Chapter 15

Monitoring

The New Zealand Transport Accident Investigation Commission argues that "the sole purpose of each (incident) report is to avoid similar occurrences" [634]. This chapter looks beyond this high-level goal to identify the problems that arise when attempting to monitor the success of any incident reporting system. For instance, it can be difficult to prevent similar occurrences when new technology and working practices introduce new froms of previous failure. It can also be difficult to ensure that any reporting system gathers sufficient information about other failures to be sure that similar failures are not going unreported. Given such uncertainty, it is particularly important that any monitoring activity justifies the investment that regulators and operators must make to sustain reporting systems. Previous chapters have described the many different activities that must be managed during the investigation, analysis and dissemination of incident reports. Domain experts must initiate follow-up interviews, site visits and data acquisition. They must filter relevant information from the mass of contextual details that are elicited in the aftermath of an adverse event. They must also ensure that the products of any root cause analysis are well-documented so that others can reconstruct the arguments that support particular recommendations. We have seen that computational tools can assist in the elicitation, classification, dissemination and retrieval of incident reports. Computer-based forms can be developed to collect initial information about an adverse event. Automated interviewing systems can prompt domain experts to consider certain causal hypotheses in the aftermath of an incident. The previous chapter has described the use of lexical retrieval tools, relational databases and case-based reasoning systems to identify patterns of failure. It is important to emphasise, however, that these tools have not been widely applied to support incident reporting systems. The costs of manually performing these various activities can, therefore, act as a significant disincentive to the creation and maintenance of many reporting schemes. It is, therefore, important that safety managers can demonstrate the 'cost-effectiveness' of any proposed system.

Who Monitors What and Why

A host of problems complicate attempts to monitor the 'cost-effectiveness' of incident reporting systems. For instance, it can be difficult to establish that any safety improvements can be attributed to a reporting system rather than to other wider changes within a company or an industry. A more detailed examination of the barriers to incident monitoring is postponed until later sections. It is first important to identify those individuals and organisations that are concerned to validate the effectiveness of a reporting system.

Safety Managers. There are many different reasons for monitoring incident data. For instance, safety managers can use information about adverse events to justify the remedial actions that are intended to prevent future failures. This also, indirectly, helps to justify the existence of the reporting system. Safety managers can also monitoring incident reports to identify the need for further contributions about certain safety concerns [169]. They can use incident information to track progress towards higher-level safety targets. They can also use monitoring data to inform employees and the public about particular safety issues and so on. For example, the Washington Metro Area Transit Authority's Safety Officer issued a press release to publicise customer injuries on escalators [856]. These accounted for 43% of all passenger injuries in the first quarter of 2002, a 5% reduction from 2001. The Safety Officer argues that this represents significant progress. He also uses this data to justify a new safety awareness campaign involving community outreach, new car cards, station posters and brochures. This example illustrates the way in which incident monitoring helps Safety Officers to recognise the scale of particular problems. They can then use the available statistics to inform the public about potential hazards. The same press release also illustrates the way in which incident monitoring can be used to provide feedback to employees about previous safety initiatives. The Safety manager describes how there were 13 reported fires and smoke incidents requiring a fire service response in the first quarter of 2001. These resulted in an average delay of under 30 minutes per incident. In the first quarter of 2002, there were only 10 such incidents with an average delay of under 20 minutes. The Safety Manager concluded that 'the reduction in the number of incidents is a result of improved maintenance measures, and interagency coordination and communications with the local fire departments' [856].

The information provided by the Washington Metro Area Transit Authority illustrates the use of incident monitoring to track relatively long term trends. Safety managers can also use this data to identify sudden increases in particular types of failure. For example, Southern California's Metrolink Rail System has used monitoring information to provide a rapid response to changes in the types of incident that are being reported [549]. In February 2000, it was realised that there had been 5 different incidents involving trains and trucks in the Southland area in a 90 day period. This formed a sharp contrast with the previous 24 months in which there had only been a single comparable incident. The five more recent incidents were 'near misses' in the sense that they resulted in relatively minor injuries. The careful monitoring of these incident statistics helped to trigger a more detailed causal analysis. This identified that a booming regional economy had resulted in an increase in freight carrier traffic. Many of the truck drivers who were brought in to satisfy this demand were unfamiliar with the Metrolink train operations. The company responded to this rapid increase in truck-train incidents by contacting local haulage firms and by a series of awareness raising initiatives including a 'Trucker on the Train' day as part of a Metrolink Rail Safety Week. This enabled truck drivers to ride with Metrolink engineers so that they can learn to avoid potential collisions.

The two previous examples have focussed on the use of incident monitoring to chart progress towards recognised safety objectives and to trigger rapid intervention when new hazards arise. This illustrates how safety managers can use data to provide others with information about adverse events and near miss incidents. It is also important to monitor the reporting system itself to ensure that submissions are handled in a timely fashion. Safety managers and the operators of reporting systems must also track any causal analysis to identify potential bias [855]. This must include some consideration of intra-rater reliability; will the same coder code similar incidents in the same way over time. Similarly, the findings of particular analysts may be compared for the same incidents to ensure inter-rater agreement. Davies, Wright, Courtney and Reid describe the results of performing this type of monitoring activity for the CIRAS voluntary reporting system that operates across Scottish railways [198]. This system is based around a classification system, similar to those described in Chapter 10.4, which provides 54 different causal categories. Two CIRAS personnel independently analysed a total of 439 incidents with 84.6% agreement over the causal classification. This involved the assignment of 1,955 codes for human factors issues alone. Such results represent a remarkable level of consistency. Chapter 11.5 has reviewed the many problems that can jeopardise agreement between independent analysts. For example, Lekberg has shown that individuals from different educational and operational backgrounds will code the same incident in different ways [484]. Monitoring techniques, such as those introduced in the following pages, provide Safety Managers with a means of assessing whether or not such factors are introducing significant biases into the analysis of and response to adverse events.

Company Monitoring. The previous section described some of the reasons why Safety Managers might choose to monitor the performance of an incident reporting system. At a corporate or organisational level, there can be more pressing requirements to track the data that can be obtained about incidents and accidents. For example, Canada's Railway Safety Act incorporates an annex that describes various requirements that must be satisfied by Railway Safety Management Systems. These include the provision that all railway companies must record safety-related information for the purpose of 'assessing its safety performance' [780]. This information should include 'accident and incident investigation reports and a description of the corrective actions taken for accidents and incidents'. Companies must also monitor accident and incident rates that should be calculated in terms of (i) employee deaths, disabling injuries and minor injuries, per 200,000 hours worked by the employees of the railway company, and (ii) train and grade crossing accidents per million train miles. The Railway Safety Management Systems annex also states that railway companies can be required to collect, maintain and submit specified performance or safety data for "the purpose of monitoring the effectiveness of its safety management system and its safety performance". These are important provisions because they specify the way in which companies must normalise their incident data to account for differences in the operating characteristics of individual companies. Clearly, raw incident frequencies for national carriers and for local railways cannot provide an adequate means of comparison. The number of journeys as well as the distance and time of travel combine to make the risk exposure radically different in each of these cases. It is possible to have a profound effect on the nature of safety statistics depending in which of these normalising factors are used. In contrast to the Canadian provisions mentioned above, the FRA calculates the total accidents and incidents rate by multiplying the number of accident and incident reports by 1,000,000 and then dividing the result by the sum of train miles and hours. This reflects a different approach in which companies do not directly perform the normalisation themselves. This is done by the regulator in assessing the performance of each operator. Such an approach raises a number of dilemmas for operating companies that must monitor accident and incident rates. In particular, they must still report normalisation statistics for periods in which they may have few or no safety related occurrences. This is necessary if regulators are to assess the performance of an industry in terms of total incidents per miles travelled, passengers carried etc. The government of South Australia has recently eased the burdens associated with the reporting of normalising factors through the development of a web page [783]. This asks operators to report how many kilometres of track they own and manage within Southern Australia. They should also report the distance, in kilometers, that their passenger or freight trains travelled within the state. This distance must be distinguished from the kilometers travelled by contract services. In addition, operators must report the number of passenger journeys in urban areas within Southern Australia given as 'a point to point journey irrespective of the number of vehicles or mode used for the trip'. Journeys in non-urban areas consist of 'a point to point journey but each change of vehicle along the route is a separate journey'. Companies must also report their total number of employees engaged in railway work in South Australia. This includes contractors and volunteers who work 'at the direction of the reporting railway' but not 'employees, contractors or volunteers of other accredited railway owners or operators who provide services to your organisation'.

Legal and regulatory provisions are not the only reasons why companies may monitor an incident reporting system and the data that it provides. There may also be strong commercial motivations. For instance, incident data is often cited when two or more companies are in competition for a particular market. Similarly, incident information will often be published if the operational activities of a company are called into question by the public, press, politicians or other pressure groups. For example, the San Jacinto Rail company is in the process of applying to transports hazardous and non-hazardous materials in the Houston area. Approximately 85% of the materials to be carried by the proposed service will be both solid and non-hazardous including polyethylene and polypropylene plastic resins. The proposals also provides for the transportation of more hazardous commodities, including isobutylene, propyleneglycol and ethylene glycol. In order to reassure potential opponents to this proposal, the rail company cited incident statistics gathered by a range of trade organisations:

> "According to research by the Association of American Railroads (AAR), 99.996% of hazardous materials moved by rail arrive at destination without incident. Over the past 20 years U.S. railroads have invested in technology and infrastructure to improve safety, reducing accidents per million train miles 66% since 1980 and 18% since 1990. Although trucks and railroads carry almost the same amount of hazardous materials, the trucking

industry has nearly 14 times more hazardous material incidents.” [724]

This quotation illustrates the way in which individual companies can draw upon the incident and accident data that is collected by trade associations. In this case, the San Jacinto proposal exploits the results of the AAR monitoring to compare the safety record of rail transportation with that of the haulage industry. The proponents of this scheme also cite incident data from similar companies. For example, the Burlington Northern and Santa Fe Railway experienced 17 releases of hazardous materials from a total of 878,428 shipments in 2000. The proposers calculate that this represents an ‘accident release ratio’ of 0.0194 accident releases per 1,000 shipments. This represents a decrease from 0.0201 accident releases in 1999.

The Nuclear Energy Institute provides a further example of how incident monitoring information can be used to support particular commercial operations [383]. A recent report stated that the US nuclear energy industry has completed more than 3,000 shipments of used nuclear fuel covering 1.7 million miles over the last 35 years without any injuries, fatalities or environmental damage from the radioactivity of the cargo. This is an interesting argument because it reverses the usual claim that reporting systems provide important information about previous failures. In contrast, the Nuclear Energy Institute stress the absence of incidents in order to reiterate the industry’s safety record. This style of analysis can seem complacent. The Institute is, however, careful to stress the more active safety measures that protect the public ‘accidents can happen and so scientists and engineers designed used nuclear fuel shipping containers to be among the safest on the road, and to protect the public against even the most unlikely accidents’ [383].

Regulatory and Investigatory Oversight. Governments are concerned to both promote and ensure

	Road	Rail	Water	Air
1991-92	2084	54	69	46
1992-93	1874	55	69	63
1993-94	1994	37	71	63
1994-95	1984	55	58	56
1995-96	1986	37	51	68
1996-97	1873	38	55	34
1997-98	1768	42	47	51
1998-99	1774	..	..	44
1999-00	1783	..	..	45
2000-01	1775	..	..	57

Table 15.1: Australian Transport Fatalities by Mode (1991-2001)

the safety of national industries. They, therefore, direct industry regulators to gather a range of statistics to monitor the performance of those industries. Some of these indicators are relatively easy to obtain. For instance, Table 15.1 presents Australian Transport Safety Bureau (ATSB) and Australian Bureau of Statistics data on fatalities in each of the major transport modes over the last decade [51]. This shows a reduction in the total number of fatalities across all modes except air transport. However, the periods in Table 15.1 reveal the lack of national, annual fatality statistics for particular industries. The lack of reliable statistics is worse for less serious incidents [51]. The Australian government has particular problems in gathering information about serious road injuries. This partly arises from the inconsistent definitions about what is and what is not reportable at a state level. Similar problems in the rail industry led to calls for a national coordinating body to receive and review incident and accident statistics [55]. The need for such a body is illustrated by the diverse legislation that covers the Australian national rail system. For example, Victoria follows a Transport (Rail Safety) Act of 1996 and enforces Transport (Rail Safety) Regulations proclaimed in 1998. Western Australia follows a Rail Safety Act of 1998 and Rail Safety Regulations of 1999 New South Wales introduced a Rail Safety Act in 1993. Section 44 of this act was affected by the Administrative Decisions Legislation Amendment Act of 1997. It also enforces Rail Safety (Offences) Regulations of 1997.

Collating statistics from different local and regional reporting systems is one of several problems that complicate the regulatory monitoring of particular industries. Chapter 1.3 has described the problems that arise when regulators become involved in both the promotion and monitoring of a safety-critical applications. There may be a temptation not to publicise adverse statistics that might affect the future success of commercial organisations. This explains why many countries deliberately separate the promotion and safety regulation of their industries. Even so, there is a temptation for regulators to focus on those statistics that illustrate the comparative safety of the industries that they support. It can be argued that increasing numbers of incidents and accidents reflect inadequate regulation as well as unsafe working practices within an industry. Many regulators are sensitive to these criticisms. The UK Health and Safety Executive's report in Signals passed at Danger (SPADs) reveals the tensions that exist when regulators monitor and publish incident information [351]. The document begins by stressing the relative safety of rail travel and overall improvements in the frequency of these incidents. In 1997-98, Her Majesty's Rail Inspectors (HMRI) received 593 reports of SPAD incidents across the UK rail network. This represented a reduction from the high of 944 incidents in 1991-92. This generally positive tone is balanced by the following paragraphs of the report which acknowledge that 'while such incidents continue to occur, there remains the possibility of one leading to a train collision and/or derailment'. They also note that the number of SPAD incidents increased to 643 in 1998-99, reversing the earlier downward trend. This careful balancing of positive and negative statistics continues throughout the report. It continues by noting that many of these 643 incidents do not threaten safety because the train stops within the 183 meter overlap to the signal which is the intended safety margin. These more positive comments are then balanced by the observation that potentially severe SPADS involving trains that run past the overlap and where there are connections ahead increased from 42 in 1997/98 to 52 in 1998/99.

The previous paragraph illustrated the *regulators dilemma*. Falling incident statistics illustrate the effectiveness of a regulator and their reporting system. However, by focusing on these figures there is a danger that the regulator may appear complacent in the face of any subsequent accidents. Conversely, rising incident statistics can be interpreted as the result of ineffective regulation even though they may indicate that sufficient information has been obtained to guide subsequent intervention. This dilemma can lead some regulators to stress the difficulty of intervening to prevent particular types of adverse events. For instance, there were 230 collisions at Canada's 22,400 public highway and railway crossings in 1998. Another 46 collisions occurred at private and farm crossings of railway lines. The National Safety Program, Direction2006, stresses that more than 50% of these incidents occurred at crossings that were equipped with automated warning devices such as flashing lights, bells and gates. There were a further 80 collisions involving trains and pedestrians. It is concluded that 'the fact that so many vehicles and pedestrians are involved in collisions with trains while either disobeying railway crossing signs and signals or trespassing on railway property underscores the need for increased enforcement' [213]. This response to the regulator's dilemma shows how adverse incident statistics can be used to justify different forms of intervention, such as 'increased enforcement' when existing measures appear to have failed. It remains to be seen whether this particular approach will have the intended effect.

It is important to stress that regulators, like companies and safety managers, often have several objectives for monitoring incident reporting systems. As we have seen, incident data can be tracked to identify areas for intervention or to monitor progress towards particular safety objectives. It is also important to monitor the performance of reporting systems and not simply the data that they produce. Regulators must account for their expenditure in terms of the 'productivity' of their reporting systems. For example, the annual report of the Chief Executive for New Zealand's Transport Accident Investigation Commission focuses on these metrics [630]. This account opens with the observation that the Commission launched 47 investigations, finalised 36 reports and promulgated 112 safety recommendations for a total cost of $1.588 million in 2000-2001. This represented an overspend of 0.1% beyond the Commission's income of $1.586 million.

The ATSB provides a further example of meta-level monitoring in which the performance of the reporting system is analysed as well as the individual incidents [51]. The 2001 annual report identifies a number of specific metrics that are to be used in assessing ATSB activities in the following twelve months. For example, one core activity was identified as the investigation of rail safety incidents to ‘identify circumstances and establish causes’. The annual report identifies quality, quantity and timeliness metrics. Quality can be assessed by ensuring ‘impartial investigations undertaken in accordance with relevant legislation/regulations and procedural guidelines’. The quantity criteria are set as ‘Findings published in up to 4 reports’. Timeliness metrics establish a median time of 27 weeks to complete investigations and finalise reports. Another key activity was to ‘Facilitate and publish rail safety statistical analysis and data collection to assist in the conduct of rail safety investigation and the development of policy and strategies’. The quality of this activity was to be assessed in terms of ‘user satisfaction with published statistical information’. The quantity criteria was again established as publishing 4 statistical reports. An associated comment noted that in the previous year; ‘work on a rail safety statistical database development continued but delay in agreement with state rail accreditation authorities delayed publication of data’. Similarly, a further core activity was to ‘publish and distribute rail safety reports’. The quality of this activity was to be assessed in terms of the acceptance and utilisation of rail safety reports by the rail industry. The plans include a commitment to publish the findings in up to 4 reports.

Political Monitoring. Politicians, typically, help to establish the regulatory structures that protect public safety. They, therefore, have a keen interest to ensure that monitoring data reflects the success of those structures. When evidence is presented about particular short-comings then there is often a rapid move to ensure that appropriate action is taken. For example, John Spellar, the Minister for Transport, recently told a rail industry safety conference that government and industry must act to reduce the 300 deaths from trespass and suicide on UK railways each year. He argued that over half of these incidents were due to malicious acts of criminal damage and that it was, therefore, necessary to introduce a coherent ‘national strategy’ to address the problem [219]. The political sensitivity over incident data also partly explains regulators’ concerns to both justify their intervention and to account for their expenditure on reporting systems. Ultimately, accidents can lead the general public to question the political structures that guide the development of safety policies at a national and an international level. For example, the Indian Government of Atal Bihari Vajpayee ordered a complete review of their national rail system in the aftermath of the Gaisal train collision in which almost 300 people died in 1999. He refused to accept the initial offer of resignation from his railways minister, Nitish Kumar, who said ‘he felt the need to punish himself for the huge loss of life’ [101]. It is instructive to note that this political reaction was triggered in spite of a relatively good safety performance across the Indian rail network. In 1997, there were 1.4 passenger deaths for each billion passenger-kilometres travelled in India compared to 1.42 in the European Union.

The events surrounding UK rail privatisation provide a more complex example of the role that politics play in monitoring the performance of incident reporting systems. The break-up of British Rail, the national rail service, was proposed throughout the late 1980’s but only emerged as a commitment in the 1992 Conservative election manifesto. A White Paper on rail privatisation was then produced following their victory under John Major in July 1992. Pressure from the Treasury resulted in a decision to separate the operation of the infrastructure from that of rail services. This led to the creation of twenty-five separate companies, including Railtrack which assumed responsibility for the rail infrastructure. The Bill was finally passed in November 1993. The first operating franchises were offered in December 1995 to SouthWest trains, LTS and Great Western. Railtrack was floated in 1996. The final operating franchise was offered to ScotRail in April 1997 shortly before the Labour party was elected to power. There then followed a succession of high-profile failures including accidents at Watford Junction in August 1996, Southall in September 1997, Ladbroke Grove in October 1999, Hatfield, in October 2000 and at Selby in February 2001.

These ‘failures’ helped to launch a series of enquiries and investigations that considered the monitoring of incidents and accidents as part of a wider review of rail safety in the UK. Previous paragraphs have cited from the Cullen report into Ladbroke Grove and the Health and Safety Executive’s report into Signals Passed at Danger. Rather than reiterate the findings of these investigations, it is also important to consider the political impact of these initiatives to monitor both incident information and the reporting systems that produce them. Don Foster, the Liberal Democrat transport spokesman, reviewed these statistics during a Commons debate into transport safety. He concluded that the way in which the Conservative government had introduced privatisation had created 'confusion between safety and other aspects of the railway - not least confusion between safety and profit' [99]. The privatisation process had created uncertainty about who was responsible for what happened when an accident occurred. Labour's junior transport minister, Keith Hill, responded by arguing that both public and private transport operators must make safety their first priority; it is 'totally unacceptable for financial interests to take precedence over safety'. His response also illustrates the way in which narrow discussions about the safety record of particular companies can be broadened by political debate. He was compelled to defend plans for the 'part-privatisation' of the London Underground and for National Air Traffic Services (NATS) in a debate on the Ladbroke Grove rail crash. The political nature of such incidents is again illustrated by the Conservative spokesperson, Shaun Woodward, who argued that 'the public not only wants us to be concerned about safety but to ensure that when we know that safety may be at risk, to take responsibility and action when and where appropriate' [99].

The Hatfield accident, in particular, helped to focus attention on the high levels of investment that were necessary to achieve acceptable safety standards throughout the UK rail infrastructure. Political pressures ultimately forced the Labour government to withdraw financial support from Railtrack. The infrastructure company was then taken into administration. This political decision had both financial and operational implications. Over 250,000 shareholders, which included 90% of the company's employees, were immediately affected by this decision. The withdrawal of government financial support for Railtrack also cast considerable uncertainty over the future of the UK rail network. In the aftermath of this action, the percentage of trains arriving 5 or more minutes late increased from approximately 25% to over 30%. Although these figures were subsequently challenged on the grounds that they reflect a seasonal increase in delays from adverse weather conditions and 'leaves on the line'. The decision also raised safety concerns that demoralised employees facing an uncertain future might exacerbate existing equipment and infrastructure problems to trigger an increase in adverse events. Hence the political intervention directly led to a request from the Department of Transport, Local Government and the Regions to the Health and safety Executive to increase their monitoring of rail incident data to ensure that the administrative procedures had not jeopardised the safety of the rail system. It remains to be seen whether the incident data will reveal the same adverse trends that many have claimed for reliability statistics.

Political interest in the data that can be obtained from monitoring reporting systems does not just focus on the need to counter potential criticisms of particular initiatives. Statistical evidence of falling reporting rates is often used to validate previous actions. It can also be used to publicise and promote a reporting system. This may indirectly increase confidence in the wider regulatory systems that protect the public. For instance, in 1999 the U.S. Transportation Secretary and Federal Railroad Administrator announced the publication of a report showing 'dramatic' improvements in railroad safety as a result of the Clinton administration's partnership with industry. In 1997-98 there was a 27% reduction in railroad employee fatalities and a 33% reduction in passenger fatalities. Highway-rail incidents declined 9% and highway-rail injuries 15%. The FRA also reported a 'dramatic' fall in six-year safety results. From 1993 to 1998, highway-rail incidents declined 28%, highway-rail fatalities 31% and highway-rail injuries 29% while railroad operations, measured in train miles, increased 11%. It was argued that the Clinton partnerships supported safety improvements by focusing attention on the 'root causes of problems' and an improved understanding of 'the nature of rail-related incidents'. The Transportation Secretary stated that:

> "'President Clinton and Vice President Gore challenged the government to do business in a new way, to work better together and get results that Americans care about. The report we are issuing today demonstrates that this approach to governing is working by dramatically increasing safety in the railroad industry." [240]

As we shall see, however, headline figures can mask other incident statistics that often contradict political claims about the safety of an industry. Closer inspection of the FRA monitoring data shows that the overall fall in accidents and incidents was largely accounted for by the drop in highway-railway incidents from 3,865 (1997) to 3,508 (1998) The same period saw an increase in train accidents from 2,397 (1997) to 2,575 (1998) mainly caused by derailments, 1,741 (1997) and 1,757 (1998), and human factors, 855 (1997) and 971 (1998) [245].

Media and Public Involvement. There are clear reasons why those who are involved in the operation and management of a reporting system should want to monitor both its output and performance. It is also important to recognise that there may be other parties, including trade associations and public pressure groups, who have an interest in tracking this information. Many of these groups have indirect access to incident information. The US Freedom of Information Act has helped to ensure that many Federal agencies provide incident information over the web. The provisions of this act have had numerous benefits. For example, much of the recent research on novel computational techniques for incident retrieval has been directly driven by these new information sources [416]. It would not have been possible to write this book ten or even five years ago when there was little or no access to such confidential databases. Even where direct access is denied, pressure groups can monitor reporting systems indirectly through official press releases and less authoritative leaks to the media. For example, the BBC reported that ScotRail were one of '10 train companies warned by the Railways Inspectorate that it was not doing enough to prevent drivers passing red lights' [112]. HMRI's figures showed 56 SPADs in May 2001 compared to only 35 in May 2000 and an average of 49 SPADs between 1995-2001. The Railway Inspectorate warned operators that they would face enforcement actions and prosecutions if their safety records did not improve. The report goes on to explain that these criticisms were triggered because the number of SPADs had *improved* but only slightly. Media organisations do not always follow the balanced approach illustrated by this example. It is also important to recognise that concerns over this publicity encouraged ScotRail to directly counter criticisms in the BBC report. The following quotation presents the response of a ScotRail spokesman to the publication of the HMRI figures. The confidential reporting system is the CIRAS scheme that has been mentioned in previous chapters and will be discussed in later sections of this chapter:

> 'I think the figures they have surround the long term average rather than last year's results. Last year we had a 22% reduction, which was better than the national average. It is a subject that is taken very seriously. It is obviously very high up our agenda. We put a great deal of effort into it and we have led in the past on man y new initiatives, including the confidential reporting system.' [112].

These comments elicited a sympathetic response from passenger 'pressure' groups. The Deputy Secretary of the Rail Passengers Committee for Scotland acknowledged that the number of SPAD incidents had fallen since rail privatisation. He also referred to initiatives by companies, such as ScotRail's defensive driver techniques, that had helped to reduce these adverse events.

This example illustrates the diverse groups that are concerned to monitor data from incident reporting systems. The SPAD frequency information was initially released by HMRI. This investigatory and regulatory organisation is primarily responsible for controlling the hazards that affect the health and safety of anyone who might be affected by the operation of Britain's railways. The BBC then identified the information as having a wide public interest. This media organisation then commissioned a report which elicited responses from the companies concerned. They countered the HMRI's interpretation of the statistics by pointing to longer term trends. Finally, a passenger group responded to ScotRail's defence of their safety record. Regulators, investigatory organisations, the media, commercial organisation and public pressure groups all contributed to the analysis of information that was initially obtained from the SPAD reporting system. Such diverse opinions illustrate the difficulty of interpreting such statistics. Many of these problems stem from the paradoxes of incident monitoring.

Paradoxes of Incident Monitoring

It can be argued that the monitoring of incident reporting systems should ensure that they help to avoid future incidents and accidents. Chapter has, however, argued that we cannot achieve absolute safety [677]. It is also important to emphasise that incident reporting systems do not operate in isolation from the rest of an organisation. A new scheme might be introduced at the same time as new processes and plant come on-line. Hence, the introduction of the reporting system may coincide with a notable increase in adverse events. **First paradox of incident monitoring:** even if a reporting system does not demonstrate a long term reduction in adverse events it can still be argued that the safety record would have been even worse if the reporting system had not been in place.

Given that we cannot achieve absolute safety, it is important to shown that the level of investment in a reporting system yields an optimal reduction in adverse events. In other words, we would like to demonstrate that additional investment would provide little additional safety information. Conversely, we might also demonstrate that savings could not be made without jeopardising important feedback about the safety of the system. Unfortunately, a number of problems complicate the task of assessing the marginal utility of investments in incident reporting systems. In particular, there is an important distinction between the numbers of incidents that occur and the numbers of submissions made to a reporting system. Chapter 4.3 has described how increased levels of funding typically elicit additional submission. **Second paradox of incident monitoring:** additional funding for incident reporting systems typically yields an increasing number of submissions as people become more aware of the system. This need not reflect a rise in the underlying number of incidents. Conversely, cuts in the funding associated with a reporting system may yield fewer contributions but this need not indicate an improvement in the underlying safety of an application. Staff may be disillusioned with the effectiveness of the reporting system.

The monitoring of incident reporting systems is further complicated by the argument that in any resource limited environment, we cannot simply consider the costs of any particular activity in isolation. In contrast, it is important to assess the opportunity cost associated with maintaining an incident reporting system. This focuses on those activities that must be sacrificed in order to support a reporting scheme. I would stress the importance of this perspective given that the individuals who help to establish and maintain reporting systems are often amongst the most highly-trained and safety conscious staff within an organisation. These individuals are often so committed to the operation of a scheme that few seem to consider whether their time and energy might not be more effectively employed in other safety-reelated tasks. **Third paradox of incident monitoring:** those individuals who are most committed to the operation and maintenance of a reporting system may not be in the best position to judge whether or not these schemes make the most use of their finite resources.

The previous paragraphs have focussed narrowly on safety improvements as the principle benefit of operating an incident reporting system. As we have seen, however, there are many other reasons why one of these schemes might be established. Regulators might require operators to support a reporting system. Reporting systems can be introduced to deflect criticism of previous safety related failures. These schemes can also be introduced to form part of a wider 'lessons learned' or quality assurance scheme. In such circumstances, safety benefits form part of wider improvements in operating practices. **Fourth paradox of incident monitoring:** incident reporting systems may continue to be maintained even though almost no safety-related contributions are submitted. The rationale for operating the system need not rely narrowly upon safety-related issues but may have more to do with wider operational and regulatory concerns.

These paradoxes make it difficult to interpret the results of any attempts to monitor the success or failure of a reporting system. For instance, a fall in the number of incidents reported might indicate disillusionment with the system, problems in submitting report forms or a genuine reduction in safety-related incidents. The difficulty of interpreting particular measures has led some organisations to adopt a broader perspective. In particular, they have sought metrics that might be used to validate the diverse range of proposed benefits from incident reporting that were enumerated in Chapter 1.3. For example, the following list extends the results of a study by the US Coast Guard [835] to identify ways of monitoring the health of their reporting systems:

- *Number of submissions received.* This is often the most convenient means of monitoring participation in a reporting system. As we have seen, however, it can be difficult to interpret the results. Low submission rates may indicate safety improvements or disinterest in the system. Similarly, increases in participation may stem from the expansion of an industry as more groups are exposed to potential hazards. There are further dangers. For instance, long-running schemes often publicise their success by reiterating the cumulative total of reports received. If one looks more closely into the nature of the reports received, it is often depressing to find that the same sorts of failures have been submitted often over decades [411]. Hence a high cumulative total and high annual participation rates may indicate the limitations of the approach rather any measure of success.

- *Change in the quality of submissions.* Rather than focusing on the raw numbers of submissions, the success of a reporting system can be assessed in terms of the quality of those submissions. This can provide feedback on whether or not potential participants can understand and follow reporting procedures. At first sight, it may be argued that such measures provides relatively little information about the safety of an underlying application. Given the problems associated with measuring the frequency of 'near misses' this approach can, however, help to minimise any potential barriers that might otherwise prevent individuals from submitting information about such events. In consequence, it can be argued that if the quality of submissions improves then we can have greater confidence in the accuracy of reporting frequencies. A high number of apparently spurious submissions might lead to some potentially valid incidents being discarded during any initial filtering. If a contributor fails to provide sufficient information about a potential incident then analysts may be forced to invest scarce resources in collecting sufficient initial information to justify subsequent investigation. In many systems, the decision may be made not to invest those resources so that additional attention can be paid to more 'clear-cut' incidents. Similarly, a high number of spurious submissions might indicate that some 'valid' incidents are not being reported because of general confusion about the purpose of the scheme.

- *Percentage of attributable reports in an anonymised system.* In systems that offer contributors the possibility of filing a report without disclosing their identity, the proportion of submissions that include contact information can provide a measure of confidence in the system. This measure can provide indirect insights into participation levels. A high proportion of unattributable reports might indicate general skepticism about the integrity and potential benefits of the system. These concerns are likely to jeopardise participation in the system. It can, therefore, be argued that the number of incidents being reported to the system is unlikely to provide an accurate impression of the total number of adverse events and 'near misses'. Conversely, a high level of attributable submissions may indicate high levels of participation. This, in turn, can increase confidence that 'near miss' incidents are being submitted and that contributions provide a more accurate measure of underlying safety.

- *Number of submissions investigated.* Chapter 9.3 has argued that participation in many reporting systems depends upon organisations acting on the information that they receive. In this view, the effectiveness of any reporting system cannot be measured simply by the number of submissions that are made. Participation levels are unlikely to be sustained if no actions are taken to investigate the safety concerns that are identified by contributors. This view is significant because it emphasises the idea that the 'health' of a reporting system will change over time. The future success of the system may, therefore, depend partly on current submission rates and partly on the way in which the system reponds to those contributions.

- *Number of reports leading to safety improvements.* The caveat that there must be some demonstrable recommendation or action taken in response to a report is significant because otherwise a rise in submissions might reflect an increase in spurious reports. The numbers of reports that trigger interventions not only provides a measure of the effectiveness of any system, this information can also be used to encourage further participation. Such information can demonstrate that reports will be acted upon. The 'safety improvements' that are derived from a reporting system can be interpreted quite broadly. For instance, the UK Rail Safety group monitors the number of enforcement actions that HMRI initiates against operating and infrastructure companies. 30 notices were issued in the second quarter of 2001/02 bring the six month total to 43 notices. This can be compared to only 31 notices for the whole of 2000-2001 [692]. It can, however, be difficult to draw firm conclusions from these figures. Not all enforcement actions are triggered by incident reports. Conversely, not all incident reports that identify potential violations will lead to enforcement actions. Similarly, a rise in the number of enforcement actions can be the result of short term initiatives by investigatory agencies rather than the result of short-term increases in the number of adverse events. For instance, the increase in enforcement actions in 2001-2002 was partly the result of actions to reduce the frequency of trespass and vandalism.

- *The number of reports submitted by particular categories of participants.* For example, the success of a reporting system might be measured for particular industry segments, regional areas, professions or staff positions. Such measures are important because most successful reporting systems achieve safety improvements in spite of 'uneven' levels of participation. For example, the Aviation Safety Reporting System gathers very few reports from Military pilots and a limited number from General Aviation. The FDA's MEDWATCH program receives proportionately less reports from nursing homes than it does from larger hospitals. In rail reporting systems, there are few reports of 'Signals Passed At Danger' in remote regions where there are few witnesses to any infringement. In such circumstances, the health of a reporting system may be judged against participation targets for particular groups of participant.

- *Change in the number of accidents.* As mentioned previously, changes in the number of submissions to a reporting system can be the result of other events that have little to do with the underlying safety of any application. It can also be difficult to gather accurate statistics about the occurrence rates for 'near miss' incidents. In consequence, the only reliable safety measure is the number of accidents within an industry. For instance, the Cullen report argued that several major accidents indicated significant flaws in existing reporting practices within the UK rail industry. As we shall see, however, structural changes in this industry created new hazards and placed new demands on the existing reporting infrastructure. This emphasises the importance of continually monitoring the performance of a reporting system against such measures. A reporting system may provide adequate information about potential hazards within one context of operation but may be ineffective in identifying potential hazards as changes occur within an industry. Further problems arise because the low frequency of accidents in many industries can prevent reliable inferences being made about the underlying safety of an application until an adverse event occurs. It can also be difficult to define what constitutes an accident. Some injuries and fatalities, for instance from suicide or trespass, are difficult for operating companies to control. The practical problems of calculating an accident rate as a means of assessing the performance of incident reporting systems can be illustrated by the UK Railway Safety Group's quarterly reviews [692]. This calculates the risk of a train accident for the previous quarter by combining the frequencies of particular contributory factors, including level crossing mis-use, irregular working activity and vandalism. The complexity of using such measures to assess overall 'safety' is illustrated by the October 2001 report. This recorded a slight increase in the accident risk even though the number of 'significant train accidents' actually fell. This apparent paradox can be explained by a rise in workforce fatalities. There are further complications. The incidence of track quality faults, wrong-side signal failures and train speeding reduced fell but the number of public accidental fatalities rose compared to in the first quarter. The practical difficulties in compiling accident statistics are exacerbated by ethical objections. Arguably the most significant criticism of accident metrics is that the performance of a reporting system is assessed in terms of the number of times it fails to protect either the workforce or the general population.

- *Change in the number of particular event types.* It may not be possible to gain an accurate assessment of the overall number of 'near miss' incidents and accidents across an industry. The problems of under-reporting and reporting bias frustrate attempts to gather such statistics. These general problems can be addressed by focusing on particular types of adverse event. Additional publicity can be provided to explain the importance of certain hazards. Automated monitoring and logging systems can be used to detect when such events have occurred. The results of these special initiatives can be compared against levels of participation to provide a measure of any previous under-reporting. Unfortunately, the effectiveness of these techniques may decline if they are used too frequently [411]. Participants may becomes immune to successive attempts to sensitise them towards particular types of failure.

- *Changes in the safety issues identified.* A danger with any reporting system is that it will continue to identify the same safety concerns that have been observed in previous incident reports. If participants continue to reiterate well known issues then it might be argued that the reporting system is ineffective as a means of addressing those issues. This view can be challenged. For instance, there may be agreement over the nature of the problem but disagreement over the recommendations proposed by incident investigators. For instance, Chapter 8.3 described how the National Transportation Safety Board (NTSB) struggled to introduce devices that were intended to address gas leaks and explosions that occurred over almost three decades. Industry representatives argued that the costs associated with such changes would not be justified by any potential benefits. It can, therefore, be argued that the continuing pattern of incidents did not simply reflect the failure of the reporting systems. Instead, it indicated the difficulty of resolving complex commercial and regulatory issues and the need to build up a body of evidence in support of the investigators' recommendations.

- *Changes in outcomes.* The success of a reporting system might be measured at a gross level in terms of a reduction of the total working days lost to industrial injuries. Similarly, it might be measured in terms of any change in particular types of injury or fatality. For instance, Table 15.2 presents five-year trend data on rail fatalities and serious injuries in Australia. These are categorised according to individual regions. Unfortunately, it can be difficult to obtain such outcome information. Some of the values in Table 15.2 denoted by the periods have been suppressed because of State privacy restraints. Although this data was published as part of a national report on transportation safety, the statistics had to be pieced together from the Australian Bureau of Statistics and the Australian Institute of Health and Welfare. These were the only sources of national rail safety data available in the absence of a national rail occurrence database.

Fatalities

	NSW	Vic.	Qld	SA	WA	Tasmania	NT	ACT	Aust.
July 1993 - June 1994	10	8	10	4	5	0	0	0	37
July 1997 - June 1998	22	12	2	1	5	0	0	0	42

Serious injuries

	NSW	Vic.	Qld	SA	WA	Tasmania	NT	ACT	Aust.
July 1993 - June 1994	80	22	24	7	10	2	0	0	145
July 1997 - June 1998	66	18	19	..	13	3	0	2	..

Table 15.2: Rail Incident Outcomes on Australian Railways (1994-1998)

Outcome measures can also be used to assess the performance of incident reporting systems in individual companies. A rising number of serious injuries might be interpreted as a failure to learn from previous incidents. Again, however, there is considerable concern over the reliability of this approach [340]. For example, it can be argued that the outcomes might have been even worse if the reporting system had not been in place. It can also be difficult to identify suitable outcome measures that might be used to assess the performance of individual firms. The outcome of an adverse event can be mitigated by the prompt intervention of medical staff. Conversely, the eventual outcome of some incidents may take many years to fully develop. There are some industries, in particular those that depend on self-employment labour, for which it has always been difficult to obtain accurate consequence statistics. Further concerns stem from the difficult of accounting for near misses with high-potential consequences. For example, no-one was killed or fatally injured by a main track derailment on Canadian railways between 1983 and 1996. During that time, there were approximately 10 derailments per year from bearing failure alone [777]. It can also be difficult to distinguish the impact of a reporting system on any changes in consequence figures. It is for this reason that the Transport Canada requires rail operators to monitor the performance of their reporting systems, in terms of the types of failure and remedial actions, as well as employee deaths, disabling injuries and minor injuries per 200,000 hours worked [780].

- *Survey results from user groups.* Given the problems associated with deriving accurate measures from either the submission rate to a reporting system and the ethical issues associated with post-hoc accident rates, it is important to find other means of assessing the effectiveness of these initiatives. Given that many reporting systems have identified failures in ill-defined concepts such as 'safety culture', it can be argued that such schemes are successful if they act to change those previous weaknesses. These schemes remind participants of previous incidents within their industry and hence can play a positive role in informing people about the potential adverse consequences of particular incidents. This 'consciousness raising' effect can be assessed by surveys of the groups who participate in a reporting system. This approach recognises that a far larger group may benefit from the publications produced by reporting system that the comparatively small number of individuals who might actually witness an adverse event and then submit a report.

- *Change in insurance premiums.* The previous measures focus on attributes of reporting systems or on the applications that they are intended to protect. It can be difficult to gather accurate figures for these direct measures. It can also be difficult to interpret what changes in these measures imply for the safety of an application. Some organisations, therefore, emphasise the indirect benefits of incident reporting systems. These include reductions in insurance premiums associated with safety-critical applications. Such 'metrics' are credible because they typically reflect the judgement of an external organisation that is strongly motivated to provide an accurate risk assessment.

- *Changes in application operating costs.* Incident reporting systems are often integrated into more general systems for quality control. Improvements in operating efficiency are often more easily measured that any improvements in safety. For example, the relatively low frequency of many safety-related events can imply that individual units will only receive a few submissions each year. It is, therefore, impossible to judge the relative success of a system from month to month. In such circumstances, organisations are often motivated to increase the scope of a 'lessons learned' system. It is hoped that by reporting lower consequence failures, potential participants will be more comfortable with the procedures that support the submission of safety-related events. It follows that even if no safety information is submitted to the system, the provision of information about other potential problems in quality control or efficiency can provide feedback about the effectiveness of the reporting system.

- *Change in the operating cost of the reporting system.* Incident reporting systems do not operate in a commercial vacuum. In consequence, many systems are assessed according to the usual financial criteria associated with any management or engineering function. Unfortunately, the problems in deriving objective measures for the success of a reporting system can make these schemes vulnerable to cost cutting. It can be difficult for safety managers to prove that cuts in the funding of a reporting scheme will jeopardise the safety of application processes. Similarly, a large increase in the number of reports processed at the same level of funding raises questions about the level of analysis that can be sustained for any particular safety issue. Such savings can be justified through increased efficiency, for instance by the introduction of information technology. Alternatively, more accurate forms of risk assessment can be used to ensure that reduced funding does not impair the organisation's response to high-criticality incidents.

- *Establishment of a self-sustaining operation.* The success of some reporting systems is measured against particular commercial or financial criteria. Increasingly, there is a view that these systems should be self-sustaining and should not be sustained by public money. The industries that benefit from the insights obtained by a reporting system should meet the costs associated with maintaining the system. This creates potential concerns. For instance, if some companies 'opt out' of the scheme then they may be isolated from any insights provided by the system. If companies are forced by the regulator to join the scheme then this can be interpreted as undue interference in the commercial operation of particular industries, especially if the regulator retains an interest in the maintenance of the scheme. Conversely, if a commercial cartel retains control of the reporting system there is a danger that the independence system can be compromised. Investigators may be unwilling to propose recommendations that have high cost implications for the rest of the industry.

- *Changes in the mode of submission.* Given the difficulties of obtaining and interpreting objective measures for any safety improvements derived from a reporting system, it is often more convenient to identify more focused objectives that relate to the way in which a particular scheme is implemented. This class of measures are often associated with the financial objectives, summarised above. For instance, the success of a reporting system can be assessed in terms of particular modes of submission. Several of the schemes described in this book have moved away from paper based submission towards telephone, fax and Internet based contributions [423]. These initiatives are intended to increase the scope of a system by cutting the costs associated with managing the collation of individual reports. The use of these metrics indicates the complexity of monitoring incident reporting systems. These changes can introduce new biases into the reporting process, it may be harder for some participants to access and use new submission techniques. By achieving particular objectives for the introduction of new technology, the reporting system may lose important safety-related information. This may, however, only be a short-term effect as more people learn how to operate the revised submission procedures. The reduced costs associated with alternative modes of submission may be necessary to support the long-term survival of the system.

- *Number of information requests.* Previous measures have focussed on the number of incidents reported or the number of investigations that have been completed. The success of a reporting system can also be assessed in terms of the information that it disseminates. Gathering information about previous failures is of little benefit if any insights are not passed on to those who are best placed to use them. For this reason, the success of a reporting system might be measured in terms of the number of information requests that are received. As with the submission metrics, more fine-grained targets might also be associated for requests from particular end-user groups within particular industries or regions.

- *Time to implement changes from first notification.* A number of temporal properties of incident reporting systems can be measured to provide insights into their efficiency in dealing with particular safety-related concerns. For example, it is possible to record the time between a request being made for incident information and that request being addressed. Such intervals are significant because any delay might compromise the safety of application processes. Similarly, the time between an initial notification and any secondary investigation could be measured to provide information about the response to a report. This is a significant concern given that safety managers have found completed report forms that have lain neglected for many months in the desks of process supervisors. These metrics, typically, introduce additional administrative overheads in terms of the resources that are required to log timing information. They are, therefore, most frequently gathered by large, distributed systems such as national Air Traffic Management reporting schemes [423].

- *Number of publications issued and acted upon.* If reporting systems disseminate most of their information through paper-based publications it can be difficult to gain a true measure of all of the individuals and organisation who may read and act upon the information that is disseminated. Each journal or bulletin can be read by several people. Conversely, there is no guarantee that the recipients of a publication from a reporting system will have actually read the information that it contains. Readership surveys provide one means of addressing these problems. Alternatively, the 'productivity' of a reporting system might be measured in terms of the raw number of publications that it produces. Unfortunately, such measures do not discriminate between active systems that continually provide new insights and those that regularly publish the same advice without seeking new remedies for past and present failures. These publication measures might be supplemented by an assessment of the other 'peripheral' activities that often provide alternative means of dissemination. Conference presentations and workshops can provide a further indication of the health of a reporting system.

- *Number of people who access computer-based resources.* It can be difficult to track all of the people who have access to the paper-based publications that are produced by an incident reporting system. Some of these problems can be addressed through the provision of computer-based resources that automatically log any requests for information. The metrics provided by these systems can help to justify any investment in computer-based resources. In particular, it is important to demonstrate that the use of electronic dissemination techniques does not hinder access to information about previous incidents. Automated logging facilities can be used to provide profile information based on the Internet Protocol address of sites that request access to the system. These addresses uniquely identify the computer that sent the request. More accurately, they identify a connection between that computer and the network because a single machine might have several network connections. In practice, however, the allocation of IP addresses to sites and the local routing of requests can limit the inferences that are made. Automated logging can provide other metrics. For instance, it is possible to identify the number of abandoned or failed requests made to a web server. This provides useful information about retrieval delays. If the number of abandoned requests rises then it may be necessary to index the data in another way or to provide additional support for the system infrastructure.

- *Changes in industry/operator participation.* Many of the proponents of incident reporting have argued that active participation from industry is required in order for these systems to be successful [845]. The imposition of reporting systems by regulatory intervention can lead to resentment and the creation of informal barriers that may discourage submissions from some employees. In contrast, the enthusiastic promotion of a reporting system can encourage participation and support the dissemination of safety-related information. In consequence, many regulators publish lists of the companies that have chosen to 'sign up' to a scheme. These lists provide a gross indication of industry participation. They provide little indication of the financial and organisational resources that each company is prepared to allocate to a reporting system. For instance, many hospitals have established incident reporting systems as a means of combating negligence claims. Many of these institutions provide limited budgets and appoint relatively junior staff to manage these schemes. In contrast, some hospitals have ensured that clinical risk managers are promoted to the highest levels within their organisational structure. Such differences make it difficult to derive accurate measures for the level of participation in incident reporting systems.

- *Levels of information sharing between companies.* Incident reporting systems have often been established with the claim that they will improve the dissemination of safety-related information between the participants in the scheme. It is, therefore, appropriate to consider how such information exchanges might be measured as a means of validating these claimed benefits. Chapter 4.3 has described how the creation of such systems can only have a limited effect on the barriers that prevent the effective dissemination of safety information. In consequence, many companies will operate their own internal schemes in parallel with industry-wide systems. This tends to ensure that only some incidents are shared in the manner proposed by the proponents of incident reporting systems. It would be very revealing to measure the differences between those incidents that are retained within a proprietary system and those that are shared in an industry wide scheme. Such initiatives would have to address the same barriers that prevent the exchange of information in the first place.

- *'Collateral' effects on industry.* Direct measures can be found for the impact that reporting systems have upon particular industries. Large numbers of similar incidents can trigger external regulatory intervention to enforce the recommendations that are made within an incident reporting system. Some health and safety organisations judge their success in terms of the numbers of prosecutions that are initiated in response to reports of adverse events. The information that is collected about 'near miss' incidents is often cited in legislative changes. In extreme cases, such reports can motivate government intervention to restructure an entire industry. The reorganisation of the UK rail infrastructure provides an example of such intervention [195]. Many of these changes cannot be initiated from within the incident reporting system itself, the influence of such schemes therefore extends well beyond those who are directly involved in operating the system.

- *Tracking of public image.* Lough has recently argued that the success of any reporting system should be measured in terms of its acceptability both by those who participate in the system and by the wider community in which it operates; "acceptability is often a reflection of high validity" [501]. His use of the term 'communittee' is interesting because it can refer to a 'communittee of practice'. His work focuses on techniques to support incident investigation by medical doctors in general practice. The term might also refer to the wider 'communittee' which includes the general public. This ambiguity is important because it identifies a dual role for incident reporting systems. On one level they can be used to derive particular insights that may prevent the recurrence of safety-related incidents. At another level, these systems act as an important means of reassuring the public that application processes are being operated in a responsible manner. This may, in part, explain why so many incident reporting systems have been established in the aftermath of major accidents. Such high-profile failures affect public confidence. Incident reporting systems satisfy their expectation that government and regulators should do something to address their safety concerns.

- *Longevity and 'technology transfer'.* The ultimate success of a reporting system can be measured in terms of its longevity. For instance, the US Aviation Safety Reporting System has continued in operation since 1976. The fact that it has survived through many changes in the fortunes both of the aviation industry and its sponsoring organisations demonstrates the perceived success of this system. The Australian Incident Monitoring System (AIMS) has a similar 'track record' within the field of patient safety, stemming from an anaesthesia project in 1989. Both of these applications have provided templates for subsequent systems. For instance, the operators ASRS argue that the success for their system led to the UK's Confidential Human Incident Reporting Program (1982), the Canadian SECURITAS system (1995), the Australian Confidential Aviation Incident Reporting system 1988, the Russian Voluntary Aviation Reporting System (1992), the Taiwan Confidential Aviation Reporting Enterprise (2000) and the Korean Confidential Aviation Incident Reporting System (2000) [60]. Similarly, the Australian Pateint Safety Foundation (APSF) which helps to administer the AIMS application has inspired the UK National Patient Safety Agency (2001) and the US National Patient Safety Foundation
indexNPSF (1998) both of which are closely involved in medical incident reporting. Imitation might provide the greatest evidence for the success of particular reporting systems.

Previous paragraphs summarise the vast range of metrics that have been proposed to support the monitoring of incident reporting schemes. Unfortunately, the strengths and weaknesses of these various measures have not been established. For example, there is no evidence to support criticisms against raw submission numbers as an indicator of the contribution to system safety. This lack of evidence is unsurprising. The relatively low frequency of accidents prevents analysts from forming the causal connections that might support statistical correlations. We might like to establish that organisations with a low number of submissions also suffer from a higher frequency of more serious accidents. Things are not so straightforward. For instance, several of the UK's rail operating companies with the best reporting record have also experiences significant safety-related problems [419]. The difficulty of establishing measures that relate incident reporting behaviour to accident frequencies is further illustrated by Wright's [875] recent work on the Heinrich ratio, summarised in Chapter 1.3. She argues that railway workers are, typically, either involved in fatalities or are witnesses to 'near-misses' [875]. There are few reports in the middle ground of more serious, non-fatal incidents. If her analysis is correct then we cannot expect there to be any clear-cut relationship between submissions and accident frequencies.

It is difficult to conduct controlled experiments in this area. Several of the metrics proposed in the previous list can be influenced by local effects. For example, the support that an organisation provides for participation in a system can be affected by the behaviour of individual managers. One could envisage a trial which compared the influence that different supervisors had upon the reporting behaviour of their workforce. It is difficult to see how such influences could be distinguished from the mass of other local factors that might also affect reporting behaviour. These include the composition of work groups as well as the submission and reporting processes that operate in individual plants. Ethical problems also complicate work in this area. If participants are informed that they are being studied then this may affect their participation in the system. Conversely, post hoc studies can compromise the confidentiality of the reporting system if they associate particular reports with particular working groups.

The lack of direct evidence to support particular metrics reflects the wider lack of research to support many other aspects of incident reporting. Considerable resources have been devoted to support the design of safety-critical systems. Far less resources have been allocated to understand why these systems fail. In consequence, the development of incident reporting systems resembles a craft skill rather than an engineering discipline. Techniques are borrowed from other systems that are perceived to be successful. Often metrics are chosen because they either validate the allocation of resources to maintain the system or because they have been used to assess other similar systems. In many cases, there is also an unquestioning assumption that incident reporting systems are 'a good thing' hence it is largely irrelevant to look for more quantitative forms of support.

With these comments in mind, the following pages focus on a number of the measures proposed in this opening section. The analysis is grouped into three parts. The following section looks in more detail at the reasons why it is important to monitor the outcomes of incident reporting. In particular, we focus on the role that these systems play in risk assessment, in systems development, in training and in operational efficiency. The subsequent section justifies attempts to monitor the reporting process itself. Particular attention is paid to changes in submission rates, to the behaviour of investigators and to the implementation of proposed changes. The closing sections of this chapter present a range of techniques that can be used to implement the metrics that are identified in the previous sections. These range from the use of computer-based audits to monitor the behaviour of incident investigators through to observational studies of the working groups that submit incident reports in the first place.

15.1 Outcome Measures

Heinrich's pioneering work in the area of safety management identified a number of tasks that safety managers must perform if incident and accident data is to inform the future operation of application processes [342].

1. collect incident and accident data;
2. analyse the data;
3. select appropriate remedies;
4. implement those remedies;
5. evaluate effectiveness of any remedies.

This approach can be criticised because it does not explicitly 'close the loop'. In other words, it is implicit that the evaluation of any remedies will help to inform the selection of future interventions. Similarly, the evaluation process might itself help to inform or direct the elicitation of incident data. Kjellen addresses some of these limitations when he argues that the monitoring of an incident reporting systems must help to identify the need for further information as well as identify priorities for intervention [444]. This iterative approach suggests means of monitoring the effectiveness of an incident reporting system. Evidence collected during the first stage of Heinrich's model can be used to provide insights into the effectiveness of previous interventions. As we have seen, however, it can be difficult to rely solely on changes in the numbers of submissions that are made to reporting systems. Contribution rates can change independently of the underlying number of safety-related incidents. The development of a reporting system can increase staff awareness of the need to report particular types of failure.

There are several alternative outcome measures that can be used to evaluate the effectiveness of incident reporting systems. Indirect observations provide feedback about those factors that have contributed to previous incidents and accidents. For example, attitudinal surveys and proficiency tests can be used to determine whether staff are better equipped to deal with situations that led to past failures. Similarly, maintenance activities can be monitored to determine whether they offer effective protection against previous incidents.

A limitation with the use of indirect measures is that previous incidents seldom recur in precisely the same way [700]. It is, therefore, important because revised training and operating procedures cannot simply be based on previous incidents, they must also consider alternative failure scenarios. This implies that incident reporting systems should not only be assessed in terms of the feedback that they provide about existing operations, they should also be evaluated in terms of the contribution that they make to feed-forward risk assessment. This has recently led to the development of accident prediction tools, such as the FRA's Highway-Rail Crossing Web Accident Prediction System (WBAPS) [246]. This uses historic data about previous incidents at particular types of rail crossings to anticipate future accidents at similar locations. Such applications raise ethical questions that complicate the monitoring of incident reporting systems. If an accident prediction proves to be correct then the overall regulatory system can be criticised for failing to prevent a failure that had been anticipated. Ideally, such incident data should direct acquisitions and design policy so that such 'anticipated accidents' are avoided. In particular, reporting systems should inform risk assessments so that the weaknesses of previous systems are not replicated in future developments. A further form of indirect monitoring is, therefore, to assess the impact that incident information has upon future systems and not simply the operation of existing applications.

15.1.1 Direct Feedback: Incident and Reporting Rates

Many industry regulators publish annual summaries that can, in part, be used to assess the performance of reporting systems. For instance, Table 15.3 provides an overview of the US Federal Railroad Administration accident and incident data for 1999 and 2000 [245]. These statistics illustrate some of the problems that arise in interpreting 'raw' information about failure rates. There was a reduction in the total number of reported casualties, from 12,632 to 12,580. At the same time, however, there was an increase in the total number of fatalities, from 932 in 1999 to 937 in 2000. It might be argued that these figures represent an improvement in the safety performance of the rail industry, as noted by the 0.5% reduction in casualties mentioned in Table 15.3. Alternatively, it can be argued that the rise in the number of fatalities represents a worsening of the overall safety record. The reduction in the total number of reported casualties, in this more negative interpretation, might reflect a reluctance to report important safety information.

The relatively small changes illustrated by these statistics can also be explained by annual fluctuations in the incident statistics rather than by changes in the underlying systems. For example, Figure 15.3 normalises accident rates against the number of miles that were travelled in 1999 and 2000. It does not, however, account for differences in the time that it took to travel those distances. Small changes in the average speed of a journey can affect the risk exposure of both staff and passengers. This may be determined by changes in the weather from one year to the next. It might, therefore, be concluded that the small fall in reported casualties might be accounted for by such factors rather than by any overall improvement in rail safety. In order to guard against such apparently 'random' effects we must also consider the issue of statistical significance. We can identify two possible dangers in the interpretation of incident statistics such as those presented in Table 15.3 [375]. A *type 1* error occurs when we decide that changes in the operation of a safety system had an effect on the overall incident data when they did not. A *type 2* error occurs when we decide that changes in the operation of a safety system had no effect on the overall incident data when they did. Significance levels provide a measure of the probability of making a type 1 error.

Data:	Jan-Dec 1999	Jan-Dec 2000	%age Change
Train accidents	2,768	2,983	7.8%
Train accidents per million train miles	3.89	4.13	6.2%
Total reported casualties	12,632 (932 fatal)	12,580 (937 fatal)	-0.5%
Trespasser fatalities	479	463	-3.3%
Employee casualties per 200,000 employee hours	3.39	3.44	1.4%
Highway-rail crossing incidents	3,489	3,502	0.4%
Highway-rail crossing fatalities	402	425	5.7%
Highway-rail crossing incidents per million train miles	4.90	4.84	-1.1%

Table 15.3: FRA Accident/Incident Statistics, February 2002

Given the fluctuations that one might expect in the contribution rate for incidents and accidents, we might therefore set stringent requirements to avoid type 1 errors. There is, however, a trade-off. The lower we set the significance threshold for type 1 errors, the greater the chance there is of making a type 2 error. Further problems affect the use of such statistical techniques. Significance levels are most easily established for carefully controlled experimental situations in which it is possible to distinguish the change, or independent variable, that is linked to any measure, the dependent variable. Unfortunately, there are likely to be many factors that have an impact on overall incident and accident rates. For example, the UK rail sector has recently gone through profound structural changes. It is, arguably, impossible to distinguish the impact of these changes from other changes, such as the introduction of the CIRAS reporting system mentioned in previous Chapters. Assuming that we witness a reduction in the number of rail incidents in the UK, how can we determine whether that improvement was due to the introduction of the reporting system or to higher level changes in the regulatory environment? The problems of obtaining and interpreting incident and accident statistics affect most of the monitoring techniques that will be described in this chapter. For now it is sufficient to observe that these problems are currently being addressed by several recent initiatives. For instance, the statistical unit within the UK Health and Safety Executive has promoted the development of professional standards for the publication of safety-related information by both public and private organisations [340].

Industry-wide incident rates are arguably at too coarse a level to support the detailed decision making that both Heinrich [342] and Kjellen [444] argue must be informed by the monitoring of adverse events. Many regulatory organisations, therefore, publish more detailed information about the incidents and accidents that are reported by particular organisations. This helps to monitor the safety performance of those companies as well as their reporting behaviour. As we have seen, a noticeably low incident rate might indicate either a strong safety record or a poor reporting culture. For instance, Table 15.4 presents incident and accident statistics from Amtrack, the US National Railroad Passenger Corporation. Not only does this table provide an overall indication of incident frequencies, it also provides a more detailed breakdown of the causal factors associated with adverse events. As we have seen in Chapter 10.4 it can be difficult to ensure the consistency and reliability of such findings. For instance, it might be argued that changes in analytical procedures explain the marked rise in human factors related incidents rather than any underlying changes in operator intervention during adverse events and near miss incidents. It is important not to underestimate these analytical effects. For example, Cullen's analysis of the Ladbroke Grove rail crash concludes that a 'no blame' culture is an essential component of rail safety. However, he also acknowledges that this approach can encourage drivers to "accept blame in order to conclude the investigation as quickly as possible" [195]. This may help to explain why 85% of Signal passed at Danger (SPADs) are reported as driver error. These observations emphasise the importance of conducting further studies to validate results such as those shown in Table 15.4. By monitoring changes in causal classification of incidents and accidents it is possible to gain important insights into the 'structural'

Type	1997	1998	1999	2000	%age change 1997-2000
TOTAL ACCIDENTS & INCIDENTS	1,413	1,341	1,265	1,603	13.45
— Fatalities	117	120	105	131	11.97
— Nonfatal	1,328	1,180	1,161	1,412	6.33
TRAIN ACCIDENTS	84	89	85	148	76.19
— Fatalities	1	.	.	.	.
— Nonfatal	74	28	41	106	43.24
— Collisions	3	4	3	8	166.7
— Derailments	51	55	46	80	56.86
— Other	30	30	36	60	100.0
— Track causes	34	29	38	75	120.6
— Human factors	12	27	23	38	216.7
— Equipment causes	8	11	5	19	137.5
— Signal causes	.	.	1	1	.
— Misc. causes	30	22	18	15	-50.0
— Yard accidents	36	41	37	72	100.0
HIGHWAY-RAIL INCS.	176	170	181	202	14.77
— Fatalities	53	50	52	56	5.66
— Nonfatal	123	125	146	90	-26.8
OTHER INCIDENTS	1,153	1,082	999	1,253	8.67
— Fatalities	63	70	53	75	19.05
— Nonfatal	1,131	1,027	974	1,216	7.52
— Employee fatalities	3	2	0	0	-100
— Employee nonfatal	898	840	914	920	2.45
— Trespasser fatalities	57	67	51	70	22.81
— Trespasser nonfatal	32	30	25	18	-43.8

Table 15.4: FRA Accident/Incident Statistics, Amtrak, June 2001

weaknesses that can affect reporting systems.

Incident and accident frequencies can mislead the unwary in other ways. Previous statistics did not account for Amtrack's exposure to certain types of hazard. In particular, the data was not normalised for the relatively large number of rail operations performed by this company. In contrast, Table 15.5 provides normalised data for Amtrak and for the Grand Trunk Western Railroad. It is important that readers understand the ways in which incident frequencies are converted into normalised statistics. For instance, if we assume that normalised rail statistics are calculated by dividing the incident frequency by the number of train miles per year then a reduction in the incident rate might stem occur in several different ways. For example, it might be the result of a fall in the incident frequency with a stable number of train miles or of an increase in the train miles with a stable incident frequency etc. In practice, the FRA calculates the total accidents and incidents rate by multiplying the number of accident and incident reports by 1,000,000 and then dividing the result by the sum of train miles and hours. Similarly, the yard accident rate is the number of train accidents that occurred on yard track multiplied by 1,000,000 and then divided by the number of yard switching train miles. The 'other track' rate is the number of accidents that did not occur on yard track multiplied by 1,000,000 divided by the total train miles minus yard switching train miles. In contrast, the train accident rate is the number of train accidents multiplied by 1,000,000 divided by the total train miles. Highway-rail incident rate is the number of incidents multilied by 1,000,000 divided by the total number of train miles. The FRA's employee 'on duty' rate is the number of reported fatal and nonfatal cases multiplied by 200,000 and then divided by the number of employee hours worked. The trespasser rate is the number of reported fatal and nonfatal incidents, excluding those associated with highway-rail incidents, multiplied by 1,000,000 and then divided by the total

Type	1997	1998	1999	2000	%age change 1997-2000
Amtrak					
Total accidents/incidents	17.95	17.00	15.51	19.57	9.02
Train accidents	2.27	2.51	2.35	4.10	80.99
Yard accidents	18.39	19.70	17.78	34.60	88.19
Other track	1.37	1.44	1.41	2.24	63.48
Highway-rail incs.	4.75	4.80	5.01	5.60	17.90
Employee on duty	4.33	3.87	4.03	4.01	-7.20
Trespassers	2.40	2.74	2.10	2.44	1.57
Passengers on train	4.65	3.44	1.97	5.33	14.59
Grand Trunk Western Railroad Incorporated					
Total accidents/incidents	19.93	19.29	17.90	17.10	-14.2
Train accidents	4.42	3.91	4.05	3.71	-16.1
Yard accidents	15.79	6.32	9.99	5.95	-62.3
Other track	0.92	3.04	1.85	2.96	220.6
Highway-rail incs.	7.07	2.60	4.62	4.82	-31.8
Employee on duty	6.35	7.00	5.52	5.79	-8.80
Trespassers	0.53	.	0.39	0.93	74.82
Passengers on train	.	.	.	.	.

Table 15.5: FRA Normalised Statistics for Two Rail Operators

train miles.

The incident rates illustrated by Table 15.5 support several different monitoring activities. For instance, regulators can make detailed comparisons between the safety performance of companies with different operating characteristics. For example, Amtrak has a relatively stable incident rate for adverse occurrences involving trespassers. Grand Trunk Western has a relatively low trespasser rate which has increased rapidly in the period between 1997 and 2000. Such differences deserve further investigation. There may be operating changes that have increased Grand Trunk Western's exposure to these forms of incident. In which case, they may need to adopt the measures that Amtrak have taken to maintain their more stable rate. Alternatively, Grand Trunk Western's lower rate, even at the 2000 level, may suggest that Amtrak could learn more from their procedures. This example provides further illustration of the need to look beyond such statistics to understand the reasons for such differences.

Previous paragraphs have argued that it is important to consider both incident frequencies and the operating characteristics that are used to derive normalised statistics. Table 15.6, therefore, provides more detailed information about the employee hours, train miles and yard operations of Amtrak and the Grand Trunk Western Railroad. As can be seen, both companies reduced their total train miles between 1997 and 2000. The Grand Western's 14.2% reduction in accidents and incidents occurred when train miles only fell by 4.66%. In contrast, Amtrak's 9.02% increase in incidents and accidents occurred over a period when their train miles fell by 2.65%. As before, however, such analysis must be treated with care. Between 1997-2000, Amtrak increased their number of employee hours by 10.03% while those of the Grant Trunk Western Railroad fell by 9.60%.

Such caveats and complexities characterise the use of normalised incident frequencies as an indicator of the success or failure of incident reporting systems. It can be very difficult to associate particular trends with changes in the underling safety of an application. This problem is even more acute when metrics are used to identify the contribution that a reporting system can itself make to the operation of a safety-critical process. On the 30th November 1999, the UK Deputy Prime Minister, John Prescott, announced that the Confidential Incident Reporting and Analysis System (CIRAS) would be ext ended from the Scottish railway system to cover the entire network [100]. In the aftermath of the Ladbroke Grove crash he said that "I am pleased to say they have taken to

Type	1997	1998	1999	2000	%age change 1997-2000
Amtrak					
Train miles	37063760	35414704	36160704	36080704	-2.65
Yard switching miles	1,957,814	2,080,704	2,080,704	2,080,704	6.28
Employee hours	41663112	43480510	45399073	45840150	10.03
Passengers transported	20555107	21246203	21544160	22985354	11.82
Passenger miles	5.26888E9	5.32419E9	5.28868E9	5.57399E9	5.79
Grand Trunk Western Railroad Incorporated					
Train miles	5,657,394	5,376,050	5,190,349	5,393,620	-4.66
Yard switching miles	1,330,157	1,425,036	1,401,708	1,344,762	1.10
Employee hours	4,124,903	4,372,190	4,418,149	3,728,758	-9.60
Passengers transported	0	0	0	0	.
Passenger miles	0	0	0	0	.

Table 15.6: FRA Normalised Statistics for Two Rail Operators

heart everything that I asked of them in the wake of that terrible tragedy and today can announce concrete results on measures that can be taken now and commitment to a programme of action for longer-term projects... I repeat my pledge to the public that the industry will make rail travel even safer". In spite of the perceived success of the CIRAS system, it is hard to demonstrate that Scottish railways have a significantly better safety record than other areas of the network. As we have seen, the region's main operating company was one of ten that were warned by the Railways Inspectorate in June 2001 that they had not done enough to combat the problem of Signals Passed At Danger [112]. Such arguments suggest that there may well have been other motives behind the expansion of the CIRAS reporting system beyond the relative safety record of the company that operated it. For example, CIRAS' original developers and operators [198] echo Clarke's argument that 'incident reporting might be viewed as an objective indicator of manager's commitment to safety' and that these 'perceptions underlie a lack of mutual trust between staff and managers, which has implications for the fostering of open and honest communications within the network and for the development of a positive safety culture' [170]. These sentiments are similar to those put forward by Cullen in his investigation into the Ladbroke Grove accident where he argues that confidential reporting systems would be unnecessary in an industry with a supporting safety culture [195]. Information about near-miss occurrences should be provided in an open manner without fear of subsequent persecution. Both arguments suggest that the potential utility of a reporting systems can be assessed in terms of the information that they provide about the safety culture in an industry.

In preparing this book, I have had many interviews with individuals who are involved in the development of the UK national rail reporting systems. In the course of these discussion, a number of criticisms have been raised about some of the arguments that are presented in the previous paragraph. These caveats illustrate the complex issues that arise during the monitoring of such applications. They also illustrate the way in which significant resources can be invested in the development of a reporting system even though there may be little consensus within an industry about the metrics that might be used to assess the success or failure of the system. Firstly, criticisms have been made about the statistics that were used by the HMRI SPAD report [112]. Secondly, performance in this area can be argued to have little connection with the information obtained from the CIRAS reporting system. Most 'Signals Passed at Danger' are observed by other rail personnel including signaling staff. They will, therefore, be notified by other means rather than the confidential incident reporting system. The success of the reporting system is, therefore, being assed in terms of safety-related incidents that it is not intended to address.

This section has identified a number of problems that frustrate the use of direct safety metrics as a means of monitoring the performance of incident reporting systems. These can be summarised as follows:

- there can be disagreement over the metrics that are used to assess the overall safety of complex applications. This creates problems when those metrics are, in turn, used to assess the contribution of a reporting system.
- it can be difficult to obtain data about the safety record of some applications even when there is agreement over the metrics to be used. Different jurisdictions can result in some data being withheld. Other organisations may under-report injuries and illnesses. This can make it difficult to assess the safety record of an industry which in tun complicates the use of direct metrics to assess the contribution of incident reporting systems.
- it can also be difficulty to identify normalising factors for statistical analysis. As we have seen, raw frequencies cannot easily be used to compare the performance of reporting systems in large and small organisations. There can be disagreements over the normalising factors to be used. It can also be difficult to collate the necessary operational statistics once those factors have been identified.
- incident reporting systems may only have an indirect effect on the metrics that are used assess the safety performance of an industry. This builds on Wright's arguments that the 'low severity' incidents described in reporting systems are very different in nature from the high-consequence accidents that are typically used to assess the overall safety performance of many industries [875].

These caveats have led some regulators to look beyond direct measures. Rather than assess the performance of a reporting system in terms of overall changes in the safety of an industry, more attention is paid to the indirect operational impact of each contribution. In other words, the success or failure of the system is assessed in terms of the different lessons that are learned from the incidents that are reported to it.

15.1.2 Indirect Feedback: Training and Operations

Previous sections have described how the Railway Group publishes regular summaries of rail safety across the UK rail network [692]. This reiterates the recommendations obtained from the CIRAS reporting system, mentioned above. For example, the survey published in October 2001 reminded managers 'at all levels in companies that are members of the Railway Group' that they should read the the publications from the national reporting system. In particular, they were advised to note the predominance of organisational problems in the incidents that were reported to the scheme. Most of these related to problems with rosters and shift patterns; 'short staffing is the most common perceived cause, and the most common consequence is fatigue'. The review also reiterated that poor communication by supervisors and management was a noted cause of many incidents. Rule violation was the most significant cause of what were described as 'workplace incidents'. The publication of this information is very significant. The majority of the Railway Group review is devoted to a statistical analysis of safety data, mainly focusing on the frequency of accidents and events that fall within the scope of a mandatory reporting system. The same approach is not used for the voluntary incident reporting system. Rather than providing statistics about contributions to the scheme or about the impact of recommendations on the frequency of accidents, the focus is on the lessons that have been learned from the system. This is an indirect approach because these lessons are intended to have a knock-on effect upon the other performance indicators.

The railway Group deliberately focuses on the high-level insights provided by the CIRAS incident reports. It does not identify particular recommendations and so there is a danger that they will have only a minimal effect on the recipients of the summary. Rather than directly measuring changes in accident and incident rates, it is possible to monitor the impact of a reporting system in terms of the changes that are made to operating practices. For example, UK reporting systems consistently revealed that track-side workers form the largest category of victims in rail related injuries and fatalities. In April 1995, these incidents led to the introduction of a relatively complex set of recommendations to segregate workers from trains. This involved a 'permit to work' scheme that ensured workers were either segregated from lines on which trains were running or that track workers were warned of approaching trains in time to move to a place of safety. Segregated worksites became known as 'green zones', while non-segregated worksites became known as 'red zones' [357]. These recommendations reduced but did not eliminate incidents involving track-side workers and so the HMRI started a further programme to review progress and to develop a strategy for future improvements. A questionnaire was developed to gather information about the effectiveness of previous recommendations based on the subsequent incident reports. These topics included the procedures used to monitor red and green zone working and the red zone risk assessment process. The results of these studies helped to identify further recommendations. In particular, it identified that some of the rail operators had provided misleading statistics when providing information about the normalising factors that, as we have seen, are important for the direct assessment of incident reporting systems:

> "A claimed 11% increase in green zone working, when analysed, represents a reduction in the proportion of green zone working because of a rise in the number of worksites (38% to 33% approx. over a twelve month period). The Railway Group Safety Performance Report 1998/99 has identified a need to provide information on an, 'exposed hours' basis for monitoring purposes" [357]

This illustrates the way in which regulatory and investigatory organisations can support investigations into the effectiveness of recommendations produced in response to previous incidents. These studies provide indirect insights into the utility of the reporting system. They can also yield additional recommendations that are intended to reduce the likelihood or mitigate the consequences of further incidents. Finally, they can also detect weaknesses in the way that a reporting system is currently being run. This is illustrated by the problems in reporting normalisation information, mentioned above. There are further examples of reporting systems being assessed in terms of the recommendations that they generate. For instance, the FRA's Switching Operations Fatality Analysis (SOFA) Working Group recently analysed 76 incident reports from January 1992 to July 1998 [243]. They also considered more limited FRA data from 1975 to 1991. The small total number of incidents and the varied circumstances of each event persuaded the Working Group that recommendations could not be based on formal statistical analysis. Instead, they used the incident data to devise 'five SOFA lifesavers'. These can be summarised as follows:

1. Notification to the locomotive engineer before fouling track or equipment. 'Any crew member intending to foul track or equipment must notify the locomotive engineer before such action can take place. The locomotive engineer must then apply locomotive or train brakes, have the reverser centered, and then confirm this action with the individual on the ground. Additionally, any crew member that intends to adjust knuckles/drawbars, or apply or remove EOT device, must insure that the cut of cars to be coupled into is separated by no less than 50 feet. Also, the person on the ground must physically inspect the cut of cars not attached to the locomotive to insure that they are completely stopped and, if necessary, a sufficient number of hand brakes must be applied to insure that the cut of cars will not move'.

2. Extra precautions when two or more train crews are working on the same track. 'When two or more train crews are simultaneously performing work in the same yard or industry tracks, extra precautions must be taken: C SAME TRACK. Two or more crews are prohibited from switching into the same track at the same time, without establishing direct communication with all crew members involved. C ADJACENT TRACK. Protection must be afforded when there is the possibility of movement on adjacent track(s). Each crew will arrange positive protection for (an) adjacent track(s) through positive communication with yardmaster and/or other crew members'.

3. Safety briefing. 'At the beginning of each tour of duty, all crew members will meet and discuss all safety matters and work to be accomplished. Additional briefings will be held any time work changes are made and when necessary to protect their safety during their performance of service '.

4. Proper communications. 'When using radio communication, locomotive engineers must not begin any shove move without a specified distance from the person controlling the move. Strict compliance with 'distance to go' communication must be maintained. When controlling train or engine movements, all crew members must communicate by hand signals or radio signals. A combination of hand and radio signals is prohibited. All crew members must confirm when the mode of communication changes'.

5. Paying proper attention to new crew members. 'Crew members with less than one year of service must have special attention paid to safety awareness, service qualifications, on-the-job training, physical plant familiarity, and overall ability to perform service safely and efficiently. Programs such as peer review, mentoring, and supervisory observation must be utilised to insure employees are able to perform service in a safe manner' [243].

These recommendations were published and then widely publicised within the US railway industry. There then followed a steady decline in switching incidents until in 2000 the FRA noted that the total number of switching-related deaths quickly exceeded those for 1999. These incidents raised questions about the working practices of crew members assigned to perform switching operations. They occurred on large and small railroads and included experienced employees with between two years to more than thirty years experience. This led to a review of the recommendations that had been derived from previous incidents. The FRA study concluded that most of the incidents 'could probably have been prevented if all employees on each railroad had strictly followed the five recommendations of FRA's Switching Operations Fatality Analysis (SOFA) Working Group and the applicable Federal and railroad company operating and safety rules to which they relate' [243].

The previous paragraph illustrates the way in which the success of a reporting systems can be assessed in terms of whether particular recommendations might have prevented recent incidents. This approach has a strong appeal. As we have seen, there are few guarantees that regulators will accept the findings of reporting agencies. Similarly, companies may fail to implement the recommendations that are identified from previous incidents. The use of more direct reporting statistics ignores the impact that such factors can have upon the effectiveness of a reporting system. Indirect forms of analysis, similar to that presented by the FRA, serve to reiterate the lessons that might have been learned if these recommendations had been implemented.

It is also important to stress the limitations of these arguments. The FRA's assertions about the effectiveness of the SOFA recommendations relies upon complex counterfactual arguments. More recent incidents would have been avoided had operating companies implemented the findings from previous incidents and accidents. Unfortunately, Chapters 9.3 and 10.4 have illustrated the dangers of this style of reasoning. In particular, it can be difficult to obtain evidence to support claims about the potential effect of recommendations that were not followed. The reiteration of well-known recommendations can also have a strong adverse effect if they are interpreted as needless reminders to 'do better next time' [411]. There is also a danger that by reiterating previous recommendations, regulators and investigators will fail to adequately consider the reasons why those findings were not followed in recent incidents. For instance, a study of incidents involving children near railways persuaded Administrator Jolene Molitoris that previous messages about the dangers of playing near railways had not been effectively communicated to the target audience. She, therefore, initiated the 1995 *Always Expect a Train* campaign using 270 television and cable markets, 673 radio markets and 194 publications [234]. As part of this work, a series of Public Service Announcements broadcast 'deliberately graphic reenactments of motor vehicle-train collisions and railroad trespassing incidents, designed to grab the viewer's attention'. A classroom teaching initiative as also created to embed safety-related information within multimedia resources on railway history and technology. These actions are instructive because they suggest a thorough re-evaluation of the way in which safety recommendations were communicated to the public. Such initiatives need not have been created if she had simply evaluated the reporting system in terms of whether previous recommendations might have prevented the incidents that were being reported. The recommendations publicised in these campaigns were essentially the same as those used in previous initiatives. In contrast, the successful implementation of the recommendations and of the incident monitoring system as a whole depended on the manner in which those recommendations were communicated to the target audiences.

These initiatives have been attributed with a 19% reduction in child-related rail 'casualties' [234]. Such statistics again raise the caveats and concerns that the previous section has raised about outcome statistics. However, the use of these figures to validate the revised recommendations is instructive because it illustrates the way in which most reporting systems are assessed both in terms of direct and indirect measures. The benefits of these systems are expressed both in terms of the particular insights that they provide and by the statistical reduction in severity or frequency rates. For instance, a fatal accident near Edson, Alberta, in early August 1996 forced Transport Canada to review their recommendations for avoiding runaway trains. In consequence, they encouraged a number of actions by the operating companies. These can be summarised as follows:

- training and education initiatives aimed at increasing employee and customer awareness of rules governing proper securement of cars;
- increased compliance monitoring by supervisors;
- increased inspection of derails to ensure proper application and positioning and to recommend locations where derails should be applied; and
- increased police monitoring of high vandalism areas [778]

The Alberta accident focussed the attention of the public and the rail industry on runaway train incidents. The high consequences of this incident led to demands for more direct evidence to demonstrate the effectiveness of these recommendations. It was insufficient simply to argue that the accident had led to the publication of the previous recommendations without also providing evidence that those recommendations were useful. Transport Canada, therefore, commissioned detailed comparisons between the number of runaway rolling stock incidents both before, between January and July, and after the publication of their recommendations, between August and December 1996. However, they anticipated that there would be a seasonal fall in the number of incidents in the winter months as the number of traffic movements fell. They, therefore, also compared this data with the number of incidents for corresponding months in 1994 and 1995. 42.3% of runaway rolling stock occurrences took place during the August-December period of 1996, compared with 44.4% during 1995 and 52.5% during 1994. These percentages are based on the total incident frequency for only the two periods that are considered in each year. The percentage of runaway rolling stock incidents that resulted in accidents in the August to December period fell from 60% in 1994 to 46.3% in 1996 and 46.5% during 1995. This created problems for the statistical analysis of the recommendations because 'the decrease in the percentage of uncontrolled movement incidents accounted for the entire decrease in runaway rolling stock occurrences that took place during the August-December period of 1996 when compared with 1994 and 1995 figures' [778].

15.1.3 Feed-forward: Risk Assessment and Systems Development

Kjellen distinguishes between four different levels of organisational learning [444]. These levels help to distinguish between different forms of metric that might be used to assess the performance of incident reporting systems:

1. *short-term learning in the workplace.* This describes immediate actions that are taken to address the direct causes of an adverse occurrence or near miss event. Kjellen argues that this form of learning only affects 'short-term memory'. In other words, any insights are likely to be forgotten as workers and supervisors move to new tasks or activities.
2. *long-term learning in the workplace.* This describes interventions that have a more sustained impact on operating practices within the particular work group or location where the incident took place. For example, they may result in the publication of revised operating guidelines or in documented modifications to particular pieces of equipment. Recommendations prevent recurrences but have limited scope and may not be effectively propagated throughout a factory or company.

3. *long-term learning in similar workplaces.* This can involve changes in the technical and administrative systems for the departments that are involved in an incident. Any recommendations will have a lasting effect and are likely to be propagated to similar departments in other areas of an organisation.

4. *long-term learning in management systems and norms.* These recommendations have profound effects on the way in which work is organised and managed. It can effect policy, goals and the specification of particular activities. The recommendations will affect most of the company and may have an impact on other organisations.

It can be argued that direct metrics, which focus on outcome measures, can be used to distinguish between these different forms of organisational learning. For example, level 1 recommendations may result in a short term fall in the accident and injury rates associated with a particular workplace. Level 4 changes will have a sustained effect on outcomes across many different sectors of an organisation. As we have seen, however, there are many factors that can confound the use of direct metrics to assess the impact of recommendations from a reporting system. In particular, the difficulty of obtaining reliable and appropriate statistical measures for safety improvements can be an obstacle to this approach. There are further problems. For instance, it can be difficult to distinguish between level 1 and 2 recommendations without careful monitoring of safety improvements over a prolonged period of time. Similarly, it can be difficult to distinguish between level 3 and 4 recommendations without reliable metrics for the performance of many different groups within an organisation. There are further problems. For example, it may take some time before particular recommendations have a discernible impact on outcome measures. These can be a delay before changes in the 'norms' and practices of senior management are effectively communicated into changes in operating procedures and acquisitions policy.

As we have seen, indirect measures do not focus on outcome metrics but, instead, concentrate on demonstrating the effective implementation of recommendations in the aftermath of an incident or accident. A range of further problems affect the use of indirect metrics as a means of assessing incident reporting systems. For example, commercial opportunities and regulatory intervention often force organisations to revise their working practices. These changes can lead to the introduction of new equipment and operating procedures. They can also invalidate many of the recommendations that were made in the aftermath of previous failures. In such circumstance, it can be difficult to distinguish between situations in which those recommendations have been 'forgotten' and situations in which previous recommendations no longer apply to present working practices. These problems are compounded by the difficulty of indirectly monitoring long-term changes in management practices. It is easier to identify level 1 changes than it is to demonstrate the effective implementation of level 4 recommendations. New piece of equipment and revise manuals are more tangible than changes in management 'norms'. Attitudinal questionnaires often focus on short-term effects and are subject to a host of biases [344]. This makes it difficult to interpret the results of such surveys, especially in the aftermath of safety-related incidents.

A number of authors, including Benner [73], have argued that evidence of long-term 'organisational' learning can be obtained by examining the impact that adverse events have upon risk assessment practices. This approach addresses many of the criticisms of direct and indirect metrics that were introduced in the previous paragraphs. For example, risk assessment practices provide a useful measure of management norms because they have a direct impact upon the allocation of finite resources. Risk assessments reflect the priorities associated with development and maintenance activities and hence indicate operational concerns at higher levels within an organisation [189]. The outcome of risk assessment procedures can also be used to predictive potential safety problems. In other words, the priorities derived from risk assessments helps to identify those areas that managers believe will pose the greatest threat to the future safety of an application. This offers the opportunity for analysis to determine whether incident statistics actually support those priorities. In contrast, direct metrics provide information about the post hoc success or failure of previous operational decisions.

A number of practical problems complicate the use of risk assessment metrics to assess the performance of incident reporting systems, For instance, many local incident reporting systems are isolated from the revenue streams that are necessary to fund large-scale safety improvements [419]. In contrast, they must fund the implementation of safety recommendations from the savings that are made through previous recommendations. This approach is intended to ensure that incident reporting systems are well integrated with wider 'lessons learned' applications. It also ensures that the reporting system is self-funding. Unfortunately, such practices also isolate the reporting system from normal risk assessment practices within the rest of the organisation. Important insights about the causes of previous incidents and accidents may not be communicated to those individuals who have the greatest influence on future acquisitions. In contrast, many other reporting systems are explicitly integrated into risk assessment systems. For instance, Transport Canada's guidance on the development of rail safety management systems identifies three stages to the risk management 'process' [781]:

1. Identification of Safety Issues and Concerns
   The first stage of risk management explicitly focuses on gathering 'input from incident/accident investigations and safety data collection and analysis'.

2. Risk Estimation
   The information from the first stage of the process is then analysed to assess the probability and severity of a potential hazard using either qualitative or quantitative techniques. Quantitative estimates can 'sometimes be developed from safety performance data, illness and injury records'. Transport Canada do, however, note that probability estimates based on 'historical data assume that future conditions will mirror those of the past'. If there is no relevant incident data then more qualitative techniques, such as event-tree analysis should be used to generate risk estimates. Event tree analysis enumerates the outcomes from a given event to map out the likely sequences of consequent events. For each event, analysts can consider the consequences of safety systems failing or succeeding in their specified function. Probabilities can then be associated with each path through what can be thought of as a form of decision tree [839].

3. Risk Evaluation
   The final stage of the management process determines which risks are tolerable, tolerable with mitigation and or unacceptable. These decisions should be guided by classification methodologies based around the Risk Assessment Matrices described in Chapter 11.5.

The Railway Safety group exploits a similar management approach to risk across the UK rail network [692]. Their three stage model manages risk by 'understanding the relationship between precursors and incidents... then by measuring the precursors and finally by applying action to the areas identified. This risk management process is supported by a Precursor Indicator Model (PIM). This relies upon 16 measures that were identified through the analysis of previous accidents and incidents. Table 15.7 enumerates these precursors. It also provides a percentage indicator that is intended to represent the 'risk' associated with each contributory factor to catastrophic rail accidents. Arguably a better description would be the percentage of major accidents in which analysts identified these precursors. The Railway Safety group interprets table 15.7 as providing a 'severity weighting' for precursors by arguing that 'a third of all injuries from train accidents are caused by category A SPADs'.

The historic severity weightings identified in Table 15.7 can be used to address some of the limitations of direct metrics. In particular, it provides a means of assessing the safety of complex systems that suffers very few catastrophic failures. We begin by pairing the severity assessments of each precursor from Figure 15.7 with the number of times that the precursor has occurred in the time period under consideration. There pairs can be used to construct a vector of the form $(\mathit{frequency}_1, \mathit{historic\_weighting}_1;\ ...;\ \mathit{frequency}_n, \mathit{historic\_weighting}_n)$. It is important to note that precursor frequencies can be obtained even though these incidents may not have led to a major accident. However, the Railway Safety group can use this information to calculate the overall risk of a major accident in the following way:

$$\mathit{system\_risk\_assessment} =$$

Precursor	Proportion of Train Accident Risk
Category A SPADs	32.84%
Level crossing misuse	22.84%
Track quality	12.86%
Irregular working	8.31%
Rolling stock failures	8.00%
Environmental factors	6.07%
Vandalism	2.91%
Structural failures	1.54%
Train speeding	0.98%
Level crossing failures	0.90%
Irregular loading of freight trains	0.83%
Wrong-side signaling failures	0.36%
Non-rail vehicles on line	0.27%
Possession irregularities	0.15%
Hot axle box	0.13%
Animals on the line	0.04%

Table 15.7: UK Railway Safety Group's Precursor Indicator Model (PIM)

$$\sum_{i=1}^{n} frequency_i.historic\_weighting_i \qquad (15.1)$$

As mentioned, the procedures used by UK Railway Safety show how incident and accident data can be used to replace direct measures of system safety for infrequent, high-consequence events. The model exploits failure information in two ways. Firstly, the weightings associated with different precursors depends on the frequency of their observation in previous accidents and incidents. This ensures that the relative importance of those weightings will change as different failures become more or less significant to the overall safety of the rail system. It is, however, also important to ensure that these weightings are derived from a relatively large sample of previous failures to ensure that adequate attention is paid to long term problems as well as more immediate changes in the precursors to incidents and accidents. Secondly, the frequency of particular precursors is introduced into the calculation of overall system risk. These precursors are directly identified from recent incident and accident reports so that any calculations reflect more immediate changes in the performance of underlying systems. The UK calculations for October 2001 reflect reductions in the frequency of track quality faults, wrong-side signaling failures and train speeding. They also reflect increases in level crossing misuse, irregular working, vandalism, level crossing incidents and rolling stock failures. These changes in frequency combined to create a slight increase in the overall accident risk in the first half of 2001/2002 [692].

Although the PIM approach illustrates both a comprehensive and effective integration of risk assessment and incident reporting, it is possible to identify a number of potential problems. As we have seen, the use of previous accident information to identify incident precursors will only provide reliable risk assessments if future incidents are similar to those that have occurred in the past. It is, therefore, essential that the components of Table 15.7 be reviewed at regular interval to ensure that they do not exclude potential causes of future failure. A further problem is that the PIM approach fails to distinguish between the different levels of severity that are associated with catastrophic accidents. This is significant because many safety objectives are expressed in terms of the number of fatalities per train miles. The UK objective is 0.3 fatalities per million train miles by 2009; in 2000-2001 the annual moving average was 0.59 while in 2001-2002 it was given as 0.52 [692]. The Rail Safety group recognise this problem and, therefore, calculate a weighting for outcomes based on previous accidents. These are expressed in terms of 'equivalent fatalities per 10 train accidents' over a specified period of time. t is important to emphasise that this metric relates to different types of accidents and not the precursors, mentioned above, that lead to those accidents. Current weightings based on a 16-year interval are given in Table 15.8. The relatively low weighting associated with 'buffer stop collisions' has led some analysts to question whether they should be included in the significant train accident statistics.

Type of Accident	Consequence Weighting
Passenger collisions	7.331
Non-passenger collisions	0.777
Passenger derailments	0.927
Non-passenger derailments	0.005
Buffer-stop collisions	0.162

Table 15.8: UK Railway Safety Group's Significant Train Accident Weightings

Table 15.8 illustrates the way in which safety improvements can have a mitigating effect on the outcomes of adverse events. It does not, however, illustrate the procedures by which precursors are associated with particular types of outcome. This is far from straightforward. Recall from Chapter 10.4 that Bayes theorem considers the probability of a given hypotheses, B, in relation to a number of alternative hypotheses, B_i where B and B_i are mutually exclusive and exhaustive:

$$Pr(B \mid A \wedge C) = \frac{Pr(A \mid B \wedge C).Pr(B \mid C)}{Pr(A \mid B \wedge C).Pr(B \mid C) + \sum_i P(A \mid B_i \wedge C).P(B_i \mid C)} \qquad (15.2)$$

Bayes' theorem can be used to assess the probability of a particular factor or precursor, B, causing a failure given that an incident report has identified that causal factor, A. Suppose we examine reports of previous buffer stop collisions to determine whether or not they were caused by train speeding, B. It is unlikely that we will have complete confidence in the intuitive causal analysis of every investigators. We, therefore, conduct a quality control exercise that performs a more detailed causal analysis for a sample of recent reports. This indicates that there is a false positive rate of 4%. In other words, 4% of the reports argue that the driver was speeding when they were not. We might also conclude that the false negative rate is 3%. This is the percentage of reports that suggest speeding was not the cause when subsequent investigations revealed that it was. The reports were, therefore, 96% accurate for incidents in which the buffer stop incidents were caused by train speeding. They were, in contrast, 97% accurate for incidents that were not caused by this precursor. To simplify the exposition, we assume that further analysis reveals that 1% of incidents were caused by train speeding. Table 15.7 provides the more accurate figure of 0.98%. We can use the previous formalisations to determine how likely it is that train speeding caused a buffer-stop incident given that a report identifies this as a cause of an incident. The probability of train speeding having caused the buffer stop is less than 20% based on a positive incident report! The following formulae adopt the convention of including the context C for the reasons given in Chapter 10.4. Dembski uses a variant on this example to demonstrate the ways in which people can fail as 'intuitive probabilists' (see [201], pp 83-84). These apparent failures are so deeply engrained that I remain unconvinced by aspects of his argument even if I can agree with the underlying mathematics!

$$Pr(B \mid A \wedge C)$$

$$= \frac{Pr(A \mid B \wedge C).Pr(B \mid C)}{Pr(A \mid B \wedge C).Pr(B \mid C) + Pr(A \mid \neg B \wedge C).P(\neg B \mid C)} \qquad (15.3)$$

$$= \frac{(0.97).(0.01)}{(0.97).(0.01) + (0.04).(0.99)} \qquad (15.4)$$

$$= 0.1968 \qquad (15.5)$$

A number of further practical issues complicate the use of incident data within the PIM approach to risk management. In particular, it is unclear how specific interventions by particular companies are to be identified from high level, industry-wide statistics. One possible mechanism is through the enforcement actions that are recorded in the Safety Group reviews. These record the date, location and nature of each violation that triggered a prosecution. Examples include 'all reasonably practicable measures have not been taken to reduce the risk of signal XXXX being passed at signal danger' and 'redundant coach left on disused track beneath bridge being vandalised and used by children'. These incidents are fed into the calculation of precursor frequencies mentioned in previous sections. The focus is, however, on the correction of violations rather than on the pro-active interventions that might further reduce potential risks.

A detailed analysis of the impact of particular mitigation or risk reduction measures arguably lies beyond the scope of the UK Safety Group's periodic reviews. Specific guidance is included within 'Focus Area' publications on reducing SPADs, Trackworker Safety, Trespass and Vandalism. Transport Canada provides a further example of the use of incident reporting data to inform risk assessments. As mentioned, their three stage risk management process is similar to that advocated by the UK Railway Safety group. This approach was employed to support the analysis of six incidents in which hot bearings had led to 'burnoffs' on the Canadian rail system. Hot journal bearings occur when inadequate wheel bearing lubrication or mechanical flaws cause an increase in bearing friction. If undetected, the resulting rise in bearing temperature can lead to a bearing burnoff which can cause a derailment. The analysis began by assessing the frequency of previous incidents. An average of approximately 10 derailments per year were linked to burnoffs between 1987 and 1993. In other words, there were only 2 burnoffs per billion freight car miles. The consequences of these events were also assessed. These incidents did not result in any fatalities; 'no passenger or member of the public has been fatally injured because of any main track derailment for over 10 years' [777]. The investigation assumed an average cost of $250,000 per derailment in terms of damage to rolling stock and infrastructure. This analysis of the risk associated with these incidents was then used to determine whether or not to invest in a number of detection and mitigation techniques. For example, the Canadian rail system already had a number of 'hot box' detectors. One proposal suggested that the number of these detectors be doubled so that the distance between consecutive units would be reduced to around 12 miles. This would involve 400 new single track and 200 double track installations at an initial cost of $90 million with a further $5 million per year for maintenance. Six detailed investigations led investigators to conclude that the additional detectors would, at best, have prevented half of these derailments. This would have brought the annual average number of burnoffs down from to 5 per year saving around $1.25M per year given the cost estimates for each previous derailment The investigators concluded that 'spending $90 million plus $5 million per year on additional detectors to save some 5 burnoffs per year without any evidence that lives would be saved or injuries prevented is clearly not a beneficial use of society's resources, is out of line with risk management expenditures in other areas, and is not a recommended course of action for the Railway Safety Directorate or the industry to pursue' [777].

The previous example illustrates a number of links between such risk assessments and the monitoring of incident reporting systems:

1. *issue identification.* Firstly, incident reporting schemes help to trigger risk management activities. Evidence from previous derailments justifies the initial investigation of bearing burnoffs as a causal factor. In this way, the incident reporting system 'earns its keep' as a necessary part of a safety management system even though it may not be possible to identify suitable metrics to support the performance of the scheme in issue identification.

2. *data sufficiency.* Secondly, the previous example illustrates the way in which risk assessment activities provide a means of assessing whether incident investigations yield sufficient insights into the causes of adverse events. In particular, the Transport Canada investigators make use of the counter factual argument that doubling the number of hot box detectors would only have prevented about half of the annual number of derailments from bearing burnoffs. Such an analysis depends upon sufficient information being available to support their conclusions. If such evidence had not been available then the risk assessment would have been seriously

flawed and the investigatory process would have been modified.

3. *investment savings.* One of the side-effects of integrating incident reporting into risk management procedures is that it provides monetary assessments of the strategic value of information that is provided about previous failures. In the previous example, it can be argued that this data helped to avoid investing more than $90 million in a scheme that would have yielded limited safety benefits. Sadly, such superficially appealing arguments can be challenged in a number of ways. They assume that the costs associated with future incidents will continue to similar to those incurred by previous failures. The $90 million costs might, however, appear to be justified if a future incident resulted in multiple fatalities.

The integration of incident reporting systems into risk management procedures provides a powerful justification for the investment that is needed to elicit and analyse information about previous failures. The previous list, therefore, shows how the utility of a reporting system can be indirectly assessed in terms of the support that it provides for risk assessment. Such approaches are unlikely to be sufficient. In particular, they provide relatively little information about the effectiveness of the reporting system in eliciting information about adverse events. Similarly, these techniques cannot easily be used to address the problems of intra and inter analyst reliability that have been identified as a potential limitation for direct, indirect and feed-forward metrics. The following section, therefore, focuses on the process measures that can be used to assess the performance of the reporting system itself rather than the utility of the information that it produces.

15.2 Process Measures

It is important to monitor the costs that a reporting system incurs as well as the benefits that it delivers in terms of safety improvements. For example, the Aviation Safety Reporting System spends about $3 million annually to analyse roughly 30,000 reports, at about $100 per case. These techniques would cost almost £50 million annually if the same techniques were applied to the 850,000 adverse events in the UK National Health Service [480]. Such figures illustrate the importance of retaining a close control of the management of incident reporting systems. It is difficult to obtain similar estimates for national rail reporting systems. The UK Health and Safety Executive argued that 'trials carried out in Scotland since 1996 (involving ScotRail, GNER, and Virgin (N)) indicate that the CIRAS confidential incident reporting system improves incident reporting as well as being financially beneficial to the companies concerned' [327]. They did not published detailed figures to support this argument and several operating companies expressed concerns about the financial overheads associated with voluntary incident reporting. Their concerns echo criticisms voiced about the development of national reporting in Australian railways. The Booz, Allen and Hamilton report recognised that individual rail operators, infrastructure managers and regulators gather data to monitor their own performance over time [55]. It was the lack of 'consolidation, consistency and analysis' at a national level that gives the greatest cause for concern because 'the industry is not yet convinced that these processes are consistently applied or completely appropriate'. In consequence, the proponents of national reporting systems have been forced to monitor the operation and management of their schemes and not simply the safety-related information that they produce.

15.2.1 Submission Rates and Reporting Costs

A number of crude measures can be used to assess the cost effectiveness of a reporting system. For example, the total investment in any scheme might be divided by the number of reports that are received each year. There are a number of potential benefits from using submission metrics to support the monitoring of incident reporting systems. In particular, this can help to identify the problems of under-reporting that were studied in Chapter 4.3. As the FRA note, employees may even neglect medical treatment rather than expose themselves to workplace harassment:

> "FRA has become increasingly aware that many railroad employees fail to disclose their injuries to the railroad or fail to accept reportable treatment from a physician because they wish to avoid potential harassment from management or possible discipline that is sometimes associated with the reporting of such injuries. FRA is also aware that in some instances supervisory personnel and mid-level managers are urged to engage in practices which may undermine or circumvent the reporting of injuries and illnesses." [235]

There are a number of problems with using submission rates as a metric for overall system performance. Most reports are received from a relatively small section of the workforce in many industries. Increases in the numbers of contributions from these employees can mask significant under-reporting in other areas. This point is illustrated by Table 15.9, which presents statistics on fatal and non-fatal injuries on US railways [245]. The majority of reports are filed by workers 'on duty' and under the employment of a rail operating company. In contract, there are relatively few reports form rail contractors even though they make up a significant proportion of the workforce employed in maintenance and infrastructure projects in this industry. 119 'workers on duty' were killed on US railways between 1997 and 2000. During this period, they suffered 33,738 non-fatal injuries. This yields a Heinrich ratio of 283.51 non-fatal injury reports per fatality. In contrast, 31 contractors were killed between 1997 and 2000. During this period only 1466 injuries were reported at a Heinrich ratio of 77.29. Contract workers might be less likely to be involved in incidents than other workers and hence we might expect a lower ratio of non-fatal injuries to fatalities. There is, however, considerable evidence to the contrary [344]. Contract workers are often less well trained and briefed on their operating tasks than full-time employees. They are also less easily integrated into working groups when they may be moved between operational responsibilities more frequently. In consequence, it can be argued that the difference in ratios illustrates a problem of under-reporting of non-fatal injuries amongst this section of the workforce. Simply dividing the overall number of non-fatal incident reports by the annual expenditure on a reporting scheme would fail to identify such structural problems. Similarly, Table 15.9 illustrates the difficulty of identifying an underlying pattern in the submission data for non-fatal incidents. Short-term reductions, for example in passenger-related incidents between 1997 and 1999, can be offset by increases elsewhere, for instance in employee on duty submissions.

	Fatalities				Nonfatal Conditions			
	1997	1998	1999	2000	1997	1998	1999	2000
Worker on duty (railroad employee)	37	27	31	24	8,295	8,398	8,622	8,423
Employee not on duty	.	2	.	1	263	219	216	286
Passenger on train	6	4	14	4	601	535	481	658
Nontrespasser	362	324	302	332	1,517	1,201	1,307	1,264
Trespasser	646	644	572	570	728	677	650	606
Worker on duty(contractor)	6	2	2	.	213	237	172	183
Contractor(other)	5	3	10	3	121	143	212	185
Worker on duty(volunteer)	.	.	.	.	3	11	4	6
Volunteer(other)	.	.	.	.	3	3	1	2
Non-trespasser, off rr prop	1	2	1	3	23	35	35	30

Table 15.9: FDR Rail Incident Reports by Worker (1997-2000)

One way of avoiding such problems is to attempt to increase contributions from particular sections of a workforce while maintaining or reducing the overall cost of operating a reporting system. This raises further problems. As we have seen in Chapter 1.3 short-term changes in submission rates can occur independently of changes in the safety of an underlying application. Awareness arising campaigns can elicit large numbers of 'low risk' contributions from a minority of the target workforce without encouraging the mass of their colleagues to report on the more serious incidents that are masked from a reporting system. Such campaigns are often costly and the effects that they achieve can be very short-lived [444, 344]. There are a number of further ways in which submission rates can fail to provide accurate indicators of the underlying safety of an application process. Such metrics are profoundly affected by changes in the criteria that are used to identify particular types of incident. For instance, the UK Railway Group widened the definition of SPAD severity Category 3 in May 2000. The new definition increased the number of incidents falling into this category. It provided a larger data group and was, therefore, argued to increase the opportunity for more meaningful analysis. Such changes created the need to revise previous data for category 3 SPADs in order to reflect the new definition and provide a consistent basis for comparison. This reclassification also illustrates how structural changes on a reporting system can have knock-on effects both on efficiency metrics and more direct forms of risk assessment. The revision indirectly increased the efficiency of the reporting system in terms of the severity of incidents being analysed within a particular budget. It also reduced overall system safety measured in terms of risk metrics, including the Precursor Indicator Model mentioned in the previous section.

The complexity of using submission and frequency metrics to assess the performance of reporting systems has persuaded many managers to concentrate on monitoring the costs of their schemes. This approach is justified by a series of interviews and focus groups that the FRA conducted to identify concerns about the regulation of the US rail system [247]. This study was intended to identify the influence of corporate culture on compliance with railroad operating rules. Part of this work focused on the attitude of operating companies to the opportunities provided by incident and accident reporting. There was a degree of skepticism about whether the costs incurred in analysing 'near accidents' could be justified by the potential insights that they provided. The rare nature of these events implies that 'one must make any number of assumptions' to identify the potential root causes and that this 'inevitably reduces the level of certainty' Many in the industry were concerned about the large numbers of near-miss events that must be investigated to gain limited insights about a small number of actual incidents; 'analysing near-incident data substantially increases the population data set from which to study'. In particular, the FRA study found that the 'systematic analysis of probable cause' for near miss incidents involving the safety conduct of locomotive engineers 'is seldom conducted' under Federal Regulation 49 CFR 240.309 [247]. To address these costs issues, the FRA report proposed that greater use be made of automated data analysis tools from reporting schemes in other industries. These tools have been reviewed in Chapter 13.5. In particular, they argued that Internet and Intranet technologies should be exploited to reduce the costs associated with the analysis of near-miss incidents.

Some of the concerns identified by the FRA stem from the difficulty of assessing the costs associated with running a reporting system. It might seem relatively straightforward to account for the fixed costs that are associated with infrastructure items, such as computer hardware. Most of this equipment serves several different purposes. Incident reporting software is often integrated into other aspects of a safety management system. It can, therefore, be difficult to distinguish the costs of running such a system from the wider overheads associated with risk analysis and assessment [781]. Similar comments can be made about the difficult of auditing variable costs. Many reporting systems rely upon the involvement of volunteers who combine incident analysis with more a more direct operational role [660]. Cost estimation for incident reporting is further complicated by the consequential overheads that are associated with some investigations. For example, the UK Health and Safety Executive make a charge for inspectors' time under The Health and Safety (Fees) Regulations 2001. Employers will be charged for the 'investigation of activities or workplaces where HSE becomes aware of an incident which has caused or is liable to cause injury to persons; and related enforcement work... including the preparation and serving of improvement or prohibition notices; assessing and issuing exemptions" [331]. Employers' scope for cost reduction is further constrained by the need to support a reporting system as a condition of operation. For example, the Safety Case for operating the London Underground contains a specific commitment to operate such a system; 'employees can raise health and safety concerns either formally with management or informally via their employee representatives... London Underground Ltd intend to join the CIRAS confidential reporting system when it is eventually rolled out across the national railway network'

[335].

It can be just as hard for regulators to justify and account for their expenditure on incident reporting systems. For instance, the FRA required $106,855,000 for 'safety and operations' in 2001. In 2002, they requested $120,583,000 while in 2003 this had risen to $122,889,000. It can be difficult to break these figures down to identify the individual sums spent on the diverse range of incident and accident reporting systems that are supported by the FRA within the US Department of Transportation. However, the 2003 budgetary request provides an important insight into the strategic importance of these systems when it justifies the additional expenditure. This additional money is intended to fund 20 new safety field inspectors because 'the number of railroad issues facing FRA is increasing and becoming more complex' [244]. It is hoped that these posts will support the 'elimination of transportation-related deaths, injuries, and incidents'. The FRA's request does not monitor the efficiency of their reporting system in terms of the cost per submission, as suggested in previous paragraphs. They do, however, cite a number of additional statistics to support their request for additional funds. These are similar to the normalising factors that are used in the calculation of direct measures for reporting system performance. The budget request focuses on the 55% increase in freight traffic since US deregulation in 1980 [244]. They also stress the unfortunate rise in rail passenger fatalities and injuries from approximately 500 per year over the past decade to over 660 in 2000. The increase in rail traffic and in adverse event is, therefore, used to justify increased regulatory expenditure on field investigations.

To summarise, performance metrics focus more on the operation of the reporting system than the direct measurement of 'system safety'. For example, the performance of a reporting system be assessed in terms of the number of incidents that are analysed within a specified budget. The efficiency of the system can be improved either by budget reductions or increases in the number of incidents that are handled by the system. Unfortunately, a number of pragmatic and theoretical concerns affect the use of such monitoring techniques. Raw data about the number of submissions made to a reporting system can be very misleading. For instance, they can hide under-reporting by particular groups. Any increase in reporting frequency might, therefore, yield few marginal benefits if those reports stem from communities that are already well represented with the system. It is also possible to distort submission statistics by changing the definition of what does and does not fall within the scope of the system. Alternatively, efficiency can be assessed in terms of savings that can be made from the operation of a reporting system. Opportunities for cost reduction may, however, be constrained by regulatory agreements that require the operation of particular schemes. The increasing complexity of many application processes also makes it difficult to reduce the costs associated with incident reporting. Both the FRA [241] and HMRI [321] have been forced to find and fund an increasingly diverse range of skills during the investigation of many recent failures. Staff costs represent the greatest single investment in most reporting systems. In consequence, managers have begun to refocus their monitoring activities on the performance of their investigators rather than on more direct metrics that can be both difficult to gather and harder to interpret.

15.2.2 Investigator Performance

Many regulatory organisations have developed detailed guidance for the investigation of adverse events and near-misses. For example, Transport Canada include such advice in their recommendations for the development of railway safety management systems [780]. Operating companies must develop 'procedures for internal and external accident and incident notification and reporting, including third-party reporting; procedures, formats and approaches (e.g., site protocol) for investigations (e.g., environmental, employee injuries, transportation of dangerous goods); a formal link to the risk management process; and procedures for reporting and documenting findings, conclusions and recommendations, and for ensuring implementation of recommendations and corrective actions'. The Safety Management System guidelines go on to argue that most train accidents can be prevented and that investigators must, therefore, identify ways of providing both 'immediate protection' and 'long-term correction'. Examples of an immediate action include the introduction of a 10 miles per hour temporary speed restriction at the site of a track geometry defect or a 40 miles per hour speed restriction on a type of car that appears to be unstable at higher speeds. These immediate protective actions must be implemented by the Investigating Team before operations are resumed. Long term 'corrections' reduce the likelihood of a similar train accident recurring in the future. Examples include the 'accelerated removal of straight plate wheels and the overhaul of trucks on a specific class of car' [780].

Such guidance reflects the way in which many regulatory and investigatory agencies have sought to support the work of incident investigators. In particular, it is typical of the way in which general recommendations are not supported by more detailed assessment criteria that might be used to monitor the performance of particular teams and individuals. This is a significant concern which is shared by many of the organisations that operate incident reporting systems [423]. Often investigators are domain experts, drawn from diverse operational areas of the industry that they are helping to protect. This offers numerous benefits in terms of their detailed understanding of industry practices. It also creates significant weaknesses because many investigators have only a rudimentary understanding of the more detailed causal analysis techniques pioneered by NASA [572], the US Department of Energy [208] and the NTSB [87]. There are further problems. The previous background of an investigator can have a powerful influence on their likely findings. Lekberg describes the correlation between an investigators area of expertise and the likely results of a causal analysis [484]. The problems of frequency and recency bias have also been described in Chapter 10.4. In particular, the fact that investigators may have already spent many years within an industry can imply a lack of understanding of more recent technical innovations. For example, new insights in the field of human factors are often slow to inform the conduct of incident investigations [700].

There are relatively few published studies into the performance of investigators. This is a significant barrier to the future development of incident reporting systems. One consequence is that many of the organisations that operate these schemes are concerned about potential inadequacies in their analysis and interpretation of individual reports. Unfortunately, the lack of previous published work in this area has created a general reluctance to assess or otherwise measure the extent of the problem. There are several exceptions. For example, the UK HMRI conducted a recent analysis of the work that investigators performed in analysing the causes of SPAD incidents. The HMRI enquiry 'looked at the results of several SPAD investigations carried out in accordance with procedure GO/RT3252 in each Railtrack zone, and it appeared that in some cases greater emphasis was placed on completing a multi-page form than getting to the root cause of the SPAD incident' [351]. They describe examples in which the same signal had been passed at danger on repeated occasions and yet the cause had not been established. It could be argued that this reflects a lack of evidence available to any investigation, however, the investigatory procedures stressed the need to reach some form of closure for each enquiry. In other cases, investigators had failed to follow through their analysis of an adverse event. For example, one investigation concluded that the driver had an 'unsuitable temperament' for driving suburban trains. He was, therefore, barred from driving these services. The HMRI inspectors argued that the investigation should have gone on to deal with the root cause of the incident which they interpreted to be the 'inadequacy of the measures for assessing the competence of drivers' [351]. The inspectors also found that SPAD investigators had difficulty in distinguishing between some of the causal categories identified in the supporting documentation. In particular, it was often unclear whether an incident was caused by 'misjudgement' or 'disregard' as it implied an assessment of driver intentions. The HMRI inspectors cite the example of a SPAD that occurred in poor weather conditions where the driver made every effort to stop the train before the signal. They argued that this was incorrectly classified as 'disregard' rather than 'misjudgement'.

Arguably the most straightforward means of assessing the performance of incident investigators is to introduce self-monitoring throughout team-based enquiries. This approach has been widely adopted and is often modelled on the 'Go Team' procedures that were initial developed by the NTSB during the 1960s and 1970s. Each Go Team is led by an Investigator-in-Charge who is a senior investigator with several years of NTSB and industry experience. Each member of the team is responsible for a defined aspect of of the investigation. For example, the 'operations specialist' will reconstruct the history of the incident including the crew members' duties for the period before the incident. The 'structures expert' will document the accident scene. They will analyse any wreckage and will also calculate impact angles and speeds. The 'human performance specialist' will make a study of crew performance and all before-the-accident factors that might be involved in human error, including fatigue, medication, alcohol. They will also consider the possible role of drugs, medical histories, training, workload, equipment design and work environment [619]. Locomotive engineers, signal system specialists and track engineers head working groups at railroad accidents. Each of these specialists heads a working group in their area. The members of the working groups are drawn from 'interested parties'. The party scheme is the key to the NTSB's efficiency; it investigate approximately 2,500 incidents per year with only 400 employees. The NTSB designates other organisations or corporations as parties to the investigation. These parties must provide specific expertise to the investigation. There is considerable freedom about who might provide such assistance; the only exclusion is that 'persons in legal or litigation positions are not allowed to be assigned to the investigation' [619].

Not only does the party system increase the efficiency of the Board's full-time staff. It also provides an important means of cross validation and of monitoring the quality of the investigation process. The head of each working group prepares a factual report and each of the parties in the group are asked to verify the accuracy of the report. Unfortunately, the formal procedures and mechanisms that support NTSB investigations are resource intensive. Most incident reporting systems lack the resources necessary to finance the involvement of more than one or two investigators in the analysis of adverse events. There are further problems. For example, the NTSB are an independent organisation that operates across several different industries. It is difficult to see how such a multi-party architecture might be used by a proprietary reporting system which focuses on adverse events within a single commercial organisation.

Fortunately, a range of alternative techniques can be used to monitor and support the performance of incident investigators. For instance, the ATSB was formed with the specific aim of pooling expertise in transport safety. It created a unified framework for the Bureau of Air Safety Investigation, elements of the Federal Office of Road Safety and the Marine Incident Investigation Unit and a new Rail Safety Unit. The intention was to 'make safety investigations even better as a result of sharing resources, ideas and techniques' [54]. The Bureau encourages investigators to move between incidents in different modes of transport. This ensures that key skills acquired in the investigation of rail incidents support the analysis of aviation or maritime incidents and vice versa. For instance, the ATSB reports that air safety investigation techniques were applied to the freight train collision at Ararat in November 1999 [54]. This exchange of expertise can also be seen to support the monitoring of individual investigators who must move beyond the immediate area of their core expertise. The Canadian Transportation Safety Board have similar objectives. Not only have they sought to improve the exchange of expertise between different investigation modes, they have also attempted to increase consistency in the resources that are available to investigators. In particular, they have developed a multi-modal Statement of Requirements that emphasises the importance of recorded information for investigative purposes in aviation, marine, rail, and pipeline incidents. The intention is to provide investigators with sufficient information to ensure that their reports achieve a level of 'reliability, comprehensiveness and timeliness' regardless of the mode of transport [88]. These multi-modal approaches are innovative and challenging. It remains to be seen whether they will realise the benefits that their proponents anticipate. They do not, however, provide a panacea for incident reporting. These approaches seem to offer more support for accident investigation because skill transfer requires the additional resources associated with maintaining a pool of investigators. Many reporting systems rely upon the analytical capabilities of one or two investigators [119].

Training programs offer alternative means of both monitoring and supporting the performance of incident investigators. Many reporting systems offer 'refresher' courses. These sessions can be used to introduce new incident and accident analysis techniques [36]. They also provide opportunities to assess and compare investigators' performance in the analysis of case study incidents. Most organisations lack the resources that are necessary to move beyond relatively ad hoc re-training programmes. In contrast, the NTSB is in the process of establishing an Academy for transport accident investigators. This is scheduled to begin operation in April 2002. It will be based in George Washington University adjacent to the U.S. Department of Transportation's National Crash Analysis Center. The intention is that the Academy will build upon existing Investigator Training Courses that are currently only held every six months. The Academy will extend the curriculum and provide a focal point for retraining. It will also provide focussed instruction in the different areas of expertise that are included within the NTSB's multi-party system, described in previous paragraphs. For instance, the reconstructed wreckage of TWA flight 800 will be held at the Academy 'so that future generations of aviation professionals and accident investigators from around the world can learn the lessons that it has to teach' [616]. The Academy will also provide a focus for investigators across the 'international investigative community' [613]. The establishment of this institution forms part of Jim Hall's response to the increasing complexity of many transportation incidents and accidents.

Previous paragraphs have identified a range of techniques that can be used to monitor the performance of incident investigators. Periodic reviews, such as that performed by the HMRI, can identify widespread failures in the analysis of particular failures. Accident investigations can also help to identify more systemic failures. For example, the Cullen report into Ladbroke Grove argued that a 'no blame' culture may paradoxically make staff more likely to accept responsibility for adverse events [195]. It is difficult to see how such reviews might be used to monitor the everyday activity of individual investigators. In contrast, many accident and incident reporting systems rely upon team-based techniques to validate the results of an investigation. The inter-modal exchange of key personnel and the NTSB's 'go team' concepts all explicitly allow for the cross-checking of any analysis before it is released beyond the agency that conducted the investigation. Finally, periodic retraining can be used to ensure that staff are brought 'up to date' with recent developments in investigatory techniques and in specialist areas such as meteorology and structural engineering. The NTSB's new investigator's Academy arguably represents the most significant recent development in this area.

The techniques described above are based around 'traditional' forms of monitoring. Team-based validation, the exchange of personnel, training and retraining have all formed core techniques of Human Resource Management for several decades. There have, however, been a number of more recent initiatives that offer new opportunities to monitor the performance of individual investigators. Many of these techniques are based around the novel computational systems that have been described in Chapter 13.5. For instance, many incident reports are now stored using relational databases. These systems enable managers to continuously monitor the performance of individual investigators. The same techniques that enable them to identify patterns of failure in a database of incident reports can also be used to identify patterns of analysis or bias in the findings of an individual investigator. In initial trials, we have begun to explore the effects that such information can have upon the overall management of a reporting system. For example, some managers have preconceived ideas about an ideal distribution of causal factors across an incident database. Individual investigators who exhibit a different pattern of analysis are encouraged to look more carefully for those factors that have been neglected in their previous reports. This can lead to potential problems if investigators feel unreasonable pressure to produce a particular pattern of causal findings almost irrespective of the incidents that they have been asked to analyse. Our initial studies have identified further effects, some of which were less easy to predict that the attempts to 'enforce' of normative causal distributions. The intention behind providing an individual with information about the results of their previous investigations is to reduce the problems of frequency and recency bias. There is, however, a danger that this information can have the opposite effect. Showing an investigator that they have identified human error in all recent incidents can reinforce rather than challenge their tendency to identify this cause!

These extensions of existing relational technology are relatively unsophisticated. More recent information retrieval systems offer a number of alternative monitoring tools. Many search engines now routinely construct user models. These models are based around information about an individual's previous retrieval requests. They can be used to make inferences about the user's future information requirements. For example, a frequently used data source may given a higher weighting than one that the user has seldom visited in their previous interactions. User models can also provide insights for the manager of a reporting system. For example, they can be used to determine whether or not an investigator has accessed information about particular aspects of an incident. An investigator's report might rule out human factors as a probable cause even though they have only performed a cursory analysis of the evidence in this area. Conversely, they might focus on particular aspects of an incident to the exclusion of alternative hypotheses.

15.2.3 Intervention Measures

Previous sections have argued that a reporting system can be assessed in terms of the recommendations that it produces. In other words, managers can monitor the effectiveness of the remedial actions that are identified in the aftermath of an adverse event. This information can be used in a number of ways. Previous sections have focussed on the direct benefits that such insights can have upon the operation of safety critical applications. They also fulfill a wider role in helping regulators, the public and government to monitor the operation of a reporting system. A successful scheme should continue to identify areas for improvement. For example, the NTSB issued a document entitled 'We Are All Safer' which stresses the impact that their investigations have had across the transportation industries [606]. They cite safety improvements from more than 150 recommendations in rail passenger car equipment and design, injury reduction and train collision avoidance. These have resulted in seats that are now secured against movement in the event of a collision or derailment. NTSB recommendations are also argued to have encouraged the installation of shatter-proof windows that can also be used as emergency exits. They have caused the installation of overhead luggage racks that have effective retention devices. The NTSB also point to the introduction of passenger emergency briefing cards and too the use of conspicuous levers to help in the operation of doors and emergency windows. The survey also directs the reader's attention towards the less 'visible' effects of their investigations. NTSB recommendations have led to the replacement of railway car construction materials to meet flammability, smoke emission, and toxicity standards. Their analysis has also encouraged the development of new procedures for emergency passenger car evacuation and revised training programs in emergency procedures for service employees. They have helped to introduce mandatory speed and signal compliance checks in certain regions. They have encouraged the use of written notifications to inform employees of speed restrictions and special permission procedures for trains entering out-of-service track sections. NTSB recommendations have also helped to introduce regular crew fitness for duty checks. They summarise the impact of their activities by emphasising the long-term effect of their findings on the industry regulator:

> "As a result of years of rail passenger safety recommendations from the Safety Board, the FRA is enacting regulations regarding passenger equipment safety standards and passenger train emergency preparedness. These regulations will implement many of the recommendations the Safety Board has made to the FRA and the railroad industry to improve the crashworthiness of rail passenger cars and locomotives." [606]

It is possible to criticise the use of such examples to indicate the 'vigour and vitality' of a reporting system. Several of the innovations identified by the NTSB were already being introduced prior to their recommendations being published. It can, therefore, be argued that their intervention helped to expedite changes that were already being made within the industry. This emphasises a point that has been made repeatedly throughout this book. Most incidents and accidents reveal problems that are already well-known by safety managers and other operational staff. Adverse events and near-miss incidents help to focus attention and increase the priorities associated with existing safety concerns. This need not undermine the argument that successful recommendations provide evidence about the health of a reporting system. The NTSB's support for existing initiatives indicates a healthy relationship with regulators and other industry bodies. Equally, there is a concern that any investigatory organisation is independent of such external influences. For instance, the NTSB review describes an incident at Silver Springs when a Maryland commuter train ignored a signal and collided with an Amtrak passenger train. This is the accident described in Chapter 2.3. The Safety Board found that the crew failed to obey the signals because of multiple distractions and the failure of federal and state regulators to analyse the human factors impact of signal modifications on that rail line. This discussion illustrates two key points. An effective reporting system cannot simply be assessed in terms of the recommendations that it issues. Any monitoring must also account for the way in which those recommendations are received and implemented by operational and regulatory organisations. If all recommendations are accepted then it can be argued that this indicates too close a relationship between the investigators and the recipients of any proposed intervention. Conversely, investigators may propose interventions that are consistently rejected by regulators or more senior safety managers. It might be argued that such situations illustrate the perseverance of an investigator fighting for necessary safety improvements. It can equally be argued that the consistent rejection of proposed recommendations illustrates a break-down in the operation of the system. In either case, serious concerns can be raised about the effectiveness and efficiency of the reporting system.

Mode	Number Issued	Percentage of Total	Acceptance Rate
Aviation	4214	36.2%	82.61%
Highway	1865	16.0%	88.48%
Intermodal	225	1.9%	76.34%
Marine	2234	19.1%	74.75%
Pipeline	1192	10.1%	85.62%
Railroad	1941	16.7%	81.57%
Total	11770	100%	81.99%

Table 15.10: NTSB Safety Recommendations Issued by Mode [617]

Table 15.10 summarises the relative acceptance rates for NTSB interventions in each of the modes of transportation that they are responsible for. As can be seen, the reputation and authority of NTSB recommendations helps to ensure relatively high levels of agreement. However, the difficulty of establishing consensus in the maritime industry, noted in Chapter 12.4, is arguably again illustrated in this table. It is also important to note the relatively low number of inter-modal recommendations in Table 15.10. This is surprising given that the NTSB is often cited as an example of inter-modal investigation techniques being used to learn lessons that are common to many different industries. There arguments supported both the creation of the ATSB and the US Chemical Safety and Hazard Investigation Board. The relatively low proportion of inter-modal recommendations might be explained by the need to identify a regulatory or industrial organisations to receive such proposals. However, inter-model recommendations can be addresses to more than one recipient. This observation may also reflect the traditional model in which recommendations are focussed towards the particular circumstances of the incident or accident that is being investigated. It remains to be seen whether the inter-modal initiatives being launched by the ATSB will yield a greater proportion of such recommendations than those of the NTSB, summarised in Table 15.10.

Table 15.11 provides a more detailed break-down of the status associated with the NTSB's recommendations following rail-related incidents and accidents. As can be seen, a relatively complex classification system is used to monitor the performance of investigations in this domain. For example, a distinction is made between 'open' and 'closed' recommendations. Open recommendations deal with issues that the NTSB considers still to pose a significant threat to future safety. In contrast, 'closed' recommendations may have been superseded by changes within the industry. Alternatively, they may have been adopted and implemented or they may simply have been rejected by their intended recipients. This detailed breakdown is necessary if recommendations are to be considered as part of the monitoring process. A relatively low acceptance rate can be indicative a large number of different situations. It might suggest a break-down in communication between investigators and industry representations, 'Open–Unacceptable Response'. It can also suggest that recommendations have been amended through successful negotiation, this might be revealed by a relatively high proportion of 'Closed–Acceptable Alternate Actions'.

It is important to emphasise that Table 15.11 provides a very high-level overview of the status of particular recommendations. A reporting system is likely to have a different impact on different aspects of an industry as diverse as the US rail system. For example, many of the innovations that were cited at the start of this section focussed on passenger transportation. More detailed data is required so that safety managers can determine whether recommendations are more likely to be implemented in this area rather than in freight distribution or in rapid transit systems. The NTSB survey does address these different areas by enumerating the particular improvements that have been triggered by their recommendations [606]. For instance, the New York City Transit

Status of Recommendation	Number
Closed–Exceeds Recommended Action	4
Closed–Acceptable Action	1136
Closed–Acceptable Alternate Action	137
Closed–Unacceptable Action	270
Closed–Unacceptable Action/Superseded	14
Closed–Reconsidered	62
Closed–Superseded	19
Closed–No Longer Applicable	110
Total Closed	1752
Open–Acceptable Response	93
Open–Acceptable Alternate Response	2
Open–Unacceptable Response	26
Open–Response Received	15
Open–Await Response	53
Total Open	189
Total Issued	1941
Acceptance Rate	81.57%

Table 15.11: NTSB Railroad Recommendation Status [618]

system has introduced standardisation braking distances and testing procedures. It has also installed speedometers and improved speed control signage. Similarly, a collision involving Greater Cleveland Regional Transit Authority (GCRTA) trains revealed that a train operator had disconnected the automatic cab signal system that provided one form of collision prevention. Coded track circuits were used to transmit speed commands to the on-board train control equipment. To avoid the speed limitation, the operator cut the cab signal to deactivate the control system. As a consequence of the NTSB recommendations, the GCRTA implemented procedures for recording the use of cab-signal cutouts to prevent unauthorised operations.

These specific examples are not supported by more detailed statistics about the frequency of recommendations or the status of those proposals that have already been made. One of the difficulties in this area is that recommendations must be addressed to the many different state and local government organisations. These bodies have the primary responsibility for the safety of the two billion passengers that use Rapid Transit systems each year. Different safety oversight procedures operate in each of these systems. This raises a further more general point; the acceptance of a recommendation by a regulator or other safety body does not imply that the recommendation will be successfully implemented at an operational level. It is, therefore, important that investigatory organisations monitor the effectiveness of their accepted proposals and not simply the overall rate of acceptance, as illustrated in Table 15.11. This is illustrated by the way in which NTSB recommendations argued for mandatory drug and alcohol testing in from the 1970s through to its introduction on US railways in 1986. By monitoring the results of these tests, it was argued that the regulations had helped to reduce substance abuse. Post-accident tests indicated that the number of employees with positive test results fell from 5.5% in 1987 to less than 1% in 1995. Random drug tests showed a similar decline from 1.04% in 1990 to 0.9% in 1995. The success of other recommendations is less easy to establish. For example, the NTSB investigated 29 locomotive derailments in 1991. Diesel fuel spills occurred from ruptured tanks and lines in more than half (56%) of these incidents. The Board issued recommendations that resulted in a joint meeting between the FRA, the Association of American Railroads and locomotive manufacturers. This resulted in a program to collect further data on fuel tank damage and fuel spills. The results of this initiative have taken time to assess because of the delay between revised equipment design and the widespread introduction of these devices across the network. Given the relatively low frequency of adverse events, it took until 1997 before the Board was called upon to investigate two passenger train derailments involving locomotives with the revised 'integral' fuel tanks. The fuel tanks on-board these trains were integrated into their frame structure rather than being suspended within the frame. Integrally tanks also provide higher ground clearance than conventional designs. Investigators concluded that the performance of these enhanced designs 'clearly outperformed frame-suspended fuel tanks' [606]. There was less fuel tank damage and no significant spillage in either of the accidents despite serious track damage. It is, however, difficult to quantify these improvements without considerable additional analysis give that it relies upon a variant of the counterfactual reasoning that has been described in Chapter 10.4. Investigators are forced to compare the consequences of incidents that might have occurred had the trains been fitted with more conventional fuel tanks.

The arguments cited in previous paragraphs have been drawn from a document that was deliberately intended to promote the investigatory work of the NTSB [606]. It, therefore, provides a relatively positive view of their role in ensuring the safety of complex applications. In contrast, it is possible to identify a more pessimistic or pragmatic view in the technical publications of other investigatory organisations. For instance, the HMRI report into SPAD investigation found that the delays in implementing previous recommendations often meant that new incidents had occurred before previous ones were adequately addressed [354]. Even though one signal at Birmingham New Street had been passed at 'Danger' seven times in eight years, it still took 12 months from the last SPAD incident to install countdown markers. In Railtrack Great Western Zone (RTGWZ), it took ten months to install long hoods on Reading signal R242. The investigators also found strong regional variations in the implementation of recommendations. This led to a meta-level finding that all 'recommendations should be time bound and operating companies should more actively track their completion by setting up their own (monitoring) systems' [354].

It is important to stress that the issues described in this section are generic. They do not simply affect the rail industry. For instance, a recent inspection of UK nuclear facilities also focussed on the efficiency and reliability of monitoring systems that are intended to support the implementation of recommendations from incident reports. They found that some recommendations from incident investigations remained 'incomplete' while others had been expedited as a matter of priority.

> "Although we found information on the state of close-out was being passed to managers, we found little evidence of it being used by managers. Often the only people we could find who were concerned about overdue recommendations were relatively junior staff tasked with keeping the action tracking database up to date. Generally we saw no effective monitoring by managers of those people responsible for closing out recommendations." [642]

Delays and regional variations in the implementation of recommendations are not simply symptomatic of inadequate monitoring. They can also reflect deeper problems, including opposition by both regulators and operators. The following section, therefore, identifies metrics that might be used to monitor the credibility and acceptance of a reporting system.

15.3 Acceptance Measures

The previous section has described ways in which the efficiency of a reporting system can be judged in terms of the recommendations that are implemented in the aftermath of adverse events. There are other ways of reaching similar assessments. For instance, it is possible to monitor the health of a reporting system in terms of those recommendations that are adopted by a regulator but which are violated by their ultimate recipients. The scale of such violations is illustrated by Table 15.12. This summarises the enforcement actions that were initiated by the UK railways Inspectorate between 1996 and 2000. As can be seen the provisional figures for 2000-2001 show a record number of enforcement actions. It might be argued that this illustrates a rising rejection both of the recommendations derived from previous incidents and of the general regulatory framework that supports the UK rail infrastructure. As with previous statistics, however, things are not so straightforward. We have described how a number of high-profile accidents together with structural changes in the infrastructure company have focussed public attention on the safety of the UK railway. It might, therefore, be argued that this rise in enforcement actions is less a reflection of the outright increasing rejection of safety regulations than it is the result of pressure being applied to the Inspectorate to increase their monitoring activities.

	96/97	97/98	98/99	99/00	00/01 (provisional)
Enforcement notices	24	33	21	45	51
Prosecutions heard	6	8	10	11	12
Total fines (£)	233,500	67,500	695,000	1,899,500	1,115,000

Table 15.12: HMRI Railway Enforcement Actions (1996-2000) [336]

This section looks beyond enforcement statistics to identify further metrics that might be used to access the acceptance of a reporting system, not simply the recommendations that it produces. Before looking in detail at these assessment techniques it is important to stress that most reporting systems rely upon the cooperation of many different groups. It, therefore, follows that some of these groups will exhibit different degrees of involvement in a reporting system. For instance, a regulator may offer strong encouragement for the introduction of a system that is opposed by line management in operating companies. Alternatively, Trades Union representatives might support the operation of a reporting system that is not fully supported by regulators. It is also important to emphasise that the public statements of support from some of these groups do not necessarily imply that other, similar organisations will share the same sentiments. For instance, the Transport Salaried Staffs' Association is an independent trade union that represents members in the railway industry, the travel trade, London Underground/Transport for London and London buses as well as road haulage, shipping and ports. They have offered strong support for the CIRAS confidential reporting system mentioned throughout this book. In a recent newsletter they summarise the recent changes that have extended this system throughout the United Kingdom. CIRAS 'is open to Railway Group members and other participating companies, and comprises a core facility supported by three regional centres'. The University of Strathclyde operates the new core facility and runs one regional centre. The other centres are run by a consultancy firm, W.S. Atkins, and a group within the UK's former Defence Evaluation and Research Agency. Railway Safety, an independent company with links to the infrastructure operator, funds the cost of the core facility while the regional centres are paid for by participating companies. The Transport Salaried Staffs' Association note that CIRAS will help employees to report concerns to the regional centres. Centre staff will then conduct follow-up interviews and provide data to a national database. The Association 'supports the important work of CIRAS and is confident that its members in the rail industry will contribute to its effective functioning' [782]. In contrast to this positive message, a number of operating companies expressed initial reluctance to join the scheme. A range of concerns focussed on the usefulness of the data that the system might produce, on the management and confidentiality of the scheme and on cost projections for a national system. This led the Deputy Prime Minister, John Prescott to state that the reporting system would be introduced 'whether or not' the train operating companies wanted it [102]. He argued that CIRAS is "an essential tool to restoring confidence in the industry and getting the actual facts of what is going on". In the aftermath of the Ladbroke Grove accident, however, several industry representatives argued that CIRAS could only provide a short term solution. The need to operate a confidential reporting system was seen as an indictment of the safety culture in the industry because many employees were reluctant to speak openly about safety concerns. A seminar held in preparation for the Cullen enquiry into this accident reached the following conclusions. The final sentence in this statement indicates some of the tensions that can arise between employers and Trades Unions in both the operation of a reporting system and the wider monitoring of safety concerns:

> "There is a problem of finding volunteers to represent the workforce as safety representatives, although this is not universally accepted. Where problems had been encountered,

> it was due to either complacency or employees who were too frightened due to potential victimisation. Trade union representatives can be seen as a nuisance factor and this is an inherent problem of the railway culture. However, it was acknowledged that unions did have a significant part to play in the area of communication, but not at the expense of the normal company communication channels." [466]

Such concerns illustrate the complex and differing attitudes that characterise reactions to voluntary, confidential reporting systems. A similar diversity of opinions can be observed in opinions about mandatory reporting schemes. For example, a review of standard setting across the UK railway found a number of conflicting attitudes towards the value of existing incident and accident investigation procedures [329]. Some of the groups contacted expressed confidence in existing arrangements. For example, the infrastructure company pointed to the introduction of fully independent Chairmen for more important internal inquiries. Others argued that investigations were still conducted with 'insufficient openness'. As a result, other groups such as insurers and consultants had to demand site access. This, in turn, was perceived to have increased concerns over liability rather than focus attention on safety improvements. The different opinions expressed about both mandatory and voluntary reporting systems illustrate the way in which different groups can express different degrees of satisfaction with the same scheme. The following sections argue that these different attitudes can also indirectly influence the effectiveness of many reporting systems. If key groups of workers remain unconvinced about the usefulness and confidentiality of a system, or of regulatory and managerial involvement, then they may be reluctant to participate in its operation. It is, therefore, important to monitoring the acceptance of a reporting system so that such problems can be both detected and addressed before they compromise the effectiveness of a reporting system.

15.3.1 Safety Culture and Safety Climate?

Safety culture forms part of the wider corporate culture that can be used to 'distinguish one organisation from another' [303]. It can be difficult to derive a precise definition of what constitutes a strong safety culture. For instance, Pidgeon and O'Leary identify four different concepts within this term: "responsibility for strategic management; distributed attitudes of care and concern throughout an organisation; appropriate norms and rules for handling hazards and on-going reflection upon safety practice" [682]. Conversely, Reason argues that an organisation embodies rather than possess a safety culture [702]. In other words, the development of a strong safety culture requires 'root and branch' changes to managerial and organisational structures. It cannot simply be grafted onto an existing institution.

Unfortunately, it can be difficult to apply these high-level observations to analyse the particular characteristics of complex, real-world organisations. These problems can be illustrated by a recent review of safety across the Irish railway system. This identified an 'improved safety culture' in the removal of fire risks and in an improved working environment in the underpart of signal cabins. The report also argues that the introduction of elected Safety representatives has also had a positive impact upon safety culture. Such specific improvements can be contrasted with more general observations. It was argued that 'the culture of safety has still not taken root in the staff at ground level' [392]. Workers continue to expose themselves to hazardous track-side conditions with relatively poor protection arrangements. The report also argues that poor morale and the breakdown of management/employee relationships have also had an adverse effect on safety culture. Such generalisations can also be contrasted with specific observations about regional differences in the safety culture within the same organisations. Workers showed greater distrust about managerial attitudes towards safety in the Dublin area than elsewhere in the railway network. This mixture of specific observations and broad generalisations is typical of the types of analysis that are used to support conclusions about the 'safety culture' within particular organisations. Unfortunately, very few studies use the same general or specific observations to support their arguments about safety culture. In consequence, it can be difficult to determine what criteria can best be used to assess the performance of a particular organisation.

In spite of the problems in defining what is meant by the term, many regulators still cite the development of a 'strong safety culture' as a primary aim. This objective is often used to justify the introduction or revision of safety regulations, including the Safety Management System requirements imposed on railway companies by Transport Canada [779]. The maintenance and acceptance of an incident reporting system is often taken to indicate a positive safety culture. There is, however, a need for more detailed metrics to show that regulatory intervention has had the intended effect. The introduction to this section has illustrated the way in which regulators might use the number of prosecutions or enforcement actions as a crude indicator of the 'safety culture' across an industry. A falling number of enforcement actions might be interpreted as evidence that the insights from incident reporting systems are being acted upon without the need for regulatory intervention. A recent survey of attitudes across the UK rail network identified these links between violations and reporting behaviour:

> "If there are rules they should be complied with... A healthy culture would accept the challenge of compliance but would not accept non-compliance. Flagrant disregard of rules needs to be sanctioned in some way. A company should not sanction people for violations that have resulted in an accident if they do not sanction the violations that have not resulted in an accident. There needs to be consistency. For sanctions to be effective the rule that has been broken needs to be seen as legitimate. There is a greater chance of this happening if the employees that operate the systems have a chance to influence and comment on the rules. Therefore if they transgress, they are breaking their own rules. It is important that rules and amendments to the rules are communicated effectively to the workforce, with an explanation of the rule or change. Outlined above are the characteristics of a safety culture moving away from the blame culture and to a more just culture. A just culture will allow for more transparency and for candid reporting, but will not condone reprehensible action." [465]

Others have identified more general links between safety culture and reporting behaviour. For instance, Lucas distinguishes between three different types of safety culture [844]. He argues that the shared perceptions and beliefs that are implicit in these different 'models' can have a profound impact on the types of incidents and accidents that an organisation might experience. Firstly, some organisations exhibit a 'traditional' safety culture. In this approach, the causes of any failure are likely to be attributed to the inattention or carelessness of individual workers. From this it follows that disciplinary actions are the mst likely remedial actions [444]. Alternatively, a 'risk management' approach to safety culture is typified by an engineering view of the human involvement in incidents and accidents. Failures are the results of a failure to correctly design the workers' tasks to their capabilities. Recommendations will focus on changes in operating procedures and on retraining. Finally, Lucas identifies a 'systemic' safety culture. The causes of an incident and accident are analysed in terms of the total working context. In addition to poor task allocation and training, recommendations will focus on mismanagement, on poor communications, low morale, inadequate feedback. The distinctions introduced by Lucas are important because they emphasise the diverse nature of 'safety cultures' within an industry. There are, however, a number of limitations. For example, the idea of a 'systemic' safety culture lacks the clarity of many other 'systemic' approaches to system failure. It is unclear how to measure or even recognise when such an approach has been adequately adopted by an organisation.

This is a significant problem given that it is often necessary to specify some timelimit by which necessary changes should be implemented throughout an organisation; 'as to the length of time within which a company or an industry can see an improvement in their safety culture, many consider that if marked results are not seen within three to five years, then it is likely that the company or industry's approach to developing a good safety culture is flawed' [465]. For instance, the UK Railway Group has devised a safety plan that is intended to take a decade to be implemented based on the premise that 'it may take up to five years for a good safety culture to develop' [465]. It is important that appropriate metrics be identified to help establish when a 'good safety culture' has been achieved. This creates problems because many of the attributes of a safety culture, such as managerial attitudes, cannot easily be assessed or validated. Kjellen, therefore, distinguishes between the abstract notion of a safety culture and the idea of a safety climate, which can be measured [444]. A safety climate denotes 'such aspects of an organisation that are possible to measure by use of a questionnaire-based survey where the results meet statistical criteria for aggregation to the organisational level'. Unfortunately, he acknowledges considerable disagreement over the dimensions that might assess the prevalent safety climate within an organisations. These include:

- management attitudes and commitment;
- involvement of employees in safety management system;
- communication about safety matters between the groups in an organisation;
- risk perception and the attribution of cause in an incident or accident investigation;
- relative priority associated to safety in comparison to other production goals;
- adherence to safety rules and attitudes to the acceptability of rule violations;
- active search for new hazards before incidents take place.

Unfortunately, he also acknowledges that although research has been conducted into safety climate for almost two decades, 'the positive effect of measuring the safety climate for use in feedback to the organisation have yet to be demonstrated' [444]. There are further problems. As we shall see, it can be difficult to demonstrate the reliability of the various instructments that might be used to measure attributes of a safety culture [191]. Surveys that reveal particular attitudes from certain members of staff at particular moments in time do not always achieve the same results when issued to other members of staff or even to the same individuals at different times [397].

The difficulty in assessing the 'safety climate' of an organisation have not dissuaded people from advocating the use of these metrics to monitor incident reporting systems. For instance, the perceived success of the CIRAS system has been cited as evidence of a poor safety culture across the UK rail network. CIRAS supports 'the silent majority who are too scared to report incidents direct to their supervisors and senior management' [466]. Staff are worried about the reaction of their supervisor. They are concerned that they will be disciplined for reporting violationsi. However, it has also been argued that CIRAS can help to correct a deficient safety culture; 'the method of incident investigation is important in developing a proper culture' [466]. This link between the establishment of an incident reporting system and the development of an appropriate safety culture has also been recognised by the US Department of Transport. The Federal Railroad Safety Enhancement Act of 1999 sought 'to reduce human-factor causation of injuries, wrecks, and deaths by improving the safety culture in the railroad industry by expanding and strengthening existing statutory protections for employee whistle blowers' [239]. Statutory protections were extended to cover the reporting of injuries to the railroad, cooperation with an FRA or NTSB safety investigation and refusing to authorise use of equipment, track, or structures that the employee reasonably believes pose an imminent danger to human life. Such initiatives reiterate the expectation that reporting systems should play a positive role in promoting an appropriate safety culture. It, therefore, follows that measurements of the safety climate might trace the impact that a reporting system has upon an organisation. For example, a series of seminars conducted in the aftermath of the Ladbroke Grove and Hatfield accidents found evidence of a 'positive and pro-active' move away from the blame culture and towards a full root cause analysis during investigations. However, they also found that 'indications that frontline staff are not convinced' by initiatives such as the CIRAS Scheme [465]. There was scepticism about whether it was genuinely confidential. This initial concern was reported to have reduced, especially, when the confidentiality of small groups had been protected by analysts generalising the details of a particular incident. There were, however, still problems amongst middle management 'where it was most needed' [466].

The problems of using safety-climate metrics to monitor the impact of a reporting system are illustrated by the different attitudes to the CIRAS system. Some industry analysts that it actually hindered, rather than supported, the development of an appropriate safety culture amongst some workers. For instance, sub-contractors were deliberately excluded from the CIRAS system. Industry surveys revealed the sense of vulnerability and the concerns that sub-contractors felt about an 'us and them' attitude [466]. These concerns were exacerbated by insecure terms and conditions that were offered in response to the increasing financial pressures on the industry. Differences in the safety culture between direct employees and sub-contracting staff were also increased by the introduction of new contractors from the construction industries who lacked specialist railway knowledge. This led to further communication breakdowns, for example in the procedures used to hand-over critical tasks and in the monitoring of safe working hours. These concerns suggest that there are important differences in the safety climate within different sectors of the same industry. Attitudes towards the effectiveness of a reporting system might be very different depending on whether one asked a direct employee or a sub-contractor who was excluded from the scheme.

The overview of UK railways, cited in the previous paragraphs, argues that 'it is the lowest level of data, such as near misses and non-compliances that do not result in an accident or even adverse effects, that are indicative of the safety culture of an organisation' [465]. This creates a potential problem. We can use safety climate metrics to monitor the impact that a reporting system has upon the safety culture within an organisation. However, the safety climate within an organisation is assessed by monitoring the submissions to a reporting system. A number of alternative metrics might be used to assess the impact of a reporting system in terms of any changes to a safety culture. For example, previous sections have summarised a broad range of direct measures that include the lost time accident rate or the severity and frequency accident rates. Unfortunately, it can be difficult to agree on and then obtain the information that will be used in this way. Alternatively, safety culture might be assessed by looking at 'the quality of the relationships within a company and between companies and how effectively they consult and involve their staff' [465]. Such metrics may ignore the relatively flexible, informal communications channels that are used in smaller working groups across the rail industry [466]. Surveys of staff attitude can also be used to assess the safety culture in an organisation. Transport Canada has produced a checklist to illustrate this approach, based on work by Reason [625]. They emphasise that their safety culture checklist provides no guarantees of immunity from accidents or incidents and that complacency is safety's 'worst enemy'. Personnel and managers change so a high score may not be sustained unless the organisation shows constant vigilance. In the following questions, a score between 16 and 20 is indicative of a safety culture that is 'so healthy as to be barely credible'. Between 11 and 15, the organisation is in good shape. A score between Between 6 and 10 is 'not at all bad, but there's still a long way to go'. A result between 1 and 5 indicates that the organisation is very vulnerable. For each of the following questions, a 'yes' answer means that 'this is definitively the case in my organisation' and adds a score of one to the running total. An answer of 'do not know' or 'maybe' adds a score of 0.5. Responding 'no' or 'this is definitely not the case in my organisation' adds zero to the total.

1. "*Mindful of danger*: Top managers are ever mindful of the human organisational factors that can endanger their operations. (Yes/No/Don't know)
2. *Accept setbacks*: Top management accepts occasional setbacks and nasty surprises as inevitable. They anticipate that staff will make errors and train them to detect and recover from them.
3. *Committed*: Top managers are genuinely committed to aviation safety and provide adequate resources to serve this end.
4. *Regular meetings*: Safety-related issues are considered at high-level meetings on a regular basis, not just after some bad event.
5. *Events reviewed*: Past events are thoroughly reviewed at top-level meetings and the lessons learned are implemented as global reforms rather than local repairs.
6. *Improved defence*: After some mishap, the primary aim of top management is to identify the failed system defences and improve them, rather than to seek to divert responsibility to particular individuals.
7. *Health checks*: Top management adopts a pro-active stance toward safety...
8. *Institutional factors recognised*: Top management recognises that error-provoking institutional factors (under-staffing, inadequate equipment, inexperience, patchy training, bad human-

machine interfaces, etc.) are easier to manage and correct than fleeting psychological states, such as distraction, inattention and forgetfulness.

9. *Data*: It is understood that the effective management of safety, just like any other management process, depends critically on the collection, analysis and dissemination of relevant information.

10. *Vital signs*: Management recognises the necessity of combining reactive outcome data (i.e., the near miss and incident reporting system) with active process information. This involves the regular sampling of a variety of institutional parameters (scheduling, budgeting, fostering, procedures, defences, training, etc.), identifying which of these vital signs are most in need of attention, and then carrying out remedial actions.

11. *Staff attend safety meetings*: Meetings relating to safety are attended by staff from a wide variety of department and levels.

12. *Career boost*: Assignment to a safety-related function (quality or risk management) is seen as a fast-track appointment, not a dead end.

13. *Money vs. safety*: It is appreciated that commercial goals and safety issues can come into conflict. Measures are in place to recognise and resolve such conflicts in an effective and transparent manner.

14. *Reporting encouraged*: Policies are in place to encourage everyone to raise safety-related issues (one of the defining characteristics of a pathological culture is that messengers are 'shot' and whistle blowers dismissed or discredited).

15. *Qualified indemnity*: Policies relating to near miss and incident reporting systems make clear the organisation's stance regarding qualified indemnity against sanctions, confidentiality, and the organisational separation of the data-collecting department from those involved in disciplinary proceedings.

16. *Blame*: It is recognised by all staff that a small proportion of unsafe acts are indeed reckless and warrant sanctions but that the large majority of such acts should not attract punishment...

17. *Non-technical skills*: Line management encourages their staff to acquire the mental (or non-technical) as well as the technical skills necessary to achieve safe and effective performance.

18. *Feedback*: The organisation has in place rapid, useful and intelligible feedback channels to communicate the lessons learned from both the reactive and pro-active safety information systems...

19. *Acknowledge error*: The organisation has the will and the resources to acknowledge its errors, to apologise for them and to reassure the victims (or their relatives) that the lessons learned from such accidents will help to prevent their recurrence." [625]

There is an assumption that this questionnaire will be answered by individual workers reflecting on their experience of the organisations that employ them. This raises interesting issues. One individual can have a very different experience of an organisation than their colleagues within the same team. Individual events can have a profound impact upon answers to general questions such as 'the organisation has the will and the resources to acknowledge its errors...'. In consequence, it may be necessary to aggregate the individual views of many different workers to obtain an overall assessment of the culture within an organisation [864]. Unfortunately, this smooths out the regional and occupational differences that have been noted throughout this section. It is for these reasons that many investigators prefer not to talk about 'safety culture'. The measurement of 'safety climate' and 'corporate culture' raise similar conflicts between the need to generalise and the need to account for differences throughout an organisation.

Previous sections have argued that there can be difficulties in the identification and collection of direct measures for the safety improvements that are attributable to incident reporting systems. For example, a rise in the number of adverse events reported through the system might indicate that the reporting system has failed to deliver necessary safety improvements. Alternatively, it can be argued that the reporting system has helped to increase submission rates or that the incident rate might have been even worse without the reporting system. Further problems complicate the use of measures that focus on the efficiency of the reporting system rather than on safety improvements. For instance, there can often be significant disagreements between an investigator's findings and those of their peers. It can be difficult to distinguish whether such differences arise from the nature of the incidents that they have been asked to investigator or from particular forms of bias that may have affected their causal analysis. The previous section has identified a further set of problems that affect the use of less direct metrics as a means of monitoring incident reporting systems. For example, it is tempting to monitor the impact that a reporting scheme has upon the safety culture within an organisation. Unfortunately, many of the metrics that are used to assess 'safety culture' are themselves derived from the reporting system, such as submission rates. Other measures, such as subjective questionnaires, raise problems because the aggregation of individual returns can hide important cultural differences within a complex organisation.

15.3.2 Probity and Equity

Reason identified three different components of a safety culture: justice, flexibility and learning [702]. Previous sections have argued that flexibility and learning are essential if complex organisations are to respond to the insights provided by incident reporting systems. In contrast, the following paragraphs focus more on the issue of 'justice'. It is important to monitor perceptions about the probity and equity of such schemes in order to assess whether such schemes retain the confidence of potential contributors. Most incident reporting systems depends upon widespread participation in order to ensure that potential insights are not missed through opposition to the scheme itself. Even where there is widespread agreement about the benefits of a proposed system there can be subtle differences of opinion. For example, the General Secretary of the Associated Society of Locomotive Engineers and Firemen wrote to the Deputy Prime Minister in the aftermath of the Ladbroke Grove accident to express his Trades Union's concerns about the future of safety in the railway industry [30] Some of his arguments focussed on the need to increase the involvement of full time officers of the Unions in cross company safety meetings. This would support the exchange of safety information and would encourage 'an open safety culture'.

The General Secretary's views are both important and influential because they represent informed opinion and carry political weight within the industry. The survey based techniques that can be used to aggregate different attitudes towards safety culture often fail to account for such strategically important opinions. In contrast, safety managers must carefully consider the political weight of such views if they are to ensure participation in a reporting system. For example, the General Secretary went on to identify Union concerns over the punitive nature of many investigations. This results in a 'secretive culture' that stifles information sharing about safety issues. In contrast, he argues that the Society's representatives should be involved in setting up a 'no blame' policy for driver retraining following adverse events. His response reiterates the Union's support for the CIRAS initiative. The crucial point to consider here is, however, that this support is offered in the context of these wider safety concerns about secrecy, punitive investigations and the lack of consultation. Again, such issues cannot easily be extracted from aggregate responses to high-level questionnaires about 'safety culture'. In particular, the Union response cites reports from the infrastructure operator that some operating companies have blocking the extension of CIRAS because they fear vindictive employees would abuse the system. He argues that the opposition from operating companies also stems from a concern that spurious submission will waste the time of their managerial staff. The General Secretary argues that these concerns reflect a negative attitude that is based on outdated prejudices. It is 'a sad reflection on management within the railway industry'. Such comments illustrate the recursive nature of many safety concerns. Not only do they reveal the attitudes of the person writing the letter, they also reveal their attitudes towards the opinions that they believe others hold about a reporting system. It is difficult to construct direct question that might elicit such information. The following sections, therefore, describe alternative qualitative techniques that might be used to monitor particular attitudes towards the 'probity and equity' of a reporting system.

Informal interviews and focus groups can provide insights into the fears that particular individuals have about voluntary and confidential reporting schemes. These same techniques can also be used to expose the attitudes of managers and regulators that might not be so readily obtained from more formal questionnaires and surveys. The recent independent review of Australian rail safety provides an example of the way in which these techniques can provide important insights into attitudes towards incident reporting [55]. The review deliberately canvassed a wide range of opinions; 'the industry is diverse, and it was expected that different organisations would have varying views'. The study consulted major track managers; all major freight and passenger operators and a sample of the smaller passenger and freight operators; all rail safety accreditation authorities; the Rail Safety Committee of Australia; representatives of new entrants, rail client groups, workers safety committees; rail heritage and tourist groups; the Industry Reference Group Chair as well as Commonwealth and State agencies. They note the willing of these groups to discuss safety issues 'forthrightly and at length' and observe that there 'were many common themes, with differences often a result of the particular circumstances of the organisation concerned'. It is interesting to note that this review avoids any attempts to define or characterise a single 'safety culture' across the Australian rail network.

The survey was deliberately intended to solicit views rather than derive numeric values. Hence the monitoring can be seen as qualitative rather than quantitative. This makes it particularly important to justify the use of particular elicitation techniques and then to validate any subsequent interpretation of the information that is received. Unfortunately, the published accounts of the review provide summary information. Relatively little information is provided about the elicitation process and conclusions are often presented without reference to the supporting data that was obtained from the parties mentioned in the previous paragraphs. For example, the independent report argues that"there is a level of co-operation which is being achieved in specific instances between the industry participants and parties such as the coroner and the occupational health and safety authorities". This includes the sharing of evidence and interview results in the aftermath of incidents and accidents. Such observations are instructive because they illustrate important strengths in some regions. However, the report does not describe the reasons why or how this level of co-operation was achieved in particular locations. Such details, arguably, lay outside the scope of the review. The high-level nature of these comments may also reflect the need to protect the confidentiality of the contributors. It does, however, illustrate the way in which qualitative reviews can lack the grounding provided by the statistical analysis of more direct monitoring techniques.

The potential need for additional evidence is also evident when the report identifies problems with the existing arrangements for incident and accident investigation. It reports a residual concern 'that specific competencies are needed to assess the cause of railway incidents and this presents a danger that evidence needed to determine the operational or technical causes may be lost, or recommendations which compromise railway operational best practice may be imposed' [55]. Such observations illustrate how the results of qualitative surveys can be used to support the monitoring of incident and accident reporting systems. Unfortunately, additional details are required if regulators and managers are to address these high level criticisms. The independent review does provide some details in a subsequent analysis of the 1996 Intergovernmental Agreement on National Rail Safety between the Commonwealth, States and the Northern Territory. The various parties expressed concerns about many aspects of incident and accident investigation:

- there was a lack of clear protocols for the parties on site and for containment of the site after an adverse event;

- there were concerns about the independence of investigations undertaken by the regulator, operators, managers and other 'interested' parties;

- there was a perception that restrictions had been placed on the means by which the results of an investigation were communicated to the industry. There had been a failure to alert the industry of potential 'hot spots';

- there was concern over the different focus of investigations between those to which 'no-blame' was attached and those aimed at prosecutions. Participants were worried about the potential for manipulation of the self-incrimination provisions of rail safety legislation;

- there was a perception that undue delays had occurred in finalising investigations or making cause information available where litigation was expected.

These are valuable observations. For example, they indicate that investigation protocols should be drafted and then publicised to increase confidence in the results of any analysis. Assurances should also be given about the independence of investigations. Unfortunately, it can be difficult to prioritise these concerns so that management and regulators can prioritise their allocation of resources. Relatively little information is provided about how whether these concerns were shared across the national system or whether they were isolated within particular geographical regions and functional groups. Some information of this nature is provided. For example, 'most (of the respondents) argue for at least a minimum national role' in the coordination of safety management activities across the rail system. Similarly, the report states that 'most respondents' recommended that this minimum role should include the collation and analysis of statistical data on incidents and accidents, feedback on the causes of accidents through safety bulletins, the coordination of major incident and accident investigations, ensuring the 'standardisation of interpretation'. The report also identifies dissenting views from this majority opinion. National operators were concerned that proposals for a national organisation have not addressed the problems of multiple jurisdictions, of inconsistent analysis and of fees to support investigatory organisations. They also argued that most rail activity focussed on intrastate business and commuter functions. It, therefore, made little sense to focus so much attention on a national body. In addition to the dissenting views of national operators, the report also identified the concerns of 'many small operators'. Some of the these companies are isolated from the mainline network and, therefore, did not consider that national regulation should affect their operations.

This report, therefore, illustrates both the strengths and weaknesses of qualitative approaches to the monitoring of incident and accident reporting systems. Surveys can be used to sample the diverse views that help to form the different safety cultures within complex organisations. Unfortunately, the lack of empirical data can make it difficult to assess whether or not a particular opinion is shared by the majority within a particular group. Confidentiality agreements can also prevent analysts from providing access to the tapes and notes that support particular conclusions. They can also isolate the reader from the contextual information that might help to interpret particular comments. Many of the findings of these enquiries can appear to be based on supposition rather that the precise statistical findings of more direct techniques. Qualitative techniques can, however, also be used to summarise a broad range of opinions that might otherwise have been hidden within particular metrics. For instance, the previous paragraph identifies important differences between national operators, small scale companies and most of the remaining groups that were surveyed.

15.3.3 Financial Support

Previous paragraphs have described how qualitative techniques yield important insights about the fears that particular groups hold about reporting systems. Workers express concerns about the probity and equity of investigations that are coordinated by management on behalf of regulatory and investigatory organisations. Managers and operating companies are worried about the undue influence of national regulatory bodies. They are also concerned that employees will waste finite managerial resources by generating spurious reports. This latter argument introduces a further means of monitoring the performance of a reporting system. It can be argued that the funding arrangements, which support a scheme, can provide valuable insights into the perceived success of the system. If companies provide financial security without without regulatory obligation then it might be argued that the scheme is well respected. Conversely, continual funding reviews and a lack of invest might be interpreted as important signs that a reporting system is failing to provide valuable insights into necessary safety improvements.

This line of argument is supported by attempts to justify continued public investment in the work of the NTSB. In 1996, $850 million was allocated to the FRA to support a regulatory rail safety program. $39 million dollars was allocated to the NTSB to ensure the safety of all forms of transportation, including oversight of the Federal Railroad Administration. To place this in perspective, the Federal Transit Administration received $4 billion to fund rail transportation infrastructure and equipment purchases. The NTSB recognise the 'substantial investment' in rail transportation safety and, therefore, acknowledge their responsibility to act as 'the eyes and ears of the American people at accident sites' [302]. Similar comments can be made about the FRA's responsibilities to elicit information about 'less-serious' incidents and near miss events.

It is possible to identify a range of different funding mechanisms that have been used to support both incident reporting systems and the wider regulatory infrastructure that supports them. For example, the UK Civil Aviation Authority is unusual because it was established with the aim that the industry should pay the cost of its own regulation. Income is partly derived from from licensing charges. Such charging schemes can lead to inequalities. For example large operators may claim that they are subsiding smaller companies if relatively more regulatory time is spent on their concerns. Conversely, smaller companies might claim that they subsidise larger operators if a unit fee is charged irrespective of the size of an organisation. These comments have also been made about the levies that support reporting schemes [444]. In order to address these concerns, the CAA also raises income from a levy on airlines that is based on on passenger kilometres. In contrast to the UK CAA, the costs of US aviation regulation are recovered from transportation users by various indirect taxation. For example, through levies placed on passenger tickets. This can lead to further problems. For example, such charges have been criticised because of the impact that they can have upon particular forms of transport. For example, Australian rail operators have pointed to the relative subsidies that road transport operators receive in relation to rail operating companies. Road operators do not meet the full costs of maintaining the national network while rail has to pay track access fees. The imposition of further overheads to support enhanced safety regulation, including the national incident reporting systems mentioned in previous sections, exacerbates this perceived imbalance [858].

The funding of incident and accident reporting on Australian railways is more complex than the previous paragraph suggests. Each operator pays a safety accreditation fee that varies between the States even though there are mutual recognition agreements. It is, therefore, possible for an operator to seek accreditation in a State that is different from the one in which they conduct the majority of their business. The inconsistencies in funding also reflect deeper variations in the safety regulations that are enforced in different areas of Australia. Previous sections have described how these include the regulations covering the reporting of adverse events. It can, therefore, be argued that funding mechanisms might provide useful metrics to help monitor incident reporting systems. A recent series of reports urged that '... the Commonwealth takes a strategic approach to provide consistency in rail safety standards and practices for the national track' [858]. It was also recommended that 'a single annual fee for accreditation should be payable only in the jurisdiction of principal activity'.

Both license-based and taxation-based models of funding create financial pressures during a 'down-turn' in the economy. Transportation companies are, typically, faced with falling revenues. They must, however, continue to meet the licensing costs that are necessary to support incident reporting systems and the wider regulatory framework. In this model, financial burdens remain on the operators. In contrast, under a taxation based scheme, both the regulator and the operator are hit hard by falling sales. This creates particular problems for incident investigators. It can take well over a year to train an analyst [198]. Skilled and experienced staff cannot easily be dismissed in response to short-term market fluctuations.

There are alternative funding mechanisms. As mentioned previously, both the UK HMRI and the Health and Safety Executive impose specific charges for the work that they conduct. Time is invoiced for each quarter or half an hour spent on an investigation. This leads to invoices that contain several thousand entries and fee recovery takes between 3 and 4% of HSE/HMRI resources [467]. It can be argued that this represent an inefficient use of scarce resources. In particular, if this model were used on a subsidised national railway then public money would simply be transferred from rail operations to rail regulation. UK railways, therefore, operate a levy scheme to support the

Office of the Rail Regulator. This overseas the economic aspects of market intervention. A similar scheme is proposed for safety regulation. This has the strong advantages of simplicity and economy in contrast to other forms of funding [467]. As we have seen, however, it raises important questions about the scale of the levy to be placed on each individual operator.

It is clearly important that those who pay for reporting systems should realise benefits that are in proportion to their investment. This creates potential conflict because there is no direct relationship between funding and control in the area of safety regulation. For instance, the International Civil Aviation Organisation (ICAO) require that the investigation of accidents and serious incidents is conducted by an independent organisation. This principle is also reflected in recent European directives on the regulation of the aviation industry (such as 94/56/EC). This distinction is not embodied within the UK rail industry where Railway Safety is a not-for-profit, wholly-owned subsidiary of the infrastructure company, Railtrack Group PLC. Railway Safety does, however, operate under a separate management structure from its sister company, Railtrack PLC, which is responsible for operating the infrastructure under the Railtrack Group. It is unclear whether such a situation could continue if the European Commission implements the proposed extension of independent requirements to other modes of transport [467].

It can be argued that a 'healthy' incident reporting system should have the same financial and operational independence as investigatory organisations within civil aviation. The rules that separate accident investigation bodies from other regulatory or commercial organisations do not extend to incident reporting systems. Most are financially dependent on the agencies that implement their recommendations. For instance, the CIRAS system was initially funded by the rail companies that operated in the region that it covered. Such close relationships can create concerns; investigators may be reluctant to propose recommendations that are unpopular with financial contributors. In consequence, a National Steering Group was established to oversee the national CIRAS system. The members of the steering group include individuals from Railtrack Safety and Standards, Railtrack Line, Railway trade unions, the Association of Train Operating and Freight Operating Companies, the Infrastructure Safety Liaison Group and an independent human factors specialist [197]. A Charitable Trust has also been created to 'promote and protect the independence and integrity' of the CIRAS system. Again the members of this trust include a representative from Railtrack Safety and Standards, Railway trade unions, a human factors academic, a member of the Rail Passenger Council, a representative of the core facility service provider, and representatives of rail employers.

By monitoring the level of funding that is made available to a reporting system, it is possible to assess the investment that companies are willing to make in these schemes. As we have seen, however, economic trends can reduce the financial support that is made available to a reporting system. The previous section has also argued that additional managerial devices must be used to ensure the independence of many schemes, especially if they receive high levels of financial support from regulatory and commercial organisations. A number of further problems complicate the use of financial metrics to assess the health of a reporting system. Incident and accident investigation require specialist skills. It can be difficult to recruit and retain necessary staff. One recent survey argued that there were no independent rail incident investigators anywhere in the UK; 'consultants who do not work for Railtrack do no exist' [467]. The lack of independent investigators is compounded by structural and organisational problems that act as barriers to recruitment even when funding exists. For example, railways are often perceived to lack the 'glamour' of other high-technology industries. This creates problems in recruiting the best graduate, technical skills. The difficulties of staff recruitment and retention are compounded by the government Civil Service pay structures that operate within the UK HMRI. When there is competition for scarce talents 'the HMRI has been limited in what it could do by a lack of good people to take work forwards' [467]. It is important to stress that these recruitment problems also affect investigatory agencies across a broad spectrum of industries, including mining, nuclear and off-shore oil production, and in many different countries not just the UK railways.

15.4 Monitoring Techniques

The previous pages in this chapter have introduced broad distinctions between the different techniques that might be used to monitor the success or failure of an incident reporting system. Particular attention has been paid to the problems of interpreting the information that is provided by many of these monitoring techniques. For instance, an increase in the financial resources that are allocated to a reporting system may not be sufficient to attract skilled personnel. Conversely, a fall in regulatory contributions can increase the independence of some reporting schemes [467]. In contrast, the remainder of the chapter focuses in more detail on a subset of these monitoring techniques. Brevity prevents a complete exposition, however, the intention is to summarise the issues that must be considered before investing in a particular approach to the validation of a reporting system.

It is important to emphasise that the particular techniques used to audit a reporting system will depend upon the scale of the scheme and the organisation that it is intended to support. This point is reiterate by Transport Canada's guidelines for the development of railway Safety Management Systems [781]. They argue that monitoring and audit frequencies should depend on the size of the railway, the risks involved in their operations and the previous safety performance of the organisation.

> "Larger railway companies will likely have the staff and expertise necessary to establish auditing processes and teams, although they may choose to hire external resources to obtain specific skills or assistance. Smaller companies that may not have the resources to conduct an audit program internally may be able to obtain assistance from a variety of sources, including senior railways with which they interchange, consultants and professional auditors." [781]

Some authors have argued for the continuous monitoring of the performance of incident reporting systems, for instance using the direct measures introduced in previous sections [444]. For small scale systems, this can divert critical resources away from the analysis of adverse events. It may, therefore, only be possible to conduct periodic monitoring every six or twelve months [119]. Fortunately, a range of computer-based monitoring systems can be used to reduce the costs and hence increase the frequency of monitoring activities. The costs associated with some monitoring techniques, such as observational analyses, can dissuade safety managers from exploiting these techniques even on larger-scale schemes. The following sections, therefore, use previous applications of these techniques to provide an impression of their relative costs and benefits for the monitoring of reporting systems.

15.4.1 Public Hearings, Focus Groups, Working Parties and Standing Committees

Many different types of meeting can be called to help monitor an incident reporting system. Most of these hearings are called in the aftermath of particular failures. They, therefore, typically considered reporting systems within the context of a wider safety management system. It is rare for public hearings, focus and working groups or standing committees to concentrate exclusively on the utility of a particular scheme. This broader focus does not, however, prevent these meetings from providing important insights about the performance of a reporting system. For example, many focus groups begin by looking at the perceived causes of a particular incident and then go on to question the reasons why lessons had not been learned from previous, similar incidents. The following sections, therefore, briefly describe the ways in which these different venues can be used to provide feedback about reporting systems.

Public hearings provide a means of assessing general attitudes towards incident and accident reporting systems. These meetings are often called to review general safety concerns in the aftermath of major failures. For instance, the FRA held a series of public hearings following a number of incidents in which passengers had been unable to escape from trains in the aftermath of a derailment or collision [236]. The catalyst for these meetings was the Silver Springs incident described above; a Maryland commuter train ignored a signal and collided with an Amtrak passenger train. Such public meetings pose a considerable challenge to those who must both organise and chair them. There is a danger that pressure groups will attempt to promote their views and exclude those of other groups with valid concerns. Equally, however, it is important that the convenor of a meeting should not be seen to stifle debate by imposing a rigid control over the proceedings. The FRA have well-rehearsed mechanisms for addressing these potential problems. The dates of a proposed public meeting are published in the Federal Register. Members of the public must then notify a clerk of their intention to speak. They must also submit three copies of their planned oral statement by a date that is specified in the call for participation. Members of the public are notified that their submission has been received by the FRA. Their written submissions are then made available for examination by other potential participants and by representatives of 'interested parties' prior to the meeting. This procedure has several merits. Firstly, it alerts the meeting chair to potential conflict. Secondly, it helps to ensure that any questions of fact can be raised and resolved before the meeting so that any subsequent debate can be based on reliable information.

Public meetings are often held to identify concerns that have not been addressed by working groups, focus groups and standing committees. For example, the FRA's public hearings were called in response to an interim report that was published by a working group on Passenger Train Emergency Preparedness. It is difficult to establish clear distinctions between these other forms of meeting. The terms 'working group', 'focus group' and 'standing committee' are often used synonymously by both regulators and operating companies. In general terms, however, a focus group can be thought of as an informal meeting that is held to consider a particular series of issues. The meeting need not arrive at a particular plan of action but may produce broad recommendations about the items being discussed. In contrast, a working group can be thought of as a more formal device to both consider particular issues and then act to resolve them. The life time of the working group usually ends with the successful resolution of the items being considered or by the implementation of their recommendations. A standing committee, typically, has greater longevity. They are often intended to provide a continuing point of reference for the consideration of long-standing issues. All three of these devices can and have been used to monitor the success or failure of incident reporting systems.

As mentioned, public meetings often attract participants that have a particular perspective of, or vested interest in, the issues that are being discussed. For example, passenger groups, environmental protection organisations, the proponents of road transport have all actively participated in recent public meetings on rail safety [338]. Such organisations are well placed to represent particular views within the wider community. They may not, however, reflect the diversity of attitudes held by the general public. In consequence, many organisations rely upon focus groups to investigate perceptions about the safety performance of particular industries. These meetings have the benefit that participants can be selected to deliberately reflect a broad cross-section of views. For example, the FRA used focus groups to assess compliance with railway operating rules. The intention was to assess whether corporate culture had an influence on potential violations [247]. This study illustrates how focus groups play a particularly important role in analysing the causes of common failures. As we have seen, incident reports can often provide information about what happened. It is far more difficult to understand why particular patterns of failure occur across an industry. The FRA in using this technique have sought to provide additional analytical information than that which is normally provided through their mandatory reports scheme.

Focus groups can be used to directly assess particular attitudes towards the operation of an incident reporting system. For instance, the US Bureau of Transportation Statistics undertook a series of workshops to identify 'stakeholder' concerns about the reliability and accuracy of accident and incident information [116]. Their concerns should not be surprising, they reiterate concerns that have been raised throughout this book. The participants drew attention to data quality. They were concerned about both the under-reporting and the over-reporting of particular types of adverse events. They were worried by the lack of uniformity in completing reports. They voiced concerns over exclusions that removed reporting requirements from some transportation workers. Typical comments include 'there needs to be better information and it needs to be of a higher quality', 'there needs to be better data on results', 'accuracy is a challenge because of budgetary problems and different interests' and 'it is difficult to get accurate, undiluted information on human error and performance' [116]. The focus groups also revealed concerns over the relevancy of data produced by the Bureau of Transportation Statistics. Industry participants were concerned to ensure that the right information was being collected and that data that was duplicative or no longer useful was not collected. This final observation is highly instructive. Focus groups are one of the few mechanisms that can be used to obtain feedback about the overheads that imposed on potential contributors by reporting requirements. Many of the other measures, such as submission rates or intervention metrics, take little account of the costs that a system might impose upon potential contributors.

Focus groups are more commonly used to discuss concerns that arise in the aftermath of high-profile accidents and incidents. Many of these concerns centre on the failure of reporting systems to prevent the occurrence, or mitigate the consequences, of the adverse event. This can be illustrated by a recent seminar held in the aftermath of the Ladbroke Grove accident. A focus group explicitly considered the role of incident reporting as part of a wider review of employee attitudes to rail safety [466]. This seminar included present and former railway staff, signalers, Control Room Operators, incident and accident investigators and project managers. All participants appeared in a personal capacity, however, and were not intended to ast on behalf of any particular organisations. A list of the questions were circulated to the participants before the meeting. They were asked to send in brief comments that were then circulated to the other members of the focus group before the meeting.

1. How concerned about safety are those who work on the railways?
2. What are the main concerns with respect to safety on the railways?
3. How important do those who work on the railways consider safety to be, relative to other issues such as punctuality and reliability of train services?
4. In practice, are safety requirements compromised by commercial considerations?
5. Has the fragmentation of the rail industry had an adverse effect on safety? If so, in what respects and for what reasons?
6. Is there uncertainty or confusion as to who is responsible for what with respect to safety on the railways?
7. On a personal level, are those who work on the railways aware of their duties and responsibilities with respect to health and safety issues?
8. Are unsafe acts and conditions tolerated on the railways? If so, do they go unreported? How can this problem, if it exists, be addressed?
9. Is there a mechanism whereby those who work on the railways can express safety concerns to those in positions of authority within their organisation? Is the mechanism effective? Are their concerns addressed and acted upon?
10. Is the confidential incident-reporting system on the railways used? Is it trusted and respected by the workforce? Is it effective?
11. How are safety issues communicated from directors and other policy-makers to the workforce? Are safety issues given enough emphasis? Are there sufficient safety-related incentives?
12. How often do those who work on the railways receive visits and/or safety briefings from supervisors and senior managers? Are the briefings effective?
13. How often do those who work on the railways receive formal training on safety-related issues? Is the training effective?
14. Does the reliance on contractors and sub-contractors for track repair and maintenance prejudice safety?
15. What is the delegates' understanding of the safety case regime? Have those who work on the railways seen their company's railway safety case or 'assurance case'? Is this a document they use or on which they rely? How does the document relate in practice to more prescriptive requirements such as the Rule Book?

16. What should be the appropriate balance between the use of broad objectives on the one hand, and detailed prescriptive rules on the other, to achieve safety on the railways?

17. What can be done, or should be done, to improve safety on the railways?

This example illustrates several important features about the use of focus groups to monitor incident reporting systems. The seminar was was not intended to reach particular conclusions on any of the questions. The intention was review employee perspectives on safety in the rail industry. This reveals an irony in the use of the term 'focus group'. These meetings frequently move from a focussed set of issues to more general and wide ranging discussions. It is, therefore, important that the facilitator or organiser retained control over any meeting without dictating the content of the discussion. Our case study meeting initially focussed on what the employees' main safety concerns, including issues of leadership, responsibility and accountability. Only then did the focus group concentrate on communication mechanisms, including incident-reporting. The meeting also focussed on many other issues ranging from the employment of contractors to training and the use of UK railway's rule book [466].

As mentioned, focus groups are often used to provide general feedback about attitudes towards a reporting system. For example, the UK meeting described an initial scepticism about whether CIRAS 'provided a genuinely confidential reporting scheme' [466]. The members of the focus group were found that 'experience of 34 months working the system showed that it was excellent, a lot had been learned and that there was no breach of confidentiality'. In contrast to focus groups, working parties are typically expected to provide detailed recommendations. For instance, the Health and Safety Executive recently established a working group to 'deal with the problem' of vehicles crashing onto railway lines from overhead bridges [113]. This group collated evidence and analysis from a large number of incident and accident investigations. Their analysis of this collated evidence recommended more barriers, improving the road layout and introducing better signs for drivers.

Working groups often coordinate their activities with those of other, broader forms of consultation. For example, previous paragraphs mentioned the public meetings that were called following reports from the FRA's Passenger Train Emergency Preparedness Working Group. It was argued that the FRA must become more proactive in order to minimize the consequences of future accidents; 'even minor incidents could easily develop into life-threatening events if they are not addressed in a timely and effective manner' [236]. The establishment of this working group might be seen as an implied criticism of existing incident and accident reporting systems. In this case, accidents such as the Silver Springs collision have demonstrated that more action needs to be taken to mitigate the consequences of any future failures. In this view, the working group is established to supplement systems that have failed to adequately address existing safety problems. Equally, however, it can be argued that the establishment of the working group illustrates the success of existing reporting systems. The need to consider emergency preparations has been established from the analysis of previous incidents.

The participants in a working group are typically chosen to ensure that a broad range of interests are represented. They, therefore, play an important role in assessing the feasibility of the recommendations that are produced from a reporting system. The expertise and experience of the participants can often help to identify implementation concerns that were not initial recognised by incident and accident investigators. For example, one recommendation from previous collisions and derailments was that the FRA should require the introduction of booklets and videotapes to illustrate equipment and describe entry and evacuation procedures on trains. The Working Group pointed out that 'that pilferage of on-board emergency equipment is a serious problem on many passenger railroads, and that specifically focusing the attention of passengers on where the equipment is located would only exacerbate the problem' [236]. They also argued that frequent travellers probably already knew where the equipment was located and would not, therefore, benefit from such additional information. This case study provides further examples of the way in which a Working Group can provide feedback on the recommendations derived from previous incident reports. For example, Amtrak used the meetings to point out the difficulties of introducing emergency preparedness booklets and videos across its network. Not only would they have to distribute this information on many thousands of rail services, they would also have to send them to emergency responders throughout the

United States. Subsequent mailings would also have to be used to ensure that any information was up to date. The FRA considered these comments to the Working Group and invited commentators to 'suggest either how Amtrak can best comply with the emergency responder liaison requirement as set forth in the proposed rule, or whether the final rule should establish a different standard for railroads that operate in territories with large numbers of potential emergency responders to contact' [236].

The FRA working group illustrates the use of such meetings to assess the recommendations that have been produced in response to previous incident reports. It illustrates the use of these meetings to look at particular safety issues, in this case emergency preparation on passenger trains. Similar techniques have been used at a higher level to review incident reporting systems within the wider context of a national regulatory framework. For example, the Ladbroke Grove and Southall rail accidents led to an industry-wide review of safety management of UK railways. A working group was established under the Department of Transport, Local Government and the Regions. This conducted a review of the Safety and Standards Directorate within the infrastructure company, Railtrack. As with the FRA case study, the findings of the working group were informed by and helped to inform public inquiries. In this case, the Department of Transport working group implemented a number of significant changes pending the recommendations of the Public Inquiry into the Ladbroke Grove accident. The scale of these changes cannot be underestimated. The working group initiated the transfer of responsibility for determining whether or not another train company was safe to operate from Railtrack to the Health and Safety Executive [689]. Railtrack's Safety and Standards Directorate were transformed into a separate, non-profit making company with an independent board of directors within the Railtrack Group. The objectives of this new organisation were to provide 'safety leadership' to the industry and take a more dynamic approach to setting and updating standards. More significantly given the focus of this book, the new Railway Safety body was to 'establish a more effective regime for safety audit, incident investigation and ensuring that corrective action from audit and investigation is taken' [689]. Previous paragraphs have described the Precursor Indicator Model (PIM) that has been developed by Railway Safety to support this more pro-active approach to safety audits. The Working Group's general review of rail safety, therefore, triggered changes that 'revolutionised' both the operation and monitoring of mandatory incident reporting across the UK rail network [692].

Regulatory and governmental agencies are responsible for commissioning most working groups. Professional organisations, industrial bodies and pressure groups have also starting investigations into the success or failure of incident reporting systems. For instance, the UK Royal Aeronautical Society's Human Factors group has established a Rail and Aviation Working group [377]. This aims to share human factors expertise, resources and best practice from the aviation communittee with representatives of the rail industry. This Working Group was explicitly established with the Royal Aeronautical Society because as 'an impartial professional charity' it can provide the intellectual resources and unbiased refereeing that may not be available from other similar bodies. Representatives are drawn from Railtrack, the Rail Industry Training Council, the Aviation Training Association, British Airways, the UK Flight Safety Committee and individual train operators amongst others. This Working Group has focussed on transferring lessons from the operation of aviation reporting schemes, in particular British Airways's BASISindexBASIS [660], into the emerging national rail systems. The objectives and even the existence of such a group provides important insights into the perceived health of existing reporting systems within the UK rail industry. The perceived need to transfer skills and techniques from the aviation domain into the railway industry implies a relatively low regard for existing rail systems. The working group description concedes that the aviation safety record is not perfect. However, there is no recognition that techniques might be propagated back from the railway domain to support aerospace safety management.

Public hearings, focus groups and working groups all tend to have a limited duration. Public hearings and focus groups are called to identify particular concerns on topical issues. Frequently, they are used to gather feedback about the management of safety in an industry following high-profile failures. For instance, they may provide insights into the reasons why reporting systems fail to prevent an adverse event. Working groups are similar in that their longevity is bounded by the publication and implementation of recommendations. In contrast, standing committees provide a common point of reference for long-standing concerns. They can be used to coordinate the work of focus groups, of public hearings and of working groups. For instance, the Royal Aeronautical Society's Human Factor's group has standing committees on crew resource management, on maintenance engineering and on air traffic management amongst other topics. These groups are intended to monitor developments, set up 'focus teams' and advise the main committee on specific Human Factors issues. The Human Factors group can itself also be seen as a standing committee; it coordinates the Rail and Aviation Working group mentioned in the previous paragraph.

The UK Railway Industry Advisory Committee provides an example of a government sponsored standing committee. It was established by the Health and Safety Commission in 1978. The Railway Industry Advisory Committee 'plays an important role in providing a consultative forum where all interests within the railway industry can meet and reach consensus on how to progress health and safety proposals and other related developments within the industry' [338]. Meetings are chaired by the Chief Inspector of Railways. Seven employers' representatives and 'balanced' by a similar number of employees' representatives who are nominated by railway trade unions. Passengers and the general public are represented by two members from the Rail Passengers Council. The membership of the committee has been reviewed and revised several times to reflect the changing structure of the industry since privatisation. This process is an important difference with the other feedback meetings mentioned in this section. The limited longevity of working groups, focus groups and public hearings makes membership changes less important than they are for standing committees. As mentioned, standing committees often coordinate the work of these other groups. The Railway Industry Advisory Committee supports a Freight Sub Group; an Occupational Health Working Group; a Prevention of Trespass and Vandalism Working Group; a Research Working Group and a Human Factors Working Group. Each of its working groups have terms of reference and plans of work that are approved by the main Committee and their Chairs report to the main RIAC Committee. Each of these groups draws upon incident and accident reports as an important means of informing their activities. For example, the Occupational Health Working Group has used analyses of previous injuries to draft of industry-specific guidance on manual handling for employers and employees on the railways. The Research Working Group has started two initiatives on track-side safety and on the effects of safety messages on influencing the behaviour of railway passengers. Each of these activities was motivated, in part, by their interpretation of recent safety statistics derived from the various industry reporting systems. The members of the Human Factors working group helped to promote the CIRAS scheme as a means of identifying the safety concerns of operators [320]. The Railway Industry Advisory Committee working groups also helped to monitor the use of the RAVERS fault tracking and reporting system following the Southall accident. This was seen as a short term solution to a situation in which most operating companies had computerised facilities to log faults and produce trend reports but 'the ability to share data nationally is being compromised by industry moves to 'stand alone' systems' [320]. It is important to note that the Railway Industry Advisory Committee blurs some of the distinctions that were introduced in previous sections. There is no suggestion that the working groups will be suspended once the human factors or maintenance issues have been satisfactorily 'resolved'! It might, therefore, be better to refer to these groups as sub-committees that will continue to support the work of the standing committee. The key point is, however, that these bodies provide many different industry stakeholders with the ability to address particular issues over a prolonged period of time. They are not simply established to address the findings of a particular investigation. It is also important to note that incident reporting systems provide a vital information resource to the members of these committees. It should not be surprising, therefore, that the Railway Industry Advisory Committee's working groups have addressed the development of various fault monitoring systems and confidential reporting schemes.

Public hearings, focus groups, standing committees and working groups provide valuable information about attitudes and opinions about particular reporting systems. They, therefore, often provide post hoc information in the sense that opinions are often formed in the aftermath of adverse events. They identify concerns without necessarily offering clear guidance about constructive solutions. There are, of course, exceptions to these generalisations. It is important, however, that safety managers and regulators have access to alternative techniques that can be used to assess the utility of particular reporting systems. Incident sampling techniques address this requirement; representative subsets of previous failures can be examined to determine whether a range of alternative techniques might have yielded further insights into the causes of adverse events.

15.4.2 Incident Sampling

The term 'incident sampling' covers a number of different techniques that extract a sample of reports that have been submitted about adverse events. These techniques differ in the criteria that are used to choose a particular subset of events. For instance, analysts may attempt to extract a random sample. Alternatively, they may base their selection on incidents from a particular functional subsystem or geographical area. Incident sampling can also be focussed on particular levels of severity. For example, monitoring activities may concentrate on those failures that had the greatest potential adverse consequences. Once the subset has been identified, each incident is analysed to assess the quality of the causal analysis, to validate any potential recommendations and so on.

A recent 'Assessment of Investigations into Signