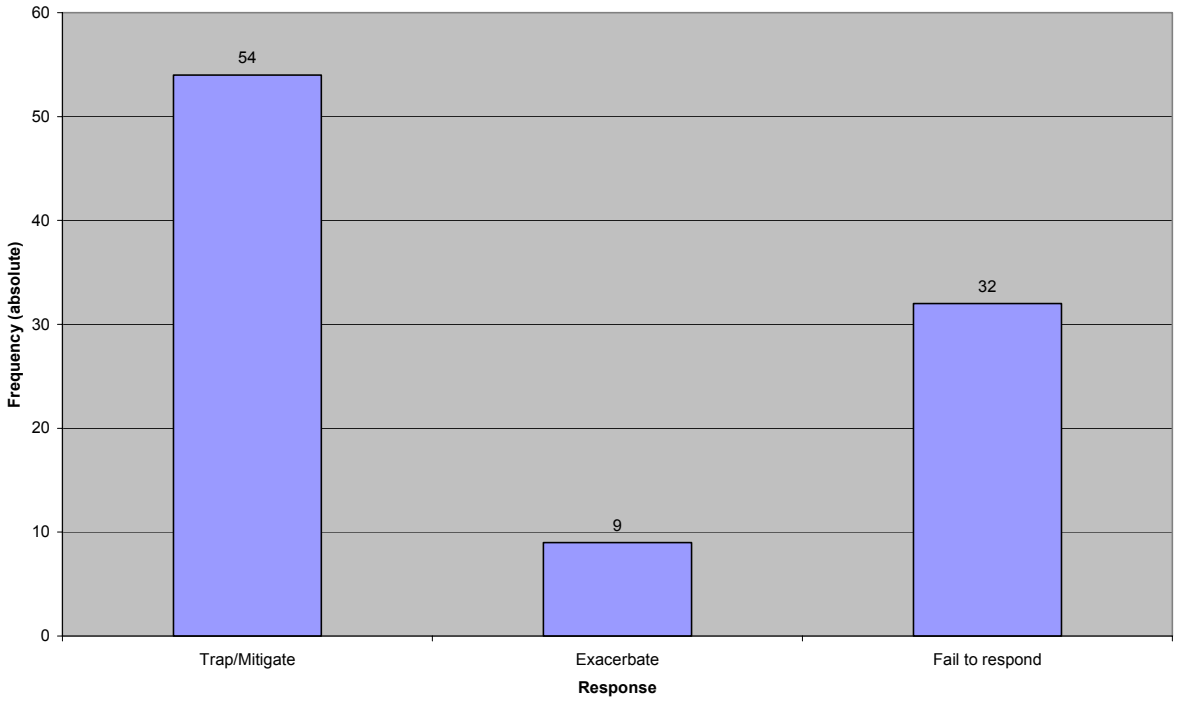


Figure 47: Distribution of response types.

In the chart below is shown the outcome, $X^2(2, N=95)=44.48$, $P<0.001$. Since the current study focuses on critical incidents it could be expected that a majority of the cases would lead to an undesired state.

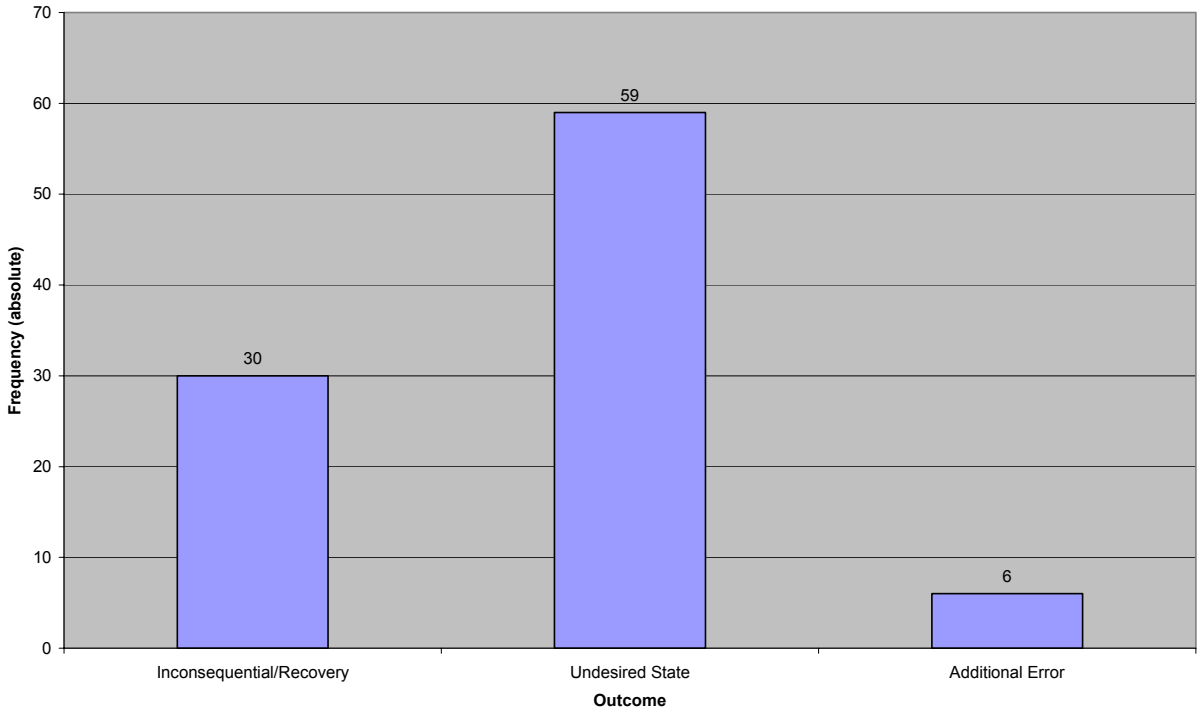


Figure 48: Distribution of outcomes.

Previously it has been mentioned that other studies have indicated that procedural violations are frequently inconsequential and, consequently, it would be expected that the framework could verify this relationship (see hypothesis 13 in section 7.3). In the chart below is shown the interaction between procedural violations (of the sub-group of errors containing decision-making errors) and outcome. The interaction is not significant, $X^2(2, N=39)=4.71$, $P=0.101$. A relatively smaller amount of non-procedural violations led to inconsequential outcomes (AR=-1.98). Conversely, a relatively smaller amount of the procedural violations led to an undesired state (AR=-2.09). It should be noted that the analysis was based on a very small sample (i.e. 39 error events) and that none of the classifiers were domain experts (and, consequently, in a lot of cases had to be omitted from the data material because it was not possible to determine whether or not something was a procedural violation). Hence, the lack of significance may be a result of a type-2 error.

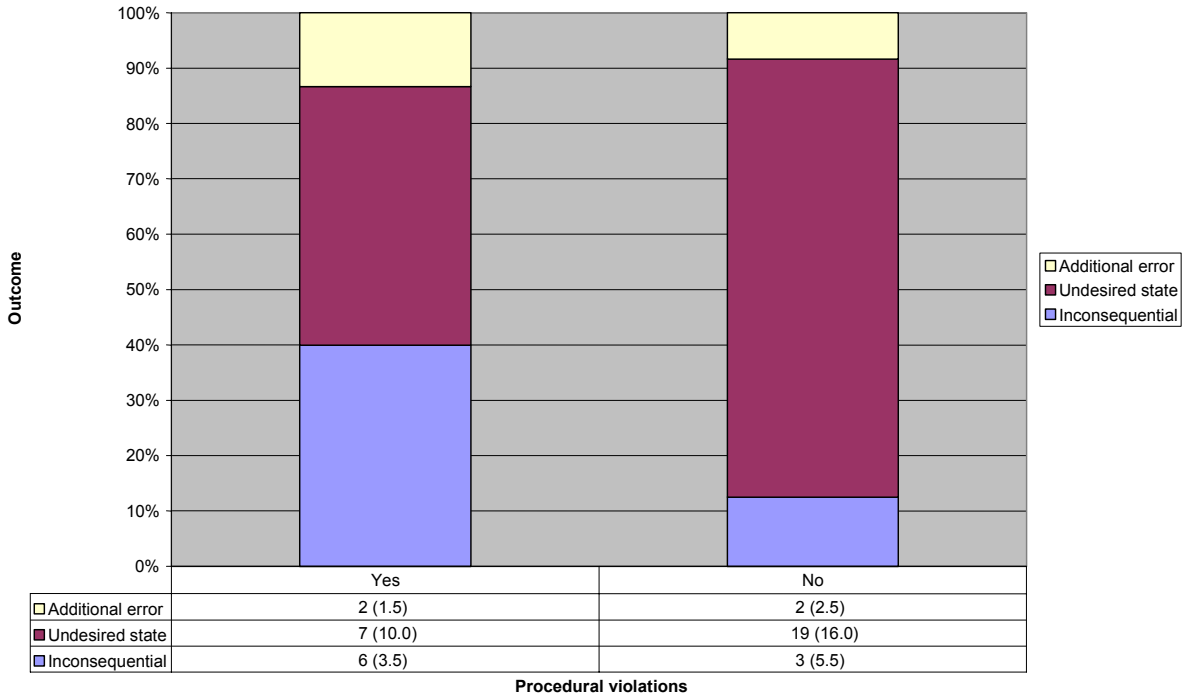


Figure 49: Interaction between procedural violation and outcome.

A related hypothesis is that errors that are ignored and tolerated are frequently inconsequential (see hypothesis 9 in section 7.3). In the figure below is shown the interaction between the decision-making associated with the correction of errors and the outcome (the group "consequential" covers both "undesired state" and "additional error"). The interaction is close to being significant, $X^2(1, N=76)=3.60$, $P=0.067$. The trend is that errors that are ignored (including intentional procedural violations) are frequently associated with inconsequential outcomes (AR=1.90) whereas errors that are responded to are frequently associated with consequences (AR=1.90).

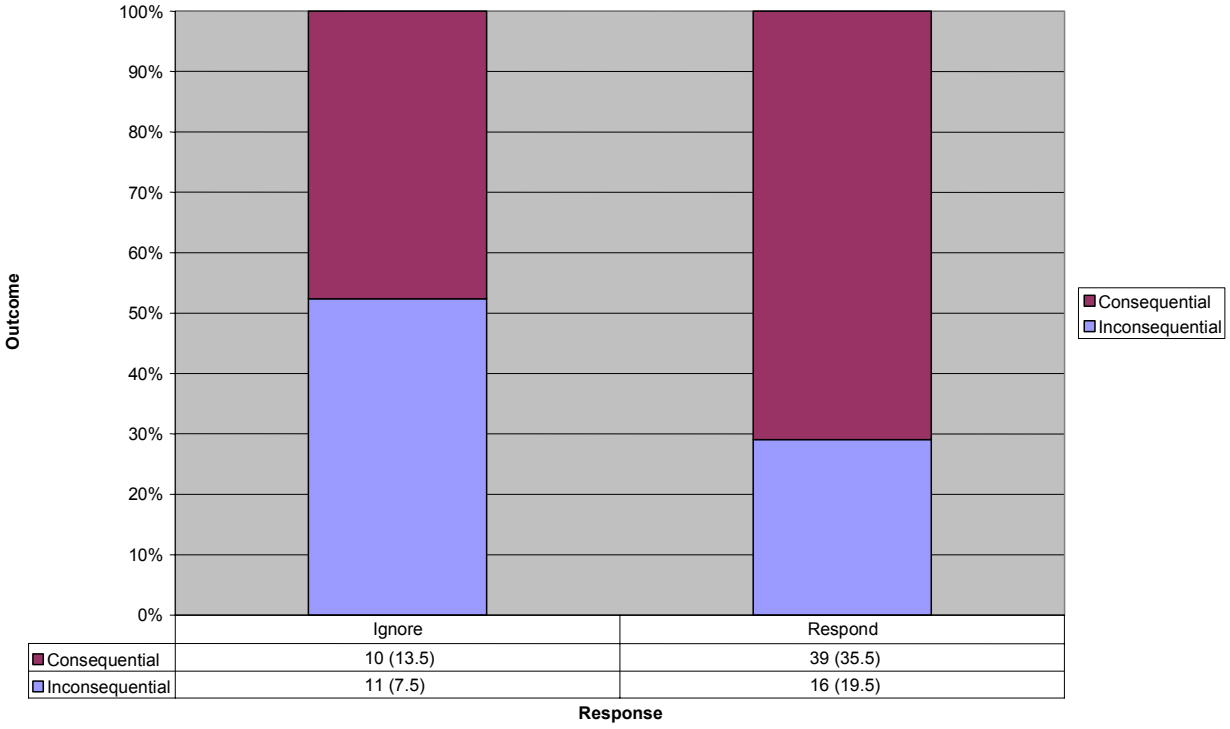


Figure 50: Interaction between ignore/respond and consequential/inconsequential.

Performance Shaping Factors

The overall distribution of the main groups of PSFs is shown below, $X^2(6, N=106)=27.53$, $P<0.001$. As can be seen most of the groups were frequently used throughout the scenarios. The most dominant groups were "Traffic, airport and airspace" and "Training and experience".

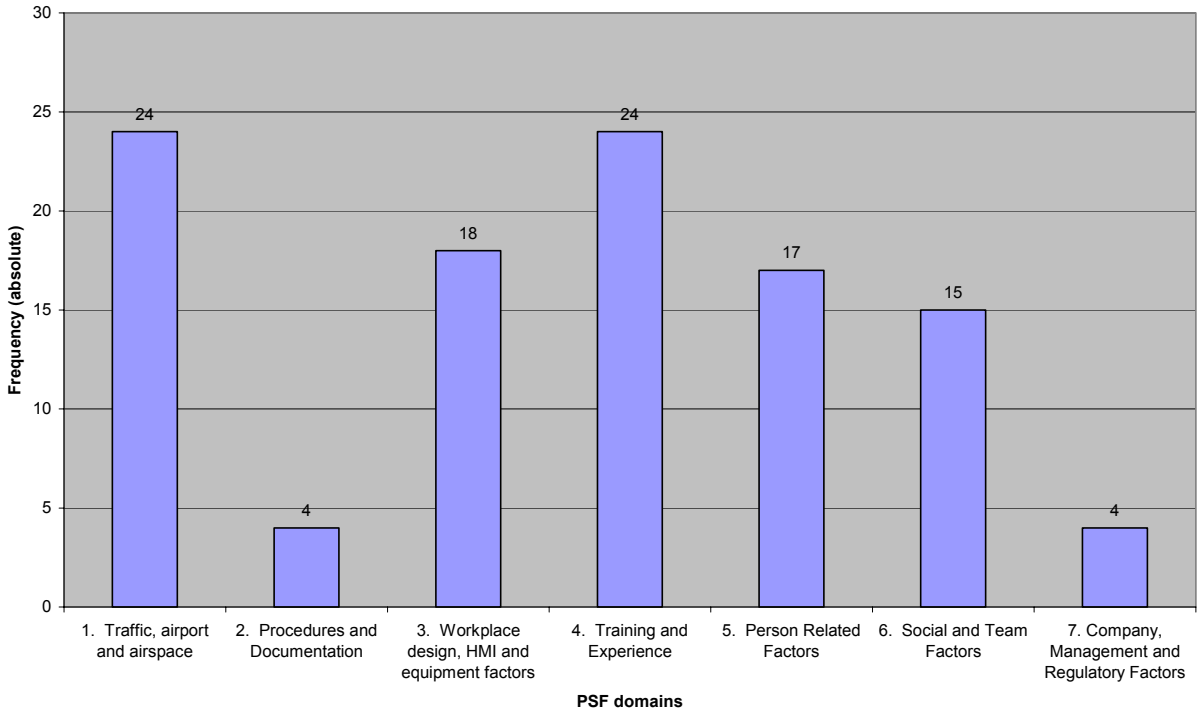


Figure 51: Distribution of main groups of PSFs.

The distribution of positive and negative PSFs is shown in the figure below. As can be seen a very large majority of the PSFs had a negative influence on the course of events (81.0%), $X^2(1, N=106)=36.26$, $P<0.001$. Similar distributions have been observed in several other studies based on the critical incident technique (Van der Schaaf & Kanse, 1999). These results indicate that it might be more difficult to subtract positive factors compared with negative factors in the incident descriptions. In other words, the negative factors are more conspicuous compared with the positive factors. Nonetheless, important lessons might still be obtained from the positive factors in relation to improving the safety barriers within the system.

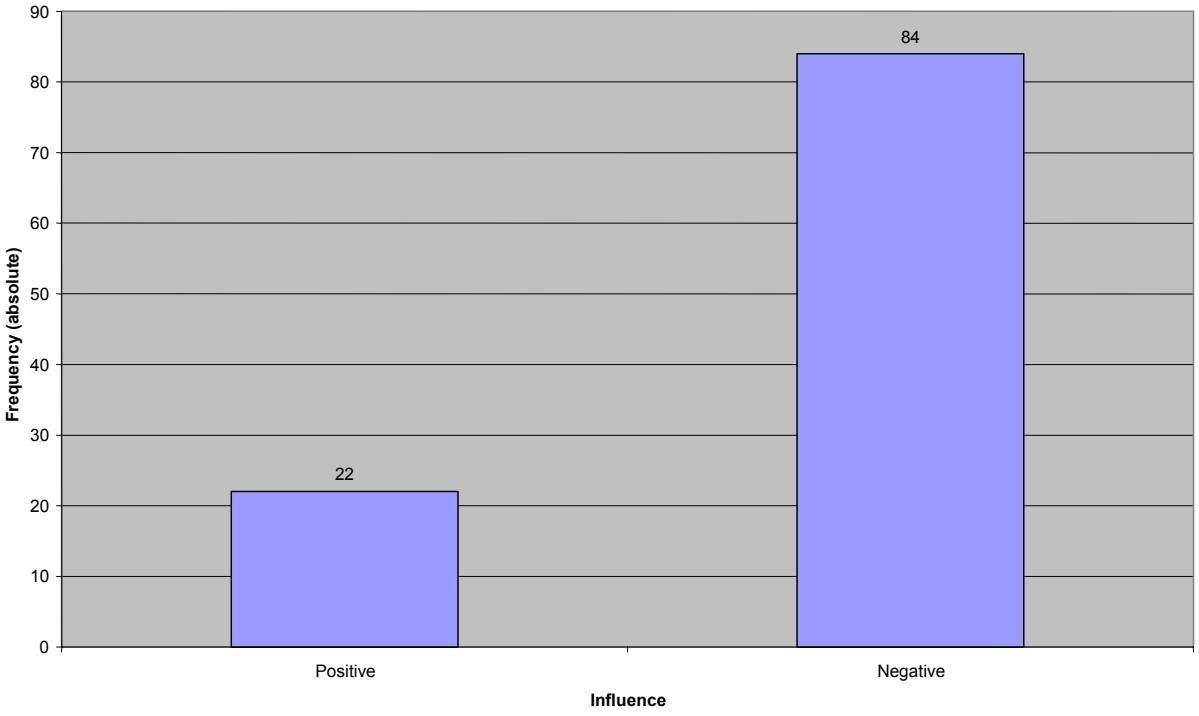


Figure 52: Distribution of positive and negative PSFs

In the figure below is shown the distribution of performance stages that the PSF influenced, $X^2(2, N=106)=3.70$, $P=0.163$. As can be seen a majority of the PSFs had a positive or negative influence on the error management stages, namely detection and correction (58.5%). The variance in the distribution is not significant and it can therefore be concluded that there is a comparable amount of PSFs within each of these three performance stages. In other words, the PSFs seem to cover factors relevant for all three performance stages.

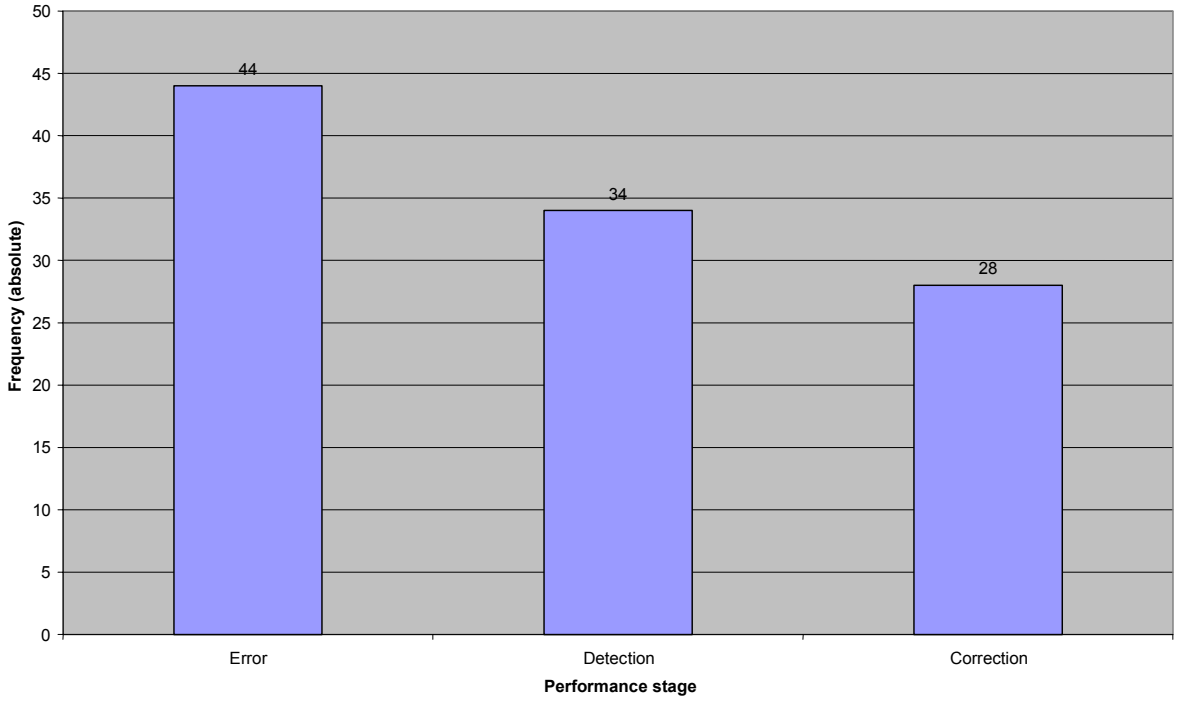
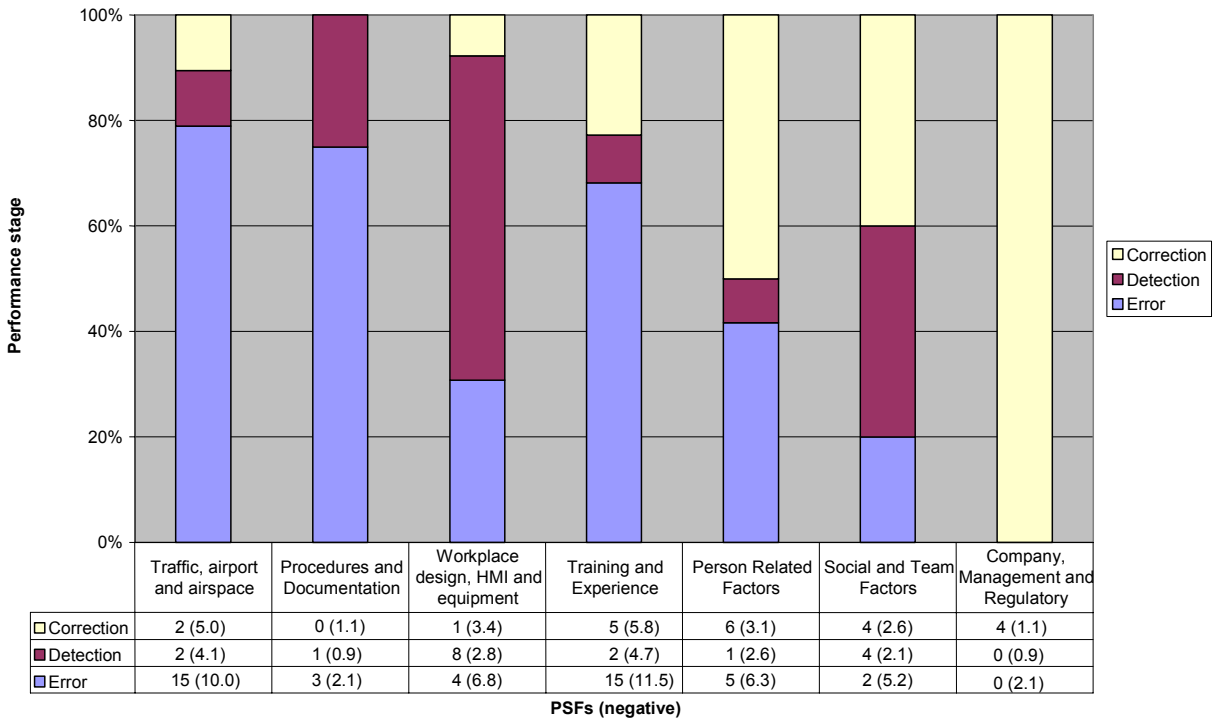


Figure 53: Distribution of PSF performance stages

There was not found any interaction between the PSFs and their positive or negative influence, $X^2(6, N=106)=7.10$, $P=0.311$. However, overall there is an interaction between main groups of PSFs and the performance stage that they influenced, $X^2(12, N=106)=36.52$, $P<0.001$. A more detailed analysis reveals that the there is no interaction when focusing on the PSFs which have a positive influence on the error management, $X^2(4, N=22)=3.35$, $P<0.669$. This might, partially, be related to the fact that the positive PSFs are far less prevalent than the negative PSFs – that is, the sample is too small. The main contribution to the interaction seems to be the negative factors, $X^2(12, N=84)=41.62$, $P<0.001$. In the chart below is shown the distribution of negative PSFs and the performance stage that they affect.



	Traffic, airport and airspace	Procedures and Documentation	Workplace design, HMI and equipment	Training and Experience	Person Related Factors	Social and Team Factors	Company, Management and Regulatory
Correction	2 (5.0)	0 (1.1)	1 (3.4)	5 (5.8)	6 (3.1)	4 (2.6)	4 (1.1)
Detection	2 (4.1)	1 (0.9)	8 (2.8)	2 (4.7)	1 (2.6)	4 (2.1)	0 (0.9)
Error	15 (10.0)	3 (2.1)	4 (6.8)	15 (11.5)	5 (6.3)	2 (5.2)	0 (2.1)

Figure 54: Interaction between negative PSFs and performance stage.

The important contributions to the interaction are elaborated below:

- "Traffic, airport and airspace" was frequently negatively associated with the error production stage (AR=2.64). This interaction is particularly related to the item referred to as "Traffic load/traffic mix/ R/T workload".
- "Workplace design, HMI and equipment factors" were frequently negatively associated with the detection stage (AR=3.83). This is, in particular, related to "Interface properties – radar display".
- "Person related factors" are frequently negatively associated with the correction phase (AR=2.03). This interaction is, in particular, associated with the items referred to as "Strategies: Risk assessments and short-cuts" and "Emotional state".
- "Social and team factors" are rarely negatively associated with the error production stage (AR=-2.18). In other words they are frequently negatively associated with the error management. This is in particular related to the items "Team climate and authority gradient" and "Verbal statements of plans/challenging plans" – in short, classical CRM issues.
- "Company, management and regulatory factors" are frequently negatively associated with the correction phase (AR=3.44). This interaction is caused by the item referred to as "Management attitudes toward human error and safety issues in general". More specifically, these factors are related to covering up the consequences of errors committed (e.g. by not using the term "avoiding action" to avoid that a report will be written).

11.3.3 Validity

Content validity refers to the comprehensiveness of the framework. A way to get an indication of this is by looking at the amount of "Unknown" classifications within the individual dimensions of the framework. The amount of "Unknown" classifications in relation to the core of the framework and the PSFs are shown in the table below:

Threat	
Threat type	
3/57	5.3 %
Threat anticipation	
0/61	0.0 %
Threat management	
0/34	0.0 %
Error	
Error producer	
0/95	0.0 %
Cognitive domain	
17/96 (2/81)	17.7 % (2.5 %)
Procedural violations	
12/51	23.5%
Detection and recovery	
Who – detection	
0/95	0.0 %
Who – correction	
0/95	0.0 %
When	
0/77	0.0 %
How – detection	
1/77	1.3 %
How – correction	
16/79 (5/68)	20.3% (7.4 %)
Response and outcome	
What – response	
0/95	0.0 %
What – outcome	
0/95	0.0 %
PSF	
PSF – main groups	
0/106	0.0 %
PSF – individual	
0/106	0.0 %
PSF influence	
0/106	0.0 %
PSF – stage	
0/106	0.0 %

Table 14: Amount of "Unknown" classifications for each dimension (study 4)

The overall picture is that very few unknown classifications were applied for most of the dimensions. However, it should be noted that the relatively high amount of unknown classifications for the cognitive domain and "how-correction" is mainly related to the fact that pilot errors were included in the analysis and, in most cases, the mechanisms behind the error and its recovery could not be determined on the basis of the available information. If the pilot-errors are removed the unknown-rate drops to 2.5%. In similar vein, if the errors corrected by the pilots are removed the amount of unknown classifications for the "how-correction"-dimension drops to 7.4%. Finally, the relative high amount of unknown classifications for the procedural violations is related to the fact that none of the classifiers were ATC domain experts and consequently a conservative approach was taken in deciding whether something was a procedural violation or not.

The detailed distribution of PSFs is shown in the table below (the relative amount of positive and negative factors is shown in the parentheses). Even though the current study was based on a very small sample, it can be seen in the table that both for the main groups and for many of the subgroups a wide range of the categories were used. A conspicuous exception is "Company, Management and Regulatory Factors" where only one subgroup was used.

1. Traffic, airport and airspace	**24 (5/19)**
a) Traffic load/ traffic mix/ R/T workload	7 (0/7)
b) Time available and degree of urgency	4 (1/3)
c) Call sign similarity	0 (0/0)
d) Air space and airport design characteristics	2 (0/2)
e) Temporary sector activities – military, parachuting, student pilot	0 (0/0)
f) Weather - clear weather, snow/ice/slush, fog/low cloud, thunderstorm, windshear	11 (4/7)
g) Other traffic, airport and airspace factors	0 (0/0)
2. Procedures and Documentation	**4 (0/4)**
a) Procedures (availability, compatibility, quality and usability)	4 (0/4)
b) Operational materials – checklists/advisory manuals/charts/notices	0 (0/0)
c) Regulations and standards	0 (0/0)
d) Other procedure and documentation factors	0 (0/0)
3. Workplace design, HMI and equipment factors	**18 (5/13)**
a) Interface properties - Radar display	5 (0/5)
b) Radar and transponder factors	6 (1/5)
c) FPS factors	0 (0/0)
d) Communication equipment	2 (0/2)
e) Warnings, alarms and automation	5 (4/1)
f) Other workplace design, HMI and equipment factors	0 (0/0)
4. Training and Experience	**24 (2/22)**
a) Knowledge/experience	16 (2/14)
b) Quality of training	3 (0/3)
c) Time since last (re)training in task	0 (0/0)
d) Informal work practice	5 (0/5)
e) Other training and experience factors	0 (0/0)
5. Person Related Factors	**17 (5/12)**
a) Vigilance (fatigue, boredom, alertness)	7 (5/2)
b) Strategies: Risk-assessment/short-cuts	4 (0/4)

c) Confidence and trust in self/others	1 (0/1)
d) Confidence in equipment and automation	1 (0/1)
e) Emotional state (calm, chock, panic, stress)	4 (0/4)
f) Other personal factors	0 (0/0)
6. Social and Team Factors	**15 (5/10)**
a) Quality of hand over /take over	0 (0/0)
b) Language/phraseology/culture issues	1 (0/1)
c) Brevity, timing, accuracy and clarity of communication	1 (0/1)
d) Sterility of environment (noise, distraction - supervisors, colleagues, visitors)	1 (0/1)
e) Team climate and authority gradient	4 (1/3)
f) Monitoring/cross-checking	3 (3/0)
g) Verbal statements of plans/challenging plans	4 (1/3)
h) Review status/modification of plans	1 (0/1)
i) Other social and team factors	0 (0/0)
7. Company, Management and Regulatory Factors	**4 (0/4)**
a) Company/commercial pressure - unsafe ops, failure to correct problems	0 (0/0)
b) Regulatory – planning, decision making, feedback	0 (0/0)
c) Management/Organisation - planning, decision making, feedback	0 (0/0)
d) Organisation of work and responsibilities	0 (0/0)
e) Training plan	0 (0/0)
f) Personnel selection plan	0 (0/0)
g) Supervision	0 (0/0)
h) Management attitudes towards human error and safety issues in general	4 (0/4)
i) Other organisational factors	0 (0/0)
PSFs Unknown	**0 (0/0)**
Unknown	0 (0/0)

Table 15: Distribution of PSF categories (study 4).

11.4 Conclusion

The critical incident technique was chosen as a method in this final study as a means to evaluate the comprehensive version of the framework. This approach was considered useful insofar as it would allow examining parts of the framework that are rarely available when using traditional sources of information (such as incident reports) and that requires having access to the underlying cognitive processes of the subjects involved in the events – in particular, when it comes to shedding light on the error management process.

In the study all of the dimensions except one produced Kappa-values above 0.50 in both the intra- and inter-rater analysis. Consequently, the comprehensive edition of the framework has been proved applicable to error management analyses. In particular, it was encouraging that the previously non-evaluated dimensions of the framework – namely threat management, problem solving associated with error recovery and the PSFs – all provided robust results.

In relation to the management of threats it was interesting to note that a majority of the threats were either not anticipated or anticipated but not responded to (in short, ineffective threat management). This is most likely a reflection of the data material used in this study – that is, it can be expected that a significant amount of threats are inadequately dealt with when focusing on critical incidents. Since it can be expected that threats in normal everyday operations would, in general, be more effectively dealt with a more comprehensive study of threat management in ATC would require real-time based observations of normal operational practice.

The analysis of the problem solving underlying the recovery process also produced some interesting results. In particular, it was demonstrated that in a large majority of error situations the underlying decision to an error recovery was either to deliberately ignore it or action was initiated without any consideration of other alternatives (i.e. "apply rule"). Only in rare cases were several options considered (i.e. "choose option) or was it necessary to create an entirely new solution to a problem never encountered before (i.e. "create solution"). These results fit well with the fact that the people who recovered the errors in the critical incidents to a large extent were highly experienced controllers who consequently had a good deal of expertise in dealing with many types of situations.

The fact that the PSFs produced a high degree of reliability was a bit surprising because the categories were not mutually exclusive and, consequently, several categories might be applicable in relation to a single factor. This makes it both difficult to obtain a measurement of reliability and to obtain a high degree of agreement. Nonetheless, a high degree of reliability was obtained both for the overall groups and the more specific subcategories. Furthermore, for the analysis of influence (positive or negative) and performance stage (error, detection and correction) yielded robust results. In this manner it has been demonstrated that the PSFs can be used consistently to the analysis of both error and error management.

The analysis of the critical incidents revealed a large majority of negative PSFs compared with positive PSFs. This is most likely because it is often more difficult to identify positive factors compared with negative factors simply because negative factors are easier to spot whereas positive factors often concern factors taken for granted. An illustrative example could be the readback procedure. This procedure requires that pilots read back the instructions that they have been given by the ATCO. If a read-back is not carried out by the pilot this might be considered a negative factor. If, on the other hand, the read-back procedure is carried out as required most people would be reluctant to describe this as a positive factor. The distribution of positive and negative factors could also be a reflection of the type of data material being used and if normal operations had been studied a more balanced distribution could be expected.

The PSFs are interesting because they directly target areas that should be dealt with to enhance safety. However, a simple summation of the different contextual categories is perhaps not sufficient to determine which areas should receive most attention. When having a large database of PSFs it seems relevant to incorporate some principles for prioritisation of the individual factors to ensure a more goal-directed and effective

improvement of safety. Some kind of weighting of the PSFs on the basis of their relative risk and how easy they can be corrected should be considered when making managerial decisions (e.g. the two dimensions can be used to create a risk-matrix, as suggested by researchers at the Human Factors Research Project at University of Texas). Such estimations should be done by or in close cooperation with domain experts. In this way factors that score high on risk and high on easy correction should be the ones that should be given highest priority when planning safety enhancing strategies.

A factor that could have improved the accuracy of the classifications in this study is to have had domain experts involved in the analysis. This is, for example, the case for the classification of threats and procedural violations. In relation to threats the involvement of subject matter experts could be relevant, because without a very thorough understanding of the domain it can be difficult to determine whether an ATCO should have been attentive to a certain factor in the internal or external environment. Without such experience and knowledge it is hard to tell whether the ATCO could reasonable be expected to have known about the threat or could have dealt with it in another way. In similar vein, whether some action or inaction is a procedural violation requires a comprehensive knowledge about both regional and international procedures. In short, the validity of the analysis would benefit from having domain experts involved in all stages of the analysis.

12 Evaluation of framework

The previous studies have been useful to shed some light on the utility of the conceptual framework. It is now reasonable to return to the product criteria described in the beginning of this thesis - namely reliability, comprehensiveness, diagnosticity and usability - and take a look at how it satisfies these. Since the whole literature review as well as the development of the framework has been focused on these issues, this will to some extent be a summary of the previously described issues.

12.1 Reliability

A critical measure of the utility of the framework was the intra- and inter-rater reliability of the classifications in the three studies based on incidents reports, a simulator study and interviews based on the critical incident technique. The reason why it is so important to obtain a satisfactory level of reliability is that if it were not possible to obtain reliable results the whole foundation of the framework as a scientific tool would be undermined. Kappa-results for all dimensions from these three studies are shown in the table below.

		Empirical studies				
		Incidents	Simulator study		Critical incident	
Main dimension	Sub-dimension	Inter-rater	Intra-rater	Inter-rater	Intra-rater	Inter-rater
Threat	Type				0.77	0.53
	Anticipation				1.00	0.80
	Management				0.79	0.75
Error	Producer		0.90	0.95	0.94	0.95
	Cognitive domain	0.81	0.86	0.69	1.00	0.52
	Procedural violation				0.81	0.72
Who	Detector	0.64	0.94	0.81	0.90	0.89
	Corrector	0.62	0.88	0.69	0.85	0.54
When	Detection	0.56	0.89	0.60	0.66	0.27
How	Detection	0.62	0.84	0.68	0.87	0.71
	Correction				0.97	0.60
What	Response	0.45	0.94	0.80	0.96	0.76
	Outcome	0.51	0.74	0.50	0.79	0.69
PSF	Groups				0.97	0.64
	Individual				0.88	0.61
	Influence				0.97	0.92
	Stage				0.68	0.67

Table 16: Kappa coefficients from study 1, 2 and 4

As can be seen in the table the overall picture from the empirical studies is that the framework did support highly robust classifications throughout all of the empirical studies. Only once did the reliability results get below the critical 0.40 cut-off point, namely for the when-dimension. The lower kappa values for the first study compared with the other two studies is related to several reasons: (1) the limited amount of information concerning error management in the incident reports; (2) it was possible to provide the second classifier with more training in the last two studies; (3) in the last study explicit efforts were made in relation to separate the segmentation phase from the classification phase – this was not done in study 1 (and was not necessary in study 2).

12.2 Comprehensiveness

Comprehensiveness is related to the extent to which the framework is able to cover all the main categories and issues associated with the area of error management. A way to determine the extent to which the categories adequately reflect the natural variation of a phenomenon is to examine the amount of unknown classifications used within the individual dimensions. A summary of unknown classifications in the simulator study and the critical incident study is shown below.

		Simulator study	Critical incident
Main dimension	Sub-dimension	Unknown classifications	Unknown classifications
Threat	Type		5.3 %
	Anticipation		0.0 %
	Management		0.0 %
Error	Producer	0.0 %	0.0 %
	Cognitive domain	4.4 %	17.7 % (2.5 %)
	Procedural violation		23.5%
Who	Detector	0.4 %	0.0 %
	Corrector	0.4 %	0.0 %
When	Detection	1.0 %	0.0 %
How	Detection	3.6 %	1.3%
	Correction		20.3 % (7.4 %)
What	Response	0.0 %	0.0 %
	Outcome	0.0 %	0.0 %
PSF	Groups		0.0 %
	Individual		0.0 %
	Influence		0.0 %
	Stage		0.0 %

Table 17: Amount of "Unknown" classifications in study 2 and 4

As can be seen the existing categories within the framework were able to account for almost all of the error events analysed in the two empirical studies. In those few cases where a relatively high level of unknown classifications was found this was mainly

related to insufficient information available in the data material (in particular from the pilot's perspective). In the case with the high amount of unknown classifications for procedural violations this was basically a reflection of the fact that the classifiers did not possess a sufficient degree of domain knowledge to make the classification in these cases. In the analysis of the PSFs similar positive results were obtained insofar as no unknown categories were used (see Table 15).

Just as important as it is to ensure that the main types of phenomena within the area of interest are covered it is also important to avoid having unnecessary dimensions and categories within the framework. In this thesis the framework has been applied to several different types of data material and the results have revealed that the variation within the individual dimensions is to some extent dependent on the type of data material being used. For example, it is clear that within the dimensions of detection stage and source little variation was found in the incident study (study 1) and the critical incident study (study 4). In these cases a large amount of errors are detected at the outcome stage and on the basis of system feedback. However, these two dimensions showed a much more varied pattern in the study based on simulator data (study 2). As a consequence of this, it can be concluded that their relevance is to some extent dependent on which type of data material that is being analysed.

In relation to the PSFs most categories within the different main dimensions were used even though the sample used in study 4 was very limited. This indicated that most of the main dimensions as well as their subcategories seemed relevant for the error management analysis. The only noteworthy exception was "Company, Management and Regulatory Factors" where only one subcategory was used from the extensive list. This might indicate that most of these categories are too abstract to be useful in understanding their effect on concrete errors. In other words, these categories are too far removed from the activities at the front-line to be able to determine a relationship with the concrete errors committed.

Finally, some indication of the comprehensiveness of the framework could also be obtained from the questionnaire study. Even though the issue of comprehensiveness was not explicitly addressed in this study several of the comments from the participants clearly revealed that they felt that the framework widely covered most of the important issues within the area of error management.

12.3 Diagnosticity

There were several requirements in relation to the diagnosticity of the framework. The first of these was that the framework should have a psychological cognitive basis that would allow insight into the underlying mechanisms of error production and recovery. The second was that the framework should adequately encompass contextual factors that influenced the error and error management events. These two issues have been given elaborate attention during the literature review and been incorporated into the framework core and the PSFs, respectively. In addition to these requirements there were a number of

hypotheses that concerned how the framework should "behave". The question was here whether the results from the pattern analysis corresponded with theoretical expectations and previous research. In this context a series of hypotheses were previously formulated:

Error:

- Hypothesis 1: *Long-term memory errors will be more frequent among novices*.

A way to explore this hypothesis is to compare the error distribution from the simulator study (study 2) and the critical incident study (study 4). In the simulator study the ATCOs were trainees whereas the ATCOs in the critical incident studies were experienced controllers. The distribution of errors in the two studies is shown in the chart below (SIM is the simulator study with the novice ATCOs; CIT is the critical incident study with the experienced ATCOs). As can be seen in the chart there is a significant variation of error types in the two studies, $X^2(4, N=391)=84.93$, $P<0.001$. In this context long-term memory errors contribute significantly to the interaction (AR=5.2).

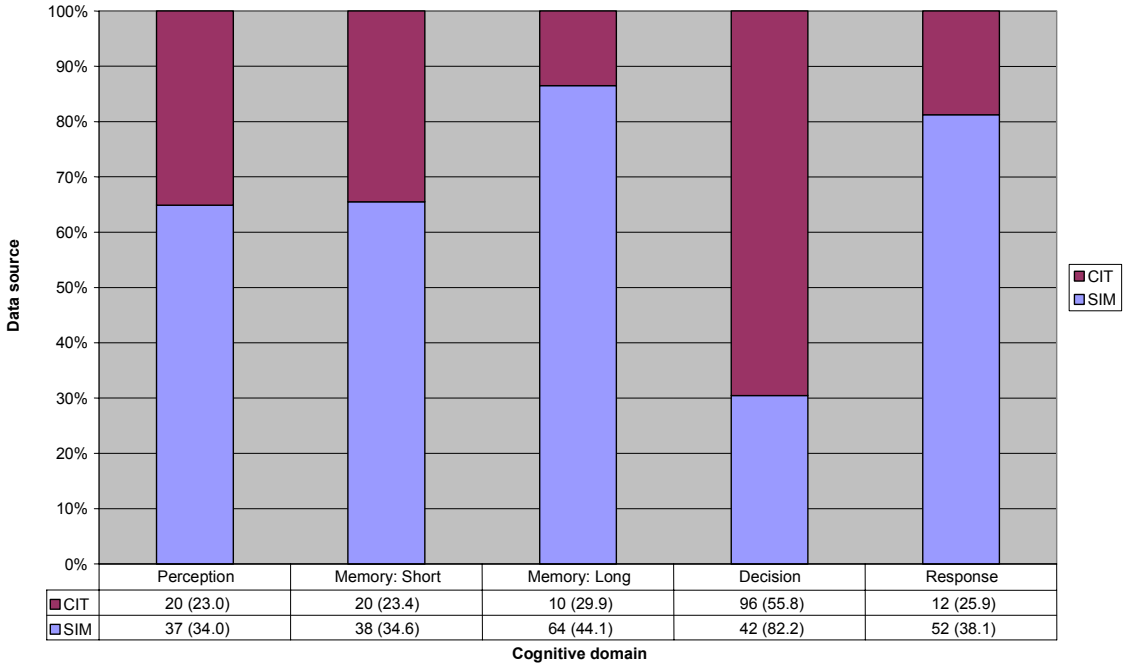


Figure 55: Interaction between cognitive domain and type of data material.

Error and error detector:

- Hypothesis 2: *Response Execution errors are most frequently self-detected*

In study 2 a large amount of response execution errors was committed and consequently this constituted the best foundation for exploring this hypothesis. Here the interaction analysis clearly demonstrated an overall interaction between error and detector (see Figure 33) and, more specifically, that the error producer has a large tendency to detect

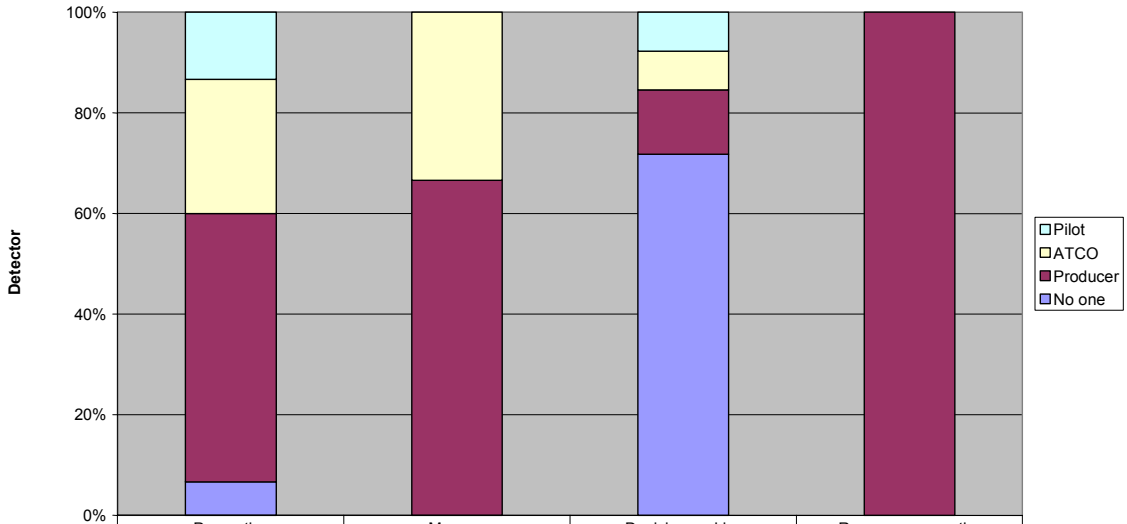
his or her own response execution errors (AR=7.64). Consequently, this hypothesis was confirmed.

- Hypothesis 3: *Decision-making errors are either not detected at all or detected by others*

This hypothesis can be explored by analysing the interaction between cognitive domain and outcome within the three empirical studies.

Study 1

In the chart below is shown the interaction between cognitive domain and detector from study 1, $X^2(9, N=64)=36.12$, $P<0.001$. The figure clearly shows that most frequently decision-making errors are detected by no one (AR=5.32).



	Perception	Memory	Decision-making	Response execution
Pilot	2 (1.2)	0 (0.5)	3 (3.1)	0 (0.3)
ATCO	4 (2.1)	2 (0.8)	3 (5.5)	0 (0.6)
Producer	8 (4.9)	4 (2.0)	5 (12.8)	4 (1.3)
No one	1 (6.8)	0 (2.7)	28 (17.7)	0 (1.8)

Cognitive domain

Figure 56: Interaction between cognitive domain and error detector.

Study 2

In the analysis of the interaction between cognitive domain and detector an overall interaction was found, but no interaction between decision-making errors and no-detection was found (please refer to Figure 33). However, the results from study 2 demonstrated an interaction between decision-making errors and detection by others insofar as decision-making errors were frequently detected by the instructor present in the scenario (AR=2.10) and these were rarely detected by the error producer (AR=-2.57).

Study 4

In this study a large majority of the errors were decision-making errors and, consequently, it was not possible to find an interaction between cognitive domain and detector. However, if we look at the distribution of error detectors associated with decision-making errors in study 4 it is clear that other people are most frequently involved in the detection of decision-making errors – in this case an ATCO colleague, $X^2(4, N=48)=32.83$, $P<0.001$. This might be a result of reporting bias insofar as people might prefer to tell about incidents where they played a positive role rather than a negative role (i.e. the participants reported frequently about episodes where they discovered errors committed by either a colleague or pilot). Nonetheless, the study confirmed together with study 2 that other people might be critical in relation to detecting decision-making errors.

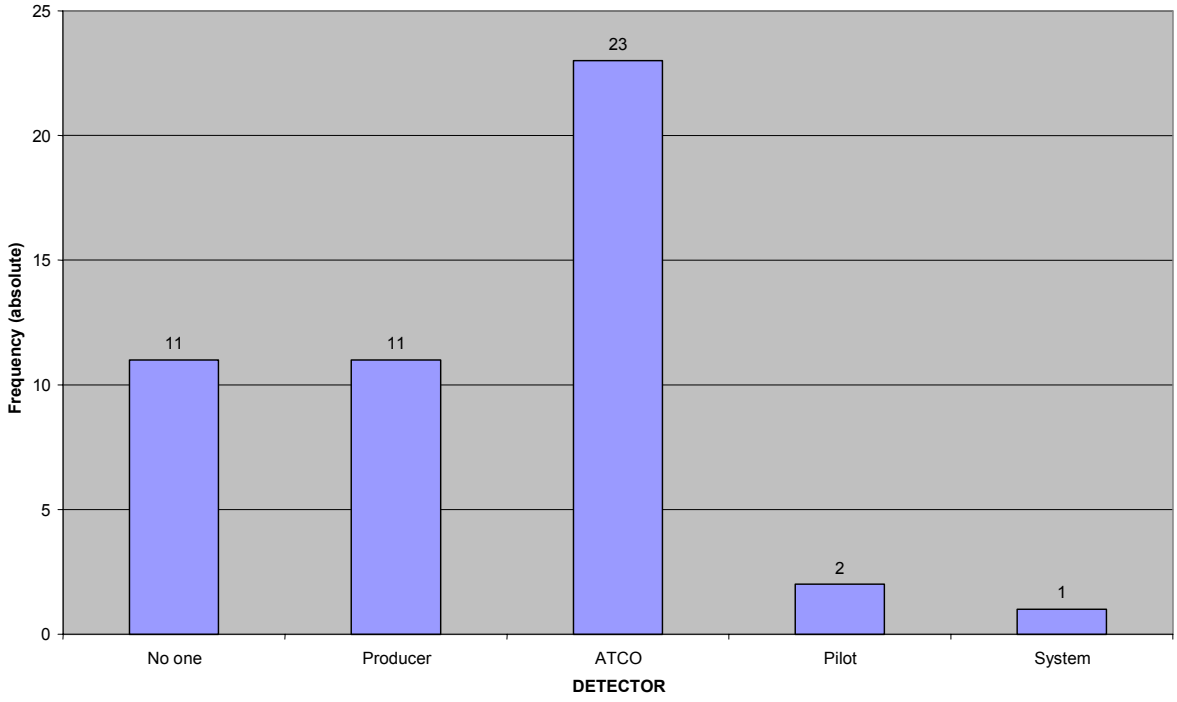


Figure 57: Distribution of error detector of decision-making errors.

In sum, on the basis of these three studies it has been demonstrated that decision-making errors are frequently detected either by no one or by another person than the error perpetrator: Study 1 showed that decision-making errors are frequently not discovered by anyone involved in the error scenarios. On the other hand, the results from the two other studies showed the detection by other people is dependent on the type of context being analysed.

- Hypothesis 4: *Long-term memory errors are either not detected at all or detected by others*.

In study 2 in the analysis of the interaction between cognitive domain and error detector (Figure 33) it was demonstrated that long-term memory errors frequently were not detected by anyone (AR=6.28) or by the instructor (AR=2.80). In this manner both parts of this hypothesis were confirmed in this study.

- Hypothesis 5: *Error detection by others depends on the amount of context-sharing.*

A way to explore this hypothesis is by analysing the errors committed by ATCOs and look at the distribution of errors detected by pilots and ATCO colleagues, respectively. The results from the three empirical studies are shown in the chart below. As can be seen in all three studies there was a clear tendency in the same direction: Errors committed by an ATCO were far more frequently detected by a colleague than a pilot, $X^2(1, N=90)=25.13$, $P<0.001$[18]. This result indicates that context-sharing is an important parameter in relation to detection by others (please note that in study 2 – the simulator study - detection by instructor has not been included so the data only cover the colleague present in the scenario).

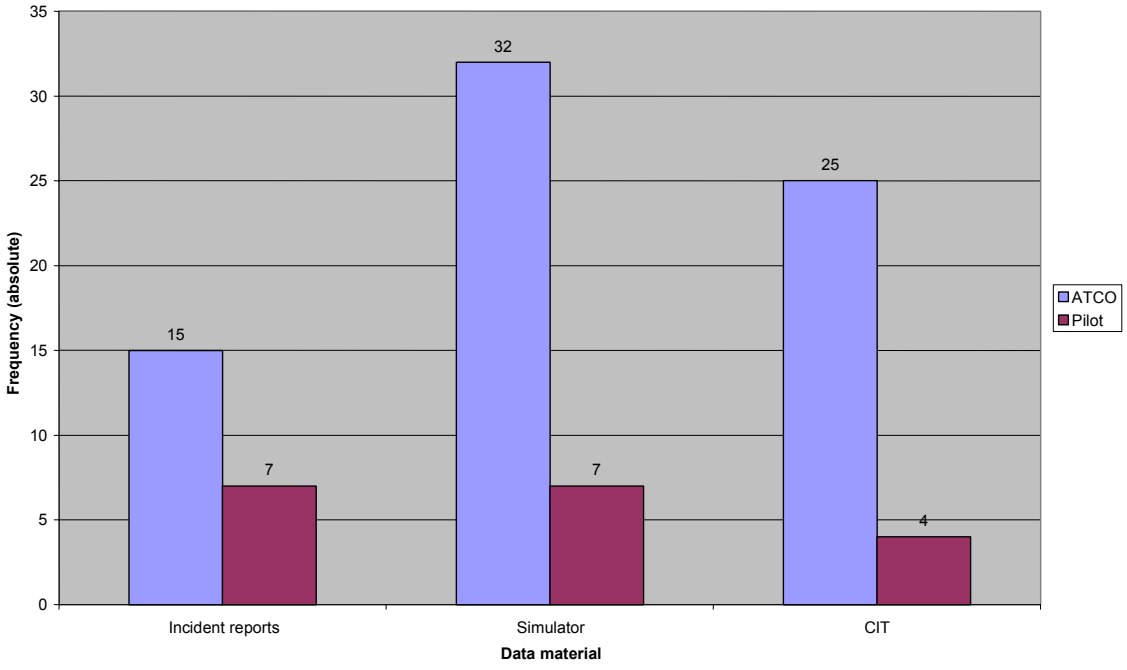


Figure 58: Detector of ATCO errors in the three studies.

[18] This analysis is based on main effect – not interaction effect. It has been done by the use of Wald Statistics.

Errors and Detection stage:

- Hypothesis 6: *Response Execution will be more frequently detected at the execution stage*.

In study 2 it was demonstrated there was an interaction between cognitive domain and detection stage (Figure 33). More specifically, the results showed that there was a significant relationship between response execution errors and detection at the execution stage (AR=7.98). Hence, the hypothesis was confirmed.

- Hypothesis 7: *Errors found in incidents reports will have a tendency to be more frequently detected at the outcome stage compared with errors committed in normal operations*.

A comparison of the detection stage of errors found in the simulator study and the errors found in the critical incident study is shown in the chart below. As can be seen there is a very clear relationship between the detection stage and the two types of data material, $X^2(2, N=346)=87.58$, $P<0.001$. More specifically, detection happened more frequently at the outcome stage in the critical incident study (AR=9.4). The conclusion should, however, be treated with some modification because it was not possible to obtain a satisfactory level of reliability in the classifications of detection stage in study 4 (i.e. the critical incident study).

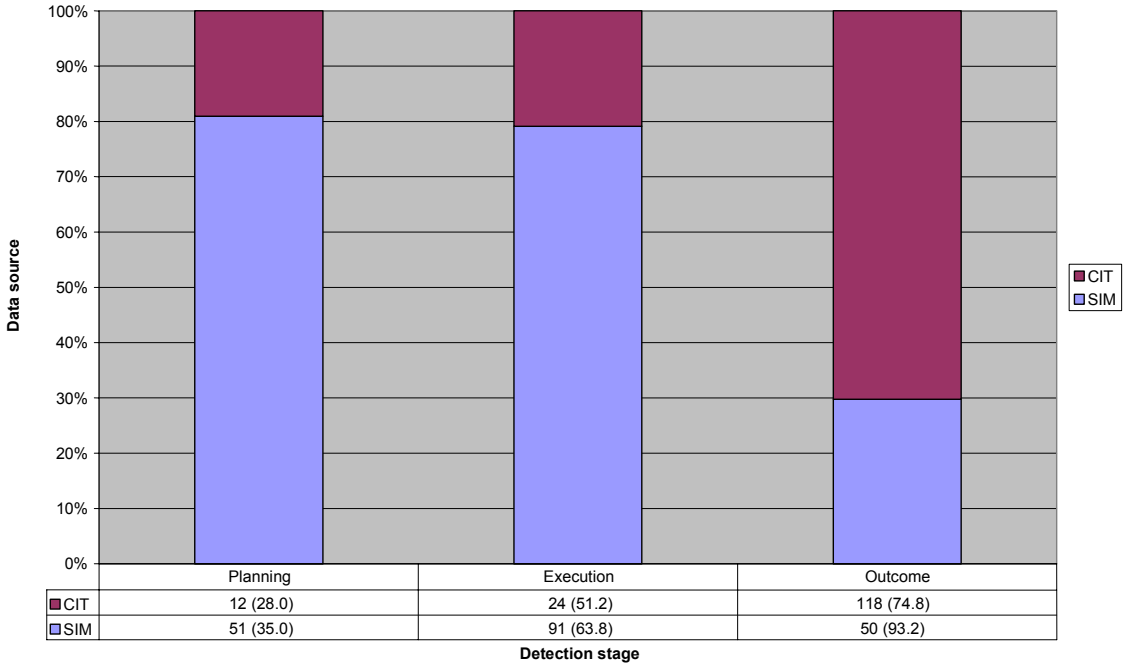


Figure 59: Interaction between detection stage and type of data material.

Error correction and problem-solving:

- Hypothesis 8: *The problem-solving process associated with error recovery will vary in such a way that 'Ignore'/'Apply rule' will be most frequent and 'Choose option'/'Create solution' the least frequent*.

In study 4 the different kinds of problem-solving were explored. Here it was clearly revealed that most error situations were associated with a non-resource demanding problem-solving strategy. The error (or its consequences) was either ignored or could be responded to by applying a straightforward resolution process (Figure 46). Only in rare cases was it necessary to use more resource-intensive processes associated with the 'Choose rule' and 'Create solution' situations.

- Hypothesis 9: *The errors that are ignored are frequently inconsequential*

This hypothesis was based on the expectation that error management to some degree is regulated on the basis of metaknowledge concerning which errors are relevant to recover and which are not relevant. This hypothesis was explored in study 4. More specifically, the interaction between the decision-making associated with the correction of errors (ignore vs. respond) and the outcome (consequential vs. inconsequential) was examined (see Figure 50). The interaction found was close to being significant, $X^2(1, N=76)=3.60$, $P=0.067$. The trend was that errors that are ignored are frequently associated with inconsequential outcomes (AR=1.90) whereas errors that are responded to are frequently associated with consequences (AR=1.90). In sum, even though the hypothesis could not be confirmed the results were close to being significant and in the expected direction. This might be a result of a type-2 error.

Errors and their consequences:

- Hypothesis 10: *Response execution errors (including speech or action errors) should be easier to detect than other errors (lapses and mistakes).*

As can be seen in Figure 55 response execution errors were relatively much more frequent in the simulator scenarios than in the critical incidents (AR=3.9). Consequently, the hypothesis was confirmed.

- Hypothesis 11: *Decision-making errors are more often associated with undesired states*.

This hypothesis can also be examined by looking at Figure 55. Here it can be seen that decision-making errors are much more frequent in the critical incident scenarios than in the simulator scenarios (AR=8.7).

- Hypothesis 12: *Most errors in everyday-life everyday situations will be inconsequential*.

A way to explore this hypothesis is by comparing the outcome of the errors found in the simulator study and the critical incident study. As it can be seen in the figure below there is a clear interaction between the type of data material and the outcome, $X^2(2, N=438)=150.87$, $P<0.001$. In the simulator study a relatively larger amount of inconsequential errors were found (AR=11.9) and, conversely, in the critical incident study a far larger amount of errors leading to an undesired state (AR=12.0).

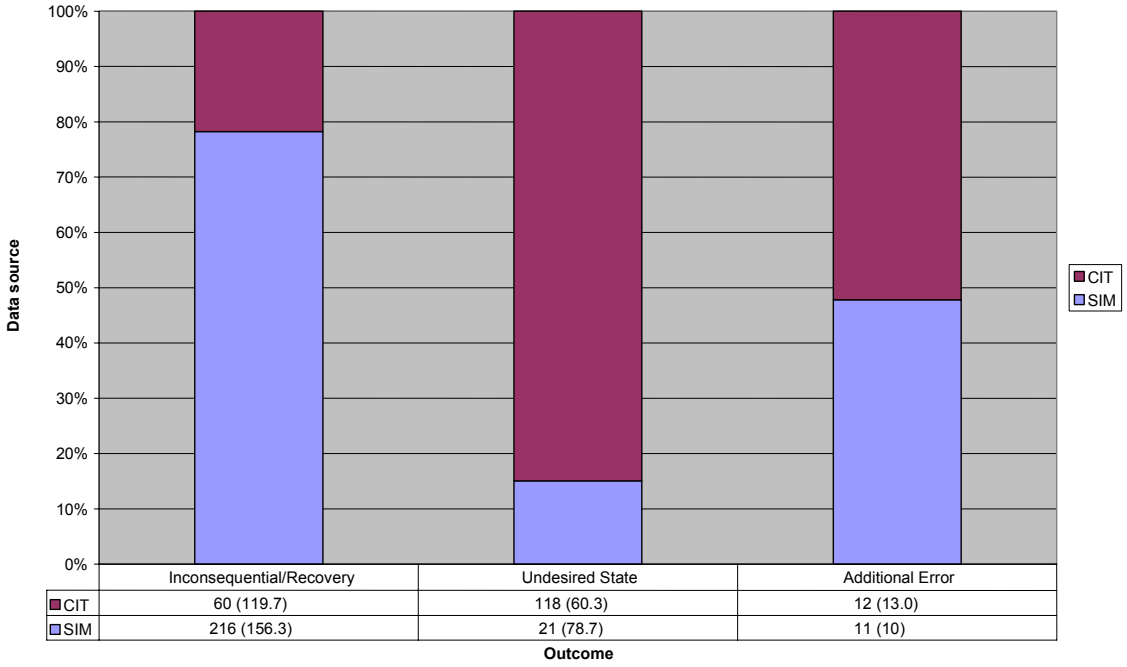


Figure 60: Interaction between outcome and type of data material.

- Hypothesis 13: *Procedural violations will frequently be inconsequential.*

Only in study 4 were procedural violations explored. It was not possible to confirm the hypothesis (see Figure 49). The results revealed that the interaction was not significant ($X^2(2, N=39)=4.71$, $P=0.101$), but this might be the result of a type-2 error. This is supported by the fact that only a very small sample was available in this study (that is, the study was based on a small sample and at the same time many cases were not used in the analysis because they were classified as unknown due to an insufficient degree of domain knowledge of the classifiers). To examine whether this was the case it was decided to identify procedural violations in the incident reports from study one and try to collapse the results from study 1 (Swedish incident reports) and study 4 (Scandinavian Critical Incidents). On this basis a total of 68 decision-making events were available. The results of the interaction between procedural violations and outcome are shown below. The interaction is significant ($X^2(2, N=68)=6.35$, $P=0.037$): Procedural violations have a

tendency to be inconsequential (AR=2.30) and non-procedural violations have a tendency to lead to an undesired state (AR=2.04).

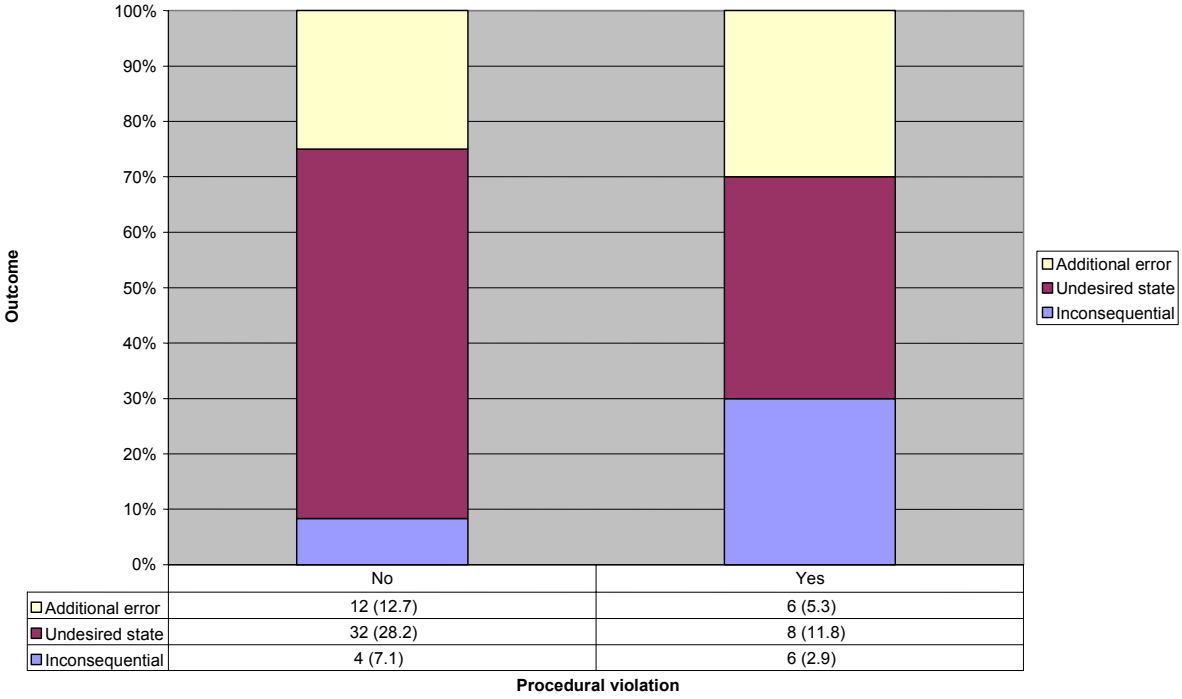


Figure 61: Interaction between procedural violation and outcome (based on study 1 and 4).

As previously suggested, the interpretation of this is that people build up a certain knowledge concerning the relevance and risk associated with different kinds of decisions, including procedural violations. Hence, some types of procedural violations will be considered less dangerous or problematic compared with others (Helmreich et al., 2001). For example, not providing traffic information after a conflict resolution or not using the term "avoiding action" when resolving a conflict will normally be less dangerous violations. On the other hand, if a controller becomes overambitous and give take-off clearances with a less than required separation between the aircraft this could easily lead to a critical situation. Consequently, the above results should not be taken as an argument for not adhering to the procedures and, as can be seen in the figure, a series of procedural violations actually lead to undesired states.

Did the results provide new insights?

The framework has not only been useful in verifying a number of *a priori* defined hypotheses, but also in uncovering previously unknown patterns in the data material from the different studies. In particular, the results from the empirical studies have provided support for the importance of linking the issues of team resource management and error management. On a positive note it was, for example, demonstrated in study 2 that

perception and working memory errors are frequently discovered through the assistance of a colleague. Another interesting result from this study was that decision-making errors are – when they get discovered - frequently detected through external communication. This result indicates the importance of clearly externalising the plans as a critical factor for detecting errors at an early point in time. On a more negative note, it was demonstrated in the analysis of PSFs in study 4 that "social and teams factors" were frequently negatively associated with error management in the critical incidents. In particular, issues such as "team climate and authority gradient" and "verbal statements of plans/challenging plans" seemed to have a negative effect on the error management. Consequently, there is empirical support that error management can be both constructively and adversely affected by the team dynamics. Such results suggest that error management should be the overarching rationale in CRM and TRM training programs (Helmreich & Merritt, 2000).

On a more general level the results from the empirical studies provided support for the notion that human operators are normally very good at detecting and correcting the errors that are committed by themselves or their colleagues (Wioland & Amalberti, 1996). This can, for example, be seen in the figure below that illustrates the distribution of error responses in the three studies[19]. The main effect for response is highly significant, $X^2(4, N=439)=33.35$, $P<0.001$. As can be seen, the amount of errors (or their consequences) that are eventually trapped or mitigated is ranging from 56.8 % (study 4) to 71.6 % (study 2). In other words, a large part of the errors are caught while there is still a chance to do something about it. Since system warnings only had a minimal contribution to the detection and correction of errors or their consequences in all three studies, these results are important to highlight the positive contribution of the human actor in relation to containing the errors that are produced. Hence, it can be concluded that error recovery is "more than sheer luck or coincidence" (Van der Schaaf, 1995).

[19] Please notice that for study 2 the instructor's contribution is included in the tabulation. However, if removing the errors that were detected and/or corrected by the instructor the relative distribution does not change very much: 123 are trapped/mitigated (68%), 0 are exacerbated (0%) and 58 are not responded to (32%).

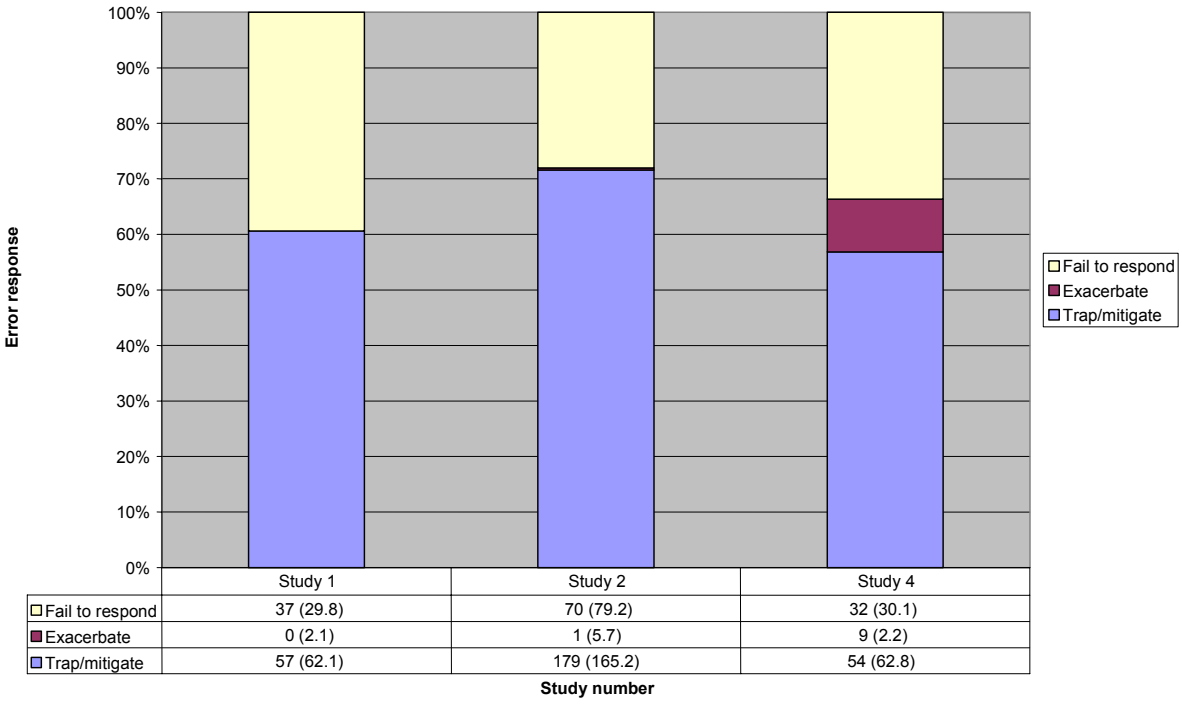


Figure 62: Distribution of responses in the three empirical studies.

Summary

The studies reported in this project have been somehow limited by the fact that only a modest sample of error events was collected from each of these studies. This is a direct consequence of the fact that it was decided to use the framework on a series of different kinds of data material and consequently the amount of data that could be obtained from the individual studies would have to be restricted. In spite of this, the exploration of the a priori defined hypotheses, as well as the new insights generated by the use of the framework, clearly revealed that the framework has a high degree of *diagnosticity*.

12.4 Usability

Even though the framework is comprehensive, attempts were made to ensure that it still maintained its usability. This included both employing categories that were easy to understand and at the same time were practically relevant.

Are the categories easy to understand and use?

Deliberate attempts were made to optimise the usability of the framework. For example, the categories associated with the detection source were chosen because they seemed to make intuitively sense. In principle, it was possible to split this dimension into much finer-grained categories as the ones described in the literature review. However, it was expected that this would jeopardise the reliability and usability of the taxonomy by introducing very subtle distinctions. It was therefore chosen to use a rougher but also easier applicable categorisation. Similar deliberations were made in relation to other dimensions such as the how-correction dimension where the problem-solving terms were chosen on the basis of using terms that are easy to grasp. Even though considerations have been given to develop a framework with a high degree of usability it should be emphasised that it is – as previously mentioned – difficult to develop a framework that displays a high level of diagnosticity and usability at the same time. Therefore, the framework will require some degree of familiarisation and training before it is possible to apply the concepts in a consistent manner.

Are the categories practically relevant?

In the questionnaire study human factors experts were asked to give their opinion about the relevance of the framework. The results revealed that both the overall structure of the framework as well as the individual dimensions received a high level of expert acceptance (all average ratings were somewhere between relevant and highly relevant). Actually, several of the comments also indicated that framework could be relevant in error management studies in other contexts – such as the maritime domain and process control.

12.5 Conclusion

The validity of the framework is largely dependent on whether it would be able to satisfy the defined criteria: reliability, comprehensiveness, diagnosticity and usability. The results from the empirical studies indicate that the framework is able to comply with these four criteria. Consequently, the framework can be said to have demonstrated a certain degree of utility in relation to error management studies.

A couple of minor modifications could be recommended on the basis of the empirical studies:

- First, the relevance of the when-dimension should be considered. In the questionnaire study it received the lowest average rating among the core components of the framework. Also some of the comments from the participants questioned the relevance of this item. Finally, the reliability analysis of this item in the critical incident study revealed a less-than-acceptable level of agreement. These results indicate that the when-dimension should be removed from the framework.

- Second, the relevance of the "Company, Management and Regulatory Factors" group of the PSFs should be considered. Even though this dimension received a high average rating in the questionnaire study, the critical incident study revealed that this group was rarely used in the classifications. Actually, only one subgroup was used in this context. Even though a small database was used in the study the results indicate that this group of PSFs might be of less relevance compared with the other PSF-dimensions in error management studies. This might, in particular, be the case in relation to incident reports where the level of inquiry into the underlying causal factors is less detailed compared with accident reports. In other words, for these organisational factors to be of relevance it would require a very thorough investigation which is normally only seen in accident reports.

13 Summary and conclusion

The purpose of this study was to develop and validate an error management framework. To achieve this goal an extensive literature review was conducted. This included both a review of existing error and error management taxonomies as well as performance shaping factors that may influence the whole process from error production to error recovery. The review of the *error* taxonomies was focused on identifying a relevant framework among the already existing frameworks. The review of the *error management* literature was focused on how it was possible to integrate empirical studies and available taxonomies into a coherent error management framework. Finally, the review of *Performance Shaping Factors* provided a foundation for generating a condensed list of positive and negative factors that could influence the error and error management process.

An error management taxonomy was generated on the basis of the literature review and analysis of different kinds of empirical material. The core of the framework consisted of the following main dimensions: A threat management, an error and an error management section. The latter was designed in such a way that it was possible to answer four main questions: The *who*-, the *how*-, the *when* and the *what*-question. In addition, a list of performance shaping factors specifically addressing the Air Traffic Control environment was presented. In this way it should be possible to answer to the *why*-question. That is, why did the error occur and why was it successfully or unsuccessfully managed.

Four different empirical studies were conducted to validate the framework. First, a pilot study was carried out on the basis of Swedish incident reports to get an initial indication of whether the core of the framework could successfully be applied to the analysis of real complex ATC scenarios. The results indicated that fairly robust analyses could be produced by the use of the conceptual framework. The second study used simulator scenarios from an ATC training curriculum. Here it was possible to analyse error events that in many ways are comparable with those that occur on a normal everyday basis in ATC. The third study was a questionnaire study where different human factors experts provided their opinion about the framework. This feedback provided the means for determining the face and content validity of the framework as well as refining the framework. The final study focused on analysing a series of cases generated on the basis of the critical incident technique. Here it was possible to test out a full-scale version of the framework.

The empirical studies revealed that robust classifications could be obtained by the use of the framework. Both on the basis of intra- and inter-rater analyses a high level of reliability was obtained. In addition, analyses of patterns of both main effects and interactions between dimensions provided interesting insights. In particular, the analysis of interaction between dimensions was useful in supporting the analysis of the criterion validity of the framework. Actually, 12 out of the 13 a priori defined hypotheses - based on theoretical expectations and previous research – were confirmed by using the framework on the data material from the empirical studies. That the framework was able

to support a series of results from numerous other studies within the field of error management is an important quality of the framework because studies within this area have been limited by not having any overarching framework to integrate the different results within. Hence, the results give credence to the utility of the framework. Furthermore, the framework includes several dimensions that have not been adequately explored in the extant literature and might therefore pave the way for conducting new studies of the error management process.

Error management is an issue of high importance in relation to ensuring safety within safety-critical work environments such as air traffic control. Suggestions concerning error management issues that should be given more attention in future research are given in the following.

Combining sources of data

Several different types of data material were used in the current project to validate the framework. It is interesting that the distributions of categories within the individual dimensions seemed to vary dependent on the type of data material. For example, in the critical incident study the pattern of error detection and correction was different from the two other studies. It is likely that this reflects a reporting bias associated with the events elicited through the critical incident technique. Many variations found were also dependent on whether or not the data material was based on observations from everyday error events (e.g. study 2) or based on critical episodes (e.g. study 1 and 4). The consequence of such variations is that conclusions about main effects based on a single type of data material should be treated with some caution. On the other hand, it can be expected that analyses based on several types of data material - and on interaction effects - should be far less vulnerable to potential distortions. Consequently, future studies could benefit from using several types of data material and on uncovering interaction effects within the framework.

Larger database

Some of the aspects of the framework were only covered in the final study based on critical incident reports – namely threat management, recovery related problem-solving and the performance shaping factors (both positive and negative). The results from the classifications revealed that these issues - which have been given insufficient attention in the extant literature - could be applied consistently to real-life cases if sufficient information was elicited. However, a limitation is that only a small database could be generated and, consequently, more extensive studies related to these dimensions and their interaction with other dimensions within the framework might be a useful scientific endeavour. Such studies should preferably focus on both normal operations and on critical incidents.

More comprehensive incident reports

To be able to expand on the knowledge about error management it is clear that incident reports can constitute an important source of information in future studies (see e.g. Van der Schaaf, 1988). However, for these reports to be useful it is desirable that more focus in incident investigations and incident reports is given to the error management part. This was emphasised by the first study based on incidents reports that contained a high level of useful information related to the errors, but little information was available about what occurred after the error – that is, the processes lying behind *how* it was discovered and recovered. To be able to derive useful information about error management it is necessary to increase the awareness of the importance of error management outside the research community. Furthermore, it is important to be able to guide investigators with a conceptual framework – such as the one proposed in this project - that provides them with some guidelines about which issues to expand on.

Error management in the future systems

The air traffic control system is undergoing many challenging changes in the near future that might have implications for both the errors that will occur and the chances of discovering and recovering from these errors. The impetus for these changes grows out of the fact that the system is currently stretched to its capacity limit and rapid increases in traffic levels are envisaged for the near future. To be able to accommodate this development it is necessary to implement new equipment as well as considering new procedures for regulating the air traffic. Some examples of future research areas where the issue of error and error management should be given careful attention are given below:

Datalink. A concept that has been given a lot of attention in recent years is datalink that allows replacing the traditional audio-voice communication between controller and pilot with an automated transfer of digital information (e.g. clearances and weather information). An important advantage of this solution is that it is possible to avoid problems with communication bottlenecks that can cause significant delays. Another benefit is that it is possible to avoid some of the notorious communication breakdowns associated with human perception and working memory (Wickens et al. 1998). This is related to the fact that both the ATCO and the pilot can have all information transferred available on a display and in this manner it is possible to avoid forgetting instructions (which especially is a risk in the case of lengthy sequences of instructions). Datalink may in similar vein be helpful in the detection of errors (e.g. having misunderstood or misread a clearance) insofar as the clearances are available for later reference. In spite of these potential advantages of datalink in relation to error and error management there are at the same time also some risks. Some examples are: (1) It can be more difficult to convey a sense of urgency in the digital communication compared with oral communication and, consequently, some instructions might not be adhered to as fast as they should be; (2) There is a risk of new types of errors – such as keystroke errors or pilots not asking for clarification – which might require unique solutions to ensure that they are trapped before leading to operational problems; (3) The party-line effect is removed insofar as pilots can

no longer listen to the communication between controllers and other pilots on the same frequency to which they are attuned. Consequently, pilots are deprived of a source of information that can be helpful in maintaining and updating their situation awareness – and, in some cases, can be a critical resource in relation to catching errors that can have consequences for the colleagues.

Free flight. Free flight is a concept that has been proposed to enhance pilots' possibilities of determining in real time optimum routes, speeds, and altitudes without being constrained by air traffic control. The idea is that the pilots should be allowed to fly to destinations directly which is far more cost- and time-effective than flying along fixed routes. Even though a high degree of autonomy is transferred to the pilots it is still the intention that the controllers should monitor the flight system and, in the case of safety critical circumstances, to be able to "bail out" pilots from these situations. Several factors might make this difficult for the controllers. First, it might be difficult to obtain the "big picture" if the controller is not actively involved in the process of controlling the aircraft (Willems & Truitt, 1999). Second, the fixed structure - which characterises the current airspace - will be replaced by a more unconstrained structure and the consequence of this is that it becomes more difficult to predict the future status of the aircraft and the chances of determining potential separation problems are thereby reduced (Endsley et al., 1997). Both of these factors may potentially reduce the ATCOs chances of effectively detecting and recovering from pilot-induced critical situations.

Innovative concepts – such as the ones described above – will require changes in procedures, displays and automation and can have significant effects on the task performed by the human controller. They have the potential for both improving safety and efficiency, but at the same time there is a risk of compromising the human operator's chances of ensuring safety. With the advent of new technologies and new operating philosophies there is a risk that new types of errors will emerge and at the same time the chances of recovery might be diminished. Since human errors cannot be completely avoided and some level of system unreliability is inevitable it is of paramount importance for safety to be maintained that the powerful human recovery abilities are not undermined. Hence, the development and evaluation of future initiatives aimed at safely enhancing the capacity of the air traffic system will require careful consideration of error and error management profiles if the strong safety record should be maintained.

14 Dansk resume (Danish summary)

Hovedformålet med denne afhandling er at udvikle, validere og evaluere en fejlhåndteringstaksonomi (”error management taxonomy”), som kan anvendes til at analysere fejlbegivenheder indenfor flyveledelsesområdet. Idéen bag taksonomien er at gøre det muligt at analysere mekanismerne bag menneskelige fejl og deres genoprettelse. På nuværende tidspunkt eksisterer der en mængde taksonomier til at beskrive mekanismerne bag menneskelige fejl, hvorimod der eksisterer kun en begrænset viden omkring mekanismerne bag opdagelsen og genoprettelsen af fejl. Dette er uheldigt da hurtige og effektive indgreb kan ofte forhindre fejl i at have alvorlige konsekvenser for systemsikkerheden. Det forhåndenværende projekt prøver derfor at tilvejebringe mere viden omkring hvordan fejl bliver indfanget. Dette vil være et vigtigt fundament for at kunne udvikle strategier rettet imod at reducere kritiske hændelser. For at dette kan lade sig gøre er det vigtigt at have et struktureret klassifikationssystem, hvor operationelle data om *udførelsen*, *opdagelsen* og *rettelsen* af menneskelige fejl kan blive analyseret inkl. de bagvedliggende omstændigheder for fejlene og deres indfangning.

Rapporten er inddelt i fire dele:

Del 1 – Baggrund. I den første del bliver betydningen af menneskelige fejl og fejlhåndtering indenfor flyveledelse gennemgået. Ligeledes bliver der opstillet nogle generelle krav til en fejlhåndteringstaksonomi. For de læsere som ikke er bekendte med flyveledelsesområdet bliver der givet en kort beskrivelse af området.

Del 2 – Litteraturgennemgang. I den anden del bliver der gennemgået den relevante litteratur for at afgøre, hvilke kategorier som skal inkluderes i fejlhåndteringstaksonomien. Fokus er på taksonomier forbundet med menneskelige fejl såvel som de begivenheder der går forud og efter fejlene. Før fejlens opståen er fokus især på, hvordan potentielt kritiske operationelle faktorer der kan lede til fejl – og måske bringe sikkerheden i fare - bevares under kontrol (”threat management”). I forbindelse med fasen efter fejlens opståen er der især fire emner, som vil blive beskrevet: *hvem* var involveret i opdagelsen og indfangningen af fejlen og/eller dens konsekvenser; *hvornår* blev fejlen eller konsekvenserne opdaget; *hvordan* blev fejlen og/eller dens konsekvenser opdaget og indfanget; og endelig *hvad* var reaktionen og udfaldet? Herudover bør det være muligt at kunne besvare *hvorfor*-spørgsmålet – nemlig hvorfor skete fejlen og hvorfor blev den effektivt eller ineffektivt håndteret? Dette kan bestemmes på basis af såkaldte ”Performance Shaping Factors” (PSFs), som kan ses som kontekstuelle faktorer, der kan have positiv eller negativ indflydelse på begivenhedsforløbet.

Del 3 – Konstruktion af taksonomien. I den tredje del vil fejlhåndteringstaksonomien blive beskrevet. Den er udviklet på basis af litteraturgennemgangen og er desuden blevet tilpasset og testet på basis af hændelsesrapporter, interviews omkring kritiske hændelser og et simulator studie (resultaterne af disse beskrives i den næste del). Systemet er organiseret omkring en fejlhåndteringsmodel. Den består af to hovedkomponenter: Kernen i systemet er udviklet på basis af litteraturgennemgangen af fejl og

fejlhåndteringstaksonomier. Listen over kontekstuelle faktorer er udviklet på basis af gennemgangen af Performance Shaping Factors.

Del 4 – Validering. Nytteværdien af fejlhåndteringstaksonomien i forbindelse med fejlbegivenhedsanalyser vil blive udforsket. Til dette formål er rammeværket blevet evalueret på basis af forskellige typer af datamateriale. For det første er værktøjet blevet anvendt på fejlbegivenheder fundet i kritiske hændelser (både svenske hændelsesrapporter og kritiske hændelser baseret på interviews med flyveledere) og i et simulatorstudie. På dette grundlag har det været muligt at få viden om det er muligt at opnå konsistente klassifikationer (på tværs af tid og personer) og yderligere har det været muligt at udforske mulighederne for at opdage interessante mønstre i disse forskellige typer af datamateriale. Systemet er også blevet evalueret v.h.a. af en række human factors eksperter, der har været involveret i forskning, som er højest relevant i forbindelse med dette projekt. På den måde har det været muligt at få en både kvantitativ og kvalitativ evaluering af systemet.

Resultaterne fra studierne rettet imod at anvende værktøjet indikerer, at det er både muligt at opnå robuste analyser på basis af systemet, og det er muligt at verificere resultater fra andre studier såvel som at tilvejebringe nye indsigter. Yderligere viser resultaterne fra spørgeskemaundersøgelsen, at eksperter mener, at værktøjet er yderst relevant i forbindelse med studier af fejlhåndtering. Kort sagt viser resultaterne, at fejlhåndteringssystemet kan være nyttig i forbindelse med fremtidige fejlhåndteringsstudier. Særligt kunne det være et nyttigt redskab i forbindelse med analyse af virkningerne af en række sikkerhedsinitiativer indenfor flyveledelse – det være sig ændringer i system design, procedurer eller træning af personale.

15 Literature

Agresti, A. (1996). An introduction to categorical data analysis. New York: Wiley.

Allwood, M.A.& Montgomery H. (1982): "Detection of errors in statistical problem solving". *Scandinavian journal of psychology*, 23, 131-139.

Allwood, M. A. (1984): "Error detection processes in statistical problem solving". *Cognitive science*, 8, 413-437.

Andersen, H.B. & Bove, T. (2001): "Validation of Human Error Classification". Contribution to System Analysis Department Annual Report 2000.

Amalberti, R. (2001): The paradoxes of almost totally safe transportation. *Safety Science*. 37 (2001) 109-126.

Amalberti, A. & Wioland, L. (1997): "Human error in aviation". Aviation Safety: Human factors, system engineering, flight operations. Soekkha, H. (Ed.). P.91-108.

Bagnara, S. & Rizzo, A. (1989): "A methodology for the analysis of error processes in human computer interaction". In: M.J. Smith & G. Salvendy (Eds.): *Work with computers: organizational, management, stress and health aspects*, pp. 605-612. Amsterdam, the Netherlands: Elsevier Science Publishers.

Bagnara, S; Rizzo, A. & Stablum, F. (1989): "Error analysis in man-machine systems". *Proceedings of the 6th Euredata conference*. Siene, Italy, March 15-17,1989.

Blickenderfer, E.; Cannon-Bowers, J.A. & Salas, E. (1998): Cross-training and Team Performance. In J.A. Cannon-Bowers & E.Salas (Eds.), Making decisions under stress: Implications for individual and team training (pp. 299-311). Washington, DC: American Psychological Association.

Bonni, D.; Jackson, A. & McDonald, N. (2001): Do I trust thee? An approach to understanding trust in the domain of air traffic control".

Bove, T. & Andersen, H.B. (2000): "Types of Error Recovery in Air Traffic Management". 3rd International Conference on Engineering Psychology and Cognitive Ergonomics

Brodbeck, F.C., Zapf, D., Prümper, J., & Frese, M. (1993). Error handling in office work with computers: A field study. *Journal of Occupational and Organizational Psychology, 66*, 303-317.

Cacciabue, P.C. (2001): "Human factors insight and data from incident reports: the case of ADREP-2000 for aviation safety assessment." *Engineering Psychology and Cognitive Ergonomics*. Vol 5. Harris, D. (ed.). Ashgate. England.

Carpenter, P.A. & Daneman, M (1981): Lexical retrieval and error recovery in reading: a model based on eye fixations. *Journal of verbal learning and verbal behaviour*, 20, 137-160

Cohen, J. (1960): "A coefficient of agreement for nominal scales". Education and psychological measurement. Vol XX., No.1.

Cooper, J.B.; Long, C.D.; Newbower, R.S. & Philip, J.H. (1982): "Critical Incidents Associated with Intraoperative Exchange of Anaesthesia Personnel". *Anaesthesiology*. Vol. 56, pp. 456-61.

D'arcy, J.F. & Della Rocco, P.S. (2001): Air Traffic Control Specialist Decision Making and Strategic Planning – A Field Survey. (DOT/FAA/CT-TN01/05). Atlantic City

International Airport: Federal Aviation Administration William J. Hughes Technical Center.

Degani, A; Chappell, S.L. & Hayes, M.S. (1991): Who and what saved the day? A comparison of traditional and glass cockpits. *Proceedings of the 6th international symposium on aviation psychology*. (pp.227-234). Jensen, R.S. (Ed.). Columbus: Ohio State University Press.

Dix, A.J.; Finlay, J.; Abowed, G. & Beale, R. (1993): "Human computer interaction." Cambridge, United Kingdom: Prentice-Hall.

Dormann, T. & Frese, M. (1994): "Error training: Replication and the function of Exploratory behavior". International Journal of Human-Computer Interaction 6(4). (pp.365-372.

Down, G. (2001): Threat and Error Management (TEM) in the Operational Environment. Slideshow. NavCanada.

Edmondson, A.C. (1996). Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *Journal of Applied Behavioral Science, 32*(1), 5-28.

Edwards, E. (1972): Man and Machine: Systems for safety. In Proceedings of British Airline Pilots Association Technical Symposium (21-36). British Airline Pilots Association: Lond.

Endsley, M. (1994): Situation Awareness in dynamic human decision-making: Theory. In R.D. Gilson, D.J.Garland & J.M. Koonce (Eds.), Situation awareness in complex systems (pp. 27-58). Daytona Beach, Florida: Embry-Riddle Aeronautical University Press.

Endsley, M. R., Mogford, R. H., Allendoerfer, K. R., Snyder, M. D., & Stein, E. S.: (1997): Effect of free flight conditions on controller performance, workload, and situation awareness (DOT/FAA/CT-TN97/12). Atlantic City International Airport: Federal Aviation Administration William J. Hughes Technical Center.

Flanagan, J.C. (1954): "The critical incident technique". *Psychological Bulleting*, 51, p. 327-358.

Fleiss, J.L. (1971): "Measuring nominal scale agreement among many raters". *Psychological Bulleting*, 76, p. 378-382.

Fleiss, J.L (1981). *Statistical Methods for Rates and Proportions.* New York: John Wiley & Sons. (Second Edition).

Fleishman, E.A. & Quantance, M.K. (1984): "Taxonomies of human performance. The description of human tasks. Academic Press. Orlando, Florida.

Frese, M. (1991): Error management or error prevention: two strategies to deal with errors in software design. Human Aspects in Computing: Design and Use of Interactive Systems and Work with Terminals. Bullinger, H.J. (ed.) Elsevier Science Publications.

Frese, M., & Van Dyck, C. (1996). Error Management: Learning from errors and organisational design. Internal paper, University of Amsterdam, Department of Work & Organisational Psychology.

Guerlain, S.A.; Smith, P.J.; Cross, S.M.; Miller, T.E.; Smith, J.W.; Svirbely, J.R. & Sachs, L. (1997): Interactive Critiqueing as a Form of Decision-Support: An Empirical Study. CSEL Technical Report #1996-14. The Ohio State University, Columbus OH.

Hawkins, F.H. (1987): Human Factors in Flight. Second edition. Edited by Harry W. Orlady. Avebury Technical.

Hayes, J.R. & Flower, L.S. (1980): Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), Cognitive Processes in Writing. Hillsdale, N.J.: Erlbaum.

Helmreich, R.L., & Merritt, A.C. (1998). Culture at work in aviation and medicine: National, organizational, and professional influences. Aldershot, U.K: Ashgate.

Helmreich, R.L., & Merritt, A.C. (2000). Safety and error management: The role of Crew Resource Management. In B.J. Hayward & A.R. Lowe (Eds.), *Aviation Resource Management* (pp. 107-119). Aldershot, UK: Ashgate. (UTHFRP Pub250)

Helmreich, R.L.; Klinect, J.R. & Wilhelm, J.A. (1999): "Models of threat, error, and CRM in flight operations". *Paper from the 10th International Symposium on Aviation Psychology*. Columbus, Ohio May 3-6, 1999.

Helmreich, R.L., Wilhelm, J.A., Klinect, J.R., & Merritt, A.C. (2001). Culture, error and Crew Resource Management. In E. Salas, C.A. Bowers, & E. Edens (Eds.), *Improving Teamwork in Organizations: Applications of Resource Management Training* (pp.305-331). Hillsdale, NJ: Erlbaum. (UTHFRP Pub254)

Hollnagel, E. (1990). The phenotype of erroneous actions: Implications for HCI design. In G. Weir & J. Alty (Eds.), *Human-computer Interaction and complex systems*. London: Academic Press.

Hopkin, V.D. (1995): "Human Factors in Air Traffic Control". Taylor & Francis Ltd. London.

Hughes, J.A.; Randall, D. & Shapiro, D. (1992): Faltering from Ethnography to Design. Proceedings from CSCW '92. New York, pp.-115-122.

Hutchins, E.L.; Hollan, J.D. & Norman, D.A. (1985): Direct Manipulation Interfaces. In Human-Computer Interaction, Vol. 1, pp. 311-338. Lawrence Erlbaum Associates, Inc.

Hutchins, E. (1994): "Cognition in the wild". MIT Press. London, England.

Isaac, A.; Shorrock, S.T.; Kirwan, B; Kennedy, R; Andersen, H.B. & Bove, T. (2000): Learning from the past to protect the future – the HERA Approach. *Paper for 24th conference of European Association For Aviation Psychology (EAAP) 2000.*

Isaac, A.; Shorrock, S.T.; Kennedy, R; Kirwan, B; Andersen, H.B. & Bove, T. (2001): The Human Error in ATM (HERA) Technique. Edition 0.2 – Second Draft. Internal document. Eurocontrol.

Isaac, A.; Shorrock, S.T. and Kirwan, B. (2002) Human Error in European Air Traffic Management: the HERA Project. *Reliability Engineering and Systems Safety* (75). 257-272.

Jambon F. (1997) Error Recovery Representations in Interactive System Development. Third Annual ERCIM Workshop on "User Interfaces for All", Obernai, France, 3-4 November 1997. p. 177-182.

Jensen, P.F. (1997): Development of a methodology for cognitive analysis of critical incidents in anaesthesia. Ph.D. Thesis. University of Copenhagen.

Johannsen, G. (1988): "Categories of human operator behavior in fault management situations". In: Task, Errors and Mental Models. Goodstein, L.P.; Andersen, H.B. & Olsen, S.E. (ed.). Taylor & Francis. London.

Jones, S.G. (1997). The last line of defence: Building a safety net for controllers. *Human Factors*, 36, 11-14.

Jones, S. (1998). Air traffic control: A starting point. In *Proceedings of the Ninth International Symposium on Aviation Psychology* (pp. 219-224). Columbus, OH: The Ohio State University.

Jones, S.G., & Tesmer, B. (1999). A new tool for investigating and tracking human factors issues in incidents. In *Proceedings of the Tenth International Symposium on Aviation Psychology* (pp. 696-701). Columbus, OH: The Ohio State University.

Jonker, H.G. (2000): Cockpit Decision Making. How the Rule of Three can Help Making Go-No-Go Decisions. Master's Thesis. Leiden University.

Kanse, L., & Van der Schaaf, T.W. (2000a). Failure Recovery in Process Industry - An Empirical Comparison of Two Models. In: P.C. Cacciabue (Ed.): *Proceedings EAM 2000, 19th European Annual Conference on Human Decision Making and Manual Control*, Ispra, Italy, June 26-28, 2000, pp. 153-163.

Kanse, L. & Van der Schaaf, T.W. (2000b): Recovery of Failures in the Chemical Process Industry. *3rd International Conference on Engineering Psychology and Cognitive Ergonomics*.

Kanse, L., & Van der Schaaf, T.W. (2000c). Recovery from failures - Understanding the positive role of human operators during incidents. In: Proceedings Human Factors and Ergonomics Society Europe Chapter Annual Meeting 2000, Maastricht, Netherlands, November 1-3, 2000, pp. 367-379.

Kanse, L., & Van der Schaaf, T.W. (2001). Factors influencing recovery from failures. CSAPC '01. *8th Conference on Cognitive Science Approaches to Process Control*. 24-26 September 2001. Universitat der Bundeswehr, Neubiberg, Germany.

Kerns, K.; Smith, P.J.; McCoy, C.E. & Orasanu, J. (1999): Ergonomic Issues in Air Traffic Management. In W. Marras and W. Karwowski (eds.). Handbook of Industrial Ergonomics. New York: Marcel Dekker, Inc., 1979-2003.

Kinney, G.C., Spahn, M.J. and Amato, R.A. (1977). The human element in air traffic control: Observations and analyses of the performance of controllers and supervisors in providing ATC separation services. METRIEK Division of the MITRE Corporation: MTR-7655.

Klein, G.A. (1989): Recognition-primed decisions. In W.B. Rouse (Ed.), Advances in Man-Machine Systems Research. 5, 47-92. Greenwich, CT: JAI Press.

Klinect, J.R.; Wilhelm, J.A. & Helmreich, R.L. (1999): "Threat and Error Management: Data from Line Operations Safety Audits". *Paper from the 10th International Symposium on Aviation Psychology*. Columbus, Ohio May 3-6, 1999.

Kontogiannis, T. (1997). A framework for the analysis of cognitive reliability in complex systems: A recovery centred approach. *Reliability Engineering & Systems Safety (RESS)*, 58, 233-248.

Kontogiannis, T. (1999): "User strategies in recovering from errors in man-machine systems." *Safety Science* 32, 49-68.

Landis, J. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, (33), 159-174.

Lenman S. & Robert, J.L. (1994): "A framework for error recovery". 12th triennial congress. Vol. 6 (p. 374-376)

Lewis, C. & Norman, D.C. (1986): "Designing for Error." User Centred System Design/new Perspectives on Human-Computer Interaction, Hillsdale (NJ), USA: Lawrence Erlbaum Associated, 1986, p.411-432.

Maurino, D.E. (1999). "Safety prejudices, training practices and CRM: A mid-point perspective". *The International Journal of Aviation Psychology* 9: (4) 413-422.

McCoy, W.E. and Funk, K.H. (1991). Taxonomy of ATC Operator errors based on a model of human information processing. In R.S. Jensen (Ed), *Proceedings of the Sixth International Symposium on Aviation Psychology*, 29 April to 2 May, Columbus, Ohio.

McGrath, J. E. (1994): "Methodology matters: doing research in the behavioural and social sciences". In: Human-Computer Interaction: Toward the Year 2000. Baecker, R.M.; Gruding, J.; Buxton, W.A.S. Buxton and Greenberg, S. (Eds.). Morgan Kaufmann Publishers, inc. San Francisco, California (p. 152-169).

Mogford, R.H.; Allendoerfer, K.R.; Snyder, M.D.; Hutton, R.J.B. & Rodgers, M.D. (1997): Application of the Recognition-Primed Decision Model to the Study of Air Traffic Controller Decision Making. *Ninth International Symposium on Aviation Psychology* (Vol. 1, pp. 739-744)

Morrow, D.; Lee, A. & Rodvold, M. (1993): Analysis of Problems in Routine Controller-Pilot Communication. *The International Journal of Aviation Psychology*, 3(4), 285-302.

Nagel. D.C. (1988): Human error in aviation operations. In *Human Factors in Aviation*, Wiener, E.L. & Nagel, D.C, eds. Academic Press: New York. pp. 263-303.

Nooteboom, S.G. (1980): Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. Fromking (Ed.), Errors in Linguistic Performance. New York, Academic Press.

Nordstrom, C.R.; Wendland, D. & Williams, K.B. (1998): "To err is human: an examination of the effectiveness of error management training." *Journal of business and psychology*. Vol. 12, No. 3.

O'Leary, M. (1999): "The British Airways Human Factors Reporting Programme". *3rd Workshop on human error, safety, and system development (HESSD '99)*, Liège (BE), 7-8 Jun 1999.

Orasanu, J. and Fischer (1997): "Finding decisions in natural environments: the view from the cockpit". Naturalistic Decision Making. Zsambok, C.E. & Klein, G. Lawrence Erlbaum Associates, Publishers. Mahwah, New Jersey.

Orasanu, J.; Fischer, U; McDonnell, L.K.; Davison, J. & Haars (1998): "How do flight crews detect and correct errors? Findings from a flight simulation study". Proceedings of the human factors and Ergonomics society 42nd annual meeting.

Orasanu, J.; McDonnell, L.K. & Davison, J. (1999): "How do flight crews detect and correct errors? I. Explanations for failures to correct errors". *Paper from the 10th International Symposium on Aviation Psychology.* Columbus, Ohio May 3-6, 1999.

Rabbitt, P.M.A. (1966): Errors and error-correction in choice-response tasks. *Journal of Experimental Psychology*, 71, 264-272.

Rasmussen, J. (1982): "Human Errors. A taxonomy for describing human malfunction in industrial installations". *Journal of Occupational Accidents*. (p.311-333).

Rasmussen, J. (1983a): “Skills, Rules and Knowledge; Signals, Signs and Symbols and Other Distinctions in Human Performance Models*”. IEEE Transactions on System, Man and Cybernetics*, Vol. SMC-13, No.3. May-June. (p.257-266).

Rasmussen, J. (1983b): "Human errors process control". Contribution to NATO Workshop on Origin of Human Error, September, Bellagio, Italy, 1983.

Rasmussen, J. (1984): “Human error data. Facts or fiction?” Invited paper presented at a seminar on Accident Research in Rovaniemi, Finland, April 1984. Risø-M-2499, 1984.

Rasmussen, J. (1987): “Approaches to the Control of the Effects of Human Error on Chemical Plant Safety”. Published in John L. Woodward (Ed.): International Symposium on Preventing Major Chemical Accidents, Washington DC, American Institute of Chemical Engineers.

Rasmussen, J. (1997): “Risk Management in a Dynamic Society: A modeling problem. *Safety Science* Vol. 27, No. 2/3, pp. 183-213.

Rasmussen, J. & Jensen, A. (1974): “Mental procedures in Real Life Tasks: A Case Study of Electronic Trouble Shooting”. *Ergonomics*, 1974, Vol. 17, No. 3.

Reber, A.S. (1985): The Penguin Dictionary of Psychology. Penguin Books. London, England.

Reason, J. (1990): “Human Error”. Cambridge University Press. Cambridge.

Reason, J. (1997): “Managing the Risks of Organisational Accidents”. Ashgate, Aldershot.

Reason, J. & Lucas, D. (1984): "Using cognitive diaries to investigate naturally occurring memory blocks". In: "Everyday memory, actions and absent-mindedness", (Eds.) Harris, J.E. & Morris, P.E. (pp. 53-70).

Rizzo, A.; Bagnara, S. & Visciola, M. (1987): "Human error detection processes*″*. *International journal of man-machine studies*, 27, 555-570.

Rizzo, A; Ferrante, D & Bagnara, S. (1995): "Handling human error". Expertise and technology. Cognition & Human-computer Cooperation. Hoc, J.M.; Cacciabue, P.C & Hollnagel, E. (eds.). Lawrence Erlbaum Associates, Publishers. Hove, UK.

Rofske-Hofstrand, R.J. & Murphy, E.D. (1998): Human Information Processing in Air Traffic Control. In Smolensky, M. & Stein, E. (Eds.): Human Factors in Air Traffic Control. Academic Press, San Diego, CA, pp. 65-113.

Rognin, L. & Blanquart, J.P. (1999): Impact of Communication on Systems Dependability – Human Factors Perspectives. In M. Felici, K. Kanoun, A. Pasquini (Eds.): Computer Safety, Reliability and Security. 18th International Conference, SAFECOMP'99, Toulouse, France, September 1999, p. 113 ff.

Rognin, L; Salembier, P & Moustapha, Z. (1998): Cooperation, interactions and socio-technical reliability in Air-Traffic Control. Comparing French and Irish settings. Ninth European Conference on Cognitive Ergonomics. University of Limerick, Ireland, September 1998

Rouse, W.B. & Morris, N.M. (1987): Conceptual Design of a Human Error Tolerant Interface for Complex Engineering Systems. Automatica. Vol. 23. No. 2, pp. 231-235.

Sanders, M.S. & McCormick, E.J. (1992): Human Factors in Engineering and Design. Seventh Edition. McGraw-Hill, Inc. New York.

Sarter, N.B. & Alexander, H.M. (2000): Error types and related error detection mechanisms in the aviation domain: An analysis of aviation safety reporting system incident reports. *The international journal of aviation psychology*, 10(2), 189-206. Lawrence Erlbaum Associates, Inc.

Sasou, K. & Reason J. (1999): "Team errors: Definition and taxonomy". *Reliability Engineering and System Safety*, 65. 1-9.

Seifert, C.M. & Hutchins, E. (1994): "Error as opportunity: learning in a cooperative task." Human-Computer Interaction. Vol. 7. (pp.409-435)

Sellen, A.J. (1994). Detection of Everyday Errors. *Applied Psychology: An International Review, 43*(4), 475-498.

Senders, J.W. and N.P. Moray (1991): *Human Error: cause, prediction and reduction*, LEA Publishers, Hillsdale, NJ.

Shappell, S.A. & Wiegmann, D.A. (1999): Human Factors Analysis of Aviation Accident Data: Developing a Needs-Based, Data-Driven, Safety Program. *3rd Workshop on human error, safety, and system development (HESSD '99)*, Liège (BE), 7-8 Jun 1999.

Shorrock, S.T. and Kirwan, B. (1998). The development of TRACEr: a technique for the retrospective analysis of cognitive errors in ATM. *Paper presented at the 2nd Conference on Engineering Psychology and Cognitive Ergonomics*. Oxford: 28 - 30 October.

Smith-Jentsch, K.A; Zeisig, R.L.; Acton, B & McPherson, J.A. (1998): Team dimensional training: A strategy for guided team self-correction. In J.A. Cannon-Bowers & E. Salas (Eds.), Making Decisions under Stress: Implications for individual and team training (pp. 271-297). Washington, DC: American Psychological Association.

Sperandio, J.C. (1971): Variation of operator's strategies and regulating effects on workload. *Ergonomics*, 14 (5), 571-577.

Sperandio, J.C. (1978): "The regulation of Working Methods as a Function of Work-load among Air Traffic Controllers". *Ergonomics*. 21, 195-202.

Swain, A.D. & Guttmann, H.E. (1983): "Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications". NUREG/CR 1278. Albuquerque, N.M.: Sandia National Laboratories, 1983.

Tullo, F. & Salmon, T. (1997): "The role of the check airman in error management". *Paper from the 9th International Symposium on Aviation Psychology.* Columbus, Ohio.

Van der Schaaf, T.W.; Lucas, D.A. & Hale, A.R. (Ed.) (1991): Near-miss reporting as a safety tool. Butterworth-Heinemann Ltd. Oxford.

Van der Schaaf, T.W. (1988). Critical incidents and human recovery: Some examples of research techniques. In: L.H.J. Goossens (Ed.): Human recovery: Proceedings of the COST A1 Seminar on Risk Analysis and Human Error. Delft: Delft University of Technology.

Van der Schaaf, T.W. (1995). Human recovery of errors in man-machine systems. In: *Proceedings of the 6th IFAC Symposium on Analysis, Design, and Evaluation of Man-Machine Systems*, June 27-29, 1995, MIT, Cambridge, MA, USA.

Van der Schaaf, T.W., & Kanse, L. (1999). Error recovery in socio-technical systems. In: *Proceedings from CSAPC '99*, Seventh European Conference on Cognitive

Science Approaches to Process Control, September 21-24, 1999, Villeneuve d'Asq, France, pp.151-156.

Van der Schaaf, T.W., & Kanse, L. (2000). Errors and error recovery. In: P.F. Elzer, R.H. Kluwe and B. Boussoffara (Eds.): Human Error and System Design and Management, pp. 27-38. London: Springer Verlag.

Van Dyck, C., Frese, M., & Sonnentag, S. (1999). Error management culture and organizational performance: On mastering the challenges of errors. Internal paper, University of Amsterdam, Department of Work & Organisational Psychology.

Volpe, C.E.; Cannon-Bowers, J.A. & Salas, E. (1996): The Impact of Cross-Training on Team Functioning: An Empirical Investigation. Human Factors, 38(1), p. 87-100.

Vortac, O.U.; Edwards, M.B. & Manning, C.A. (1995): "Functions of External Cues in Prospective Memory". *Memory*, 3 (2), 201-219.

Wickens, C.D. (1987): "Information Processing, Decision-Making, and Cognition". In: Handbook of Human Factors. Salvendy, G. (eds.). John Wiley & Sons. New York. (p.72-107).

Wickens, C.D. (1992): Engineering Psychology and Human Performance. Harper Collins Publishers. 2nd Edition. New York.

Wickens, C.D.; Mavor, A.S. & McGee, J.P. (eds) (1997): Flight to the Future. Human Factors in Air Traffic Control. National Academy Press. Washington, D.C.

Wickens, C.D.; Mavor, A.S.; Parasuraman, R. & McGee, J.P. (eds) (1998): The Future of Air Traffic Control. Human Operators and Automation. National Academy Press. Washington, D.C.

Wickens, C.D. & McCloy, T.M. (1993): ASRS and aviation psychology. *Proceedings of the 7th international symposium on aviation psychology*. Jensen, R.S. (Ed.)

Wiegmann & Shappell (1997): Human Factors analysis of postaccident data. *The International Journal of Aviation Psychology*. Vol 7(1), (p. 67-81).

Wiegmann and Shappell (2002): Human Error Perspectives in Aviation. *The International Journal of Aviation Psychology*, 11(4), pp. 341-357.

Willems, B. & Truitt, T. R (1999). Implications of reduced involvement in en route air traffic control (DOT/FAA/CT-TN99/22). Atlantic City International Airport: Federal Aviation Administration William J. Hughes Technical Center.

Wioland, L., & Amalberti, R. (1996*). When errors serve safety: towards a model of ecological safety.* Paper presented at CSEPC '96, Cognitive Systems Engineering in Process Control, Kyoto, Japan, November 1996, pp. 184-191.

Wioland, L & Amalberti, R (1998): "Human error management: towards an ecological safety model. A case study in an Air Traffic Control microworld." *Paper presented at the 9th European Conference on Cognitive Ergonomics* - Limerick University, Ireland August 24-26.

Wioland, L. & Doireau, P. (1995): "Detection of human error by an outside observer: a case study in aviation". *5th European conference on cognitive science approaches to process control*. Espoo, Finland, Aug. 30 - Sept. 1, 1995. Norros, L. (Ed.)

Wioland, L.; Skaaning, G. & Kaarstad, M. (1999): "Error detection and recovery - results from experiment 1997". HWR-622. OECD Halden Reactor Project.

Woods, D.D. (1984). Some results on operator performance in emergency events. *Institute of Chemical Engineers Symposium Series, 90*, 21-31.

Woods, D.D.; Johannesen, L; Cook R. & Sarter, N. (1994): "Behind human error: Cognitive Systems, Computers, and Hindsight". Soar Cseriac 94-01, Dayton: Ohio.

Zapf, D. & Reason, J.T. (1994). Introduction: Human errors and error handling. *Applied Psychology: An International Review, 43*(4), 427-432.

Appendixes

Appendix A: Glossary

ACC	Area Control Centre
Accident	An event leading to physical harm or damage, brought about unintentionally.
ADREP-2000	ADREP-2000 is a classification system that has been proposed by the International Civil Aviation Organisation (ICAO) for structuring the data analysis of aviation accidents.
APP	Approach Controller
AR	Adjusted Residual
ATC	Air Traffic Control
ATCO	Air Traffic Controller
ATM	Air Traffic Management
BASIS	British Airways Safety Information System
CAA	Civil Aviation Authorities
Cognitive theory	Theories of or pertaining to the mental processes of perception, memory, judgement, and reasoning.
Construct validity	Construct validity is probably the most difficult type of validity to establish and cannot be done within a single research study. It refers to the extent to which evidence points to the construct or concept being useful in a scientific endeavour.
Content validity	Relates to whether the methodology adequately represents the variety and balance of the field it purports to examine.
Criterion validity	A criterion can be seen as an already validated and accepted standard to which a measurement or methodology can be compared.
Critical incident	An unintended event which could have reduced, or did

	reduce the safety margin of the system
Critical Incident Technique	Interview technique originally developed by the Flanagan (1954) as a systematic effort to gather incidents of effective and ineffective behaviour with respect to a designated activity.
CRM	Crew Resource Management
Error	Any action (or inaction) that potentially or actually results in negative system effects given the situation that other possibilities were available. This includes any deviation from operating procedures, good working practice or intentions.
External validity	Can the results be generalised to other situations or domains?
Face validity	Does the framework seem reasonable, using 'common sense'?
Failure	Can be a technical or human failure
Fault	Equipment breaking down or ceasing to function
FL	Flight Level
FPB	Flight Progress Board
FMS	Flight Management System
FPS	Flight Progress Strip
Framework model	A frame or structure composed of a hierarchical set of theories which fit together and make up a coherent overall theory.
HCI	Human-Computer Interaction
Hearback	The procedure associated with listening to instruction read back by the pilot.
HERA	Human Error Reduction in ATM. Error taxonomy developed for the European organisation for Air Traffic Control (EUROCONTROL).

HRA	Human Reliability Assessment
ICAO	International Civil Aviation Organisation
IFR	Instrument Flight Rules
ILS	Instrument Landing System
Incident	A critical occurrence that could have led to an accident if not recovered or spontaneously resolved.
Interactive critiquing	A concept that has been proposed to overcome the problems associated with the traditional cooperative architecture of decision aids. Instead of having the human to critique the computer the computer system will be assigned with the role of critiquing the system user's problem-solving.
Internal validity	Are the results valid within a particular setting?
LOSA	Line Operations Safety Audit. Method where experts observe and collect data about crew behaviour and situational factors on normal flight.
Mistake	An error caused by an act where the intention itself was wrong.
PSF	Performance Shaping Factors
Readback	The procedure required by pilots when they have been given an instruction to repeat the instruction.
Reliability	Consistency of a methodology in providing the same results, e.g. among different observers (Inter-observer reliability) or repeatedly with the same observers (Intra-observer reliability).
Risk	A chance or possibility of danger, loss or injury or other adverse consequence
R/T	Radio/Telephone
SHEL-model	In this model the focus is on the human component (i.e. the liveware) and its interaction with other main components within a socio-technical system. The components are given in the SHEL-acronym: Software,

	Hardware, Environment and Liveware.
Slip	An error caused by an act where the intention was correct but the actual action was wrong.
SRK framework	A framework describing cognitive control mechanisms. Contains three distinct cognitive levels of problem solving activities: Skill-based, rule-based, and knowledge based performance.
SSR	Secondary Surveillance Radar, a radar-type system that requires a transponder to transmit a reply signal.
STCA	Short-Term Conflict Alert
Taxonomy	Classification, e.g. the systematic classification of phenomena into groups or taxa.
TCAS	Traffic Collision Avoidance System
Threat	Threats are operational factors that have the potential of jeopardising safety and require active operator involvement to maintain safety.
Threat Management	Threat management is the act of anticipating and minimising the potential consequences of threats on flight safety
TRM	Team Resource Management
TWR	Tower Controller
Validation	The process of determining the validity of a framework.
Validity	Validity is related to the degree to which a study accurately reflects or assesses the specific concept that a researcher is attempting to measure.
VFR	Visual Flight Rules

Appendix B: PSF Taxonomies

HERA PSFs
• **Traffic and Airspace** are factors associated with the airspace which can influence the ATCOs task such as excessive traffic load, air space design characteristics and poor or unpredicted weather. • **Pilot - Controller Communications** describe factors that can influence breakdowns in the communication between the pilot and the controller such as pilot/language difficulties and pilot breaches of R/T standards. • **Procedures and Documentation** concern everything from problems with procedures (e.g. incomplete, poor, unclear or contradictionary) to inappropriate regulations and standards. • **Training and Experience** are factors that affect the individual ATCO's abilities and resources to respond to the demands encountered such as inexperience on position, inadequate specialist training (e.g. emergency training and TRM training) and inadequate time on position due to other duties. • **Social and Team Factors** are different kinds of factors associated with the communication and interaction between ATCOs such as unclear hand over/takeover, trust in others (over/under/miss) and inadequate assertiveness. • **Workplace Design, HMI and Equipment Factors** are the influences that are associated with the technical and design related aspects of the work environment and include radar failure and HMI-deficiencies (e.g. visibility and consistency). • **Ambient Environment** concerns disturbing factors such as noise, distraction (e.g. by supervisor or colleagues) and lighting (e.g. illumination and glare). • **Person Related Factors** concerns characteristics associated with the individual controllers such as fatigue, boredom, complacency and confidence in self and others. • **Organisational Factors** are factors associated with broader organisational issues that directly or indirectly can affect the working conditions of the controllers such as problems in the work environment (e.g. general understanding/manning levels, work scheduling and poor relations/confidence with management) and companies/commercial pressure.

ADREP-2000
Liveware: • Personal physical and sensory limitations • Human physiology • Psychological limitations • Personal workload management • Experience, knowledge and regency **Liveware (Human)-Environment Interface:** • Physical environment • Psychosocial factors • Company, management, manning and regulatory issues • Operational task demands **Liveware (Human)-Hardware/Software Interface:** • Human and hardware interface • Inadequate information/data sources • Human firmware/software interface • Automation/automatic systems • Automatic defences/warnings • Operational material **Liveware (Human)-System Support Interface:** • Human/system interface procedures • Human/system interface training **The Liveware (Human)-Liveware (Human) Interface:** • The interface between humans in relation to communication • The interface between humans in relation to interaction/team skills crew/team resource management training • The interface between humans in relation to supervision • The interface between humans in relation to regulatory requirements

Recovery Influencing Factors
• Factors relevant for prioritisation of recovery related tasks • Occurrence related factors • Person related factors • Social factors • Organisational factors • Technical/workplace/situational of factors

ASAP Contributing Factors (Cockpit Crew Factors)
• Physiological State • Memory/recall • Experience • Procedures selected • Procedural compliance • Setting priorities • Distributing workload • Situational awareness • Assessing threats to safety • Verbal statements of plans/challenging for clarification • Review status/modification of plans • Monitoring/cross-checking of communications or settings

BASIS
• **Crew actions.** This group includes, among other things, the error types from Reason's model of human performance and Helmreich's CRM team skills. • **Personal influences.** This group is concerned with the subjective state of the individual actor and includes e.g. boredom, personal stress and tiredness. • **Organisational influences.** These factors are under the control and responsibility of the company such as training and technical support. • **Informational influences.** These are influences related to the operational information and materials such as standard operating procedures and electronic checklists • **Environmental influences.** These are influences outside the control of the company such as the ATC services and technical failures.

Appendix C: Interview guide

Briefing
"The present study will seek to gain more knowledge of how potential critical situations are captured in operational practice. In other words, it is the positive side of the critical incidents that we are interested in - that is, how you managed to detect and recover (which requires skills and knowledge), and not so much what caused the problem in the first place. The main purpose of this interview is to gather and analyse a record of ATM-related potential critical incidents and how they were discovered and recovered. In other words, I am very much interested in hearing your descriptions of *specific situations* which could have led to a dangerous outcome if not discovered. For the specific situations I focus on it is important that (1) You played a central role in the discovery and/or recovery of the situation. (2) The situation occurred recently (e.g. within a year). It should be emphasised that *a substantive negative outcome is not required*. The focus is on events that have the potential of negatively impacting safety, but do not fall within the category of "reportable" incidents or accidents Please notice that even though I do have some basic knowledge concerning ATM, I am in no way expert and it may therefore be necessary to explain some of the concepts and abbreviations to us. For practical reasons I would like to emphasise that this interview is strictly confidential and therefore no data that can be used to identify you or your colleagues will be reported. In particular, no feedback will be given to management about particulars of reports that could identify you or your colleagues or the ATM centre in question. The interview is expected to last between 45 and 60 minutes. Do you have any questions before we begin?"

Questions
1: "Can you think of an experience that stands out in your mind as a critical situation? Could you please tell me what happened with as much detail as possible (what, where, when, who)? It would be helpful if you can make a drawing of the scenario." *If the anecdote is not usable:* "Do you have any other examples of critical situations?" *Note 1: Be sure to get background information - such as the type of ATM position involved, the time of the day, number of people in the position and the workload.* *Note 2: When having heard the critical incidents, echo back a version to make sure you understand all the details and can relate them clearly.*
2: "Why do you think the situation occurred? Were you aware in advance that problems could arise and, if so, did you make any initiatives to avoid or minimise them?"
3: "Please tell me about the discovery of the problem. Who discovered the problem? When was the problem discovered? What source(s) of information was used (radar, strips, communication with pilots/controllers, etc.)? How was the problem realised? Where any attempts made at finding the root cause of the problem? Was it a problem you have (frequently) encountered in the past?"
4: "Who recovered the situation? How was the situation recovered (e.g. What were your specific goals at the time? Did it require a routine or a more 'creative' response? Were you reminded of any previous experience? What specific training or experience was necessary or helpful in making the decision? Were other courses of action considered and, if so, why were they rejected? Was the time pressure and risk high?)"
5: "What was done and what were the consequences (e.g. violation of separation standards, inconvenience to other people or inconsequential)? What do you think would have happened if no recovery was made?"
7: "What could improve to prevent this event from reoccurring in the future? That is, do you think there are any lessons learned from this incident?"
8: "If I should have some follow-up questions is it okay if I contact you again?" *Note: Get E-mail, telephone number, etc.*
9: "I don't have any further questions. Is there anything that you would like to add or ask before we close this interview?"

Debriefing
A more elaborated description of the purpose and design of the interview is given, if wanted, after the tape recorder is stopped.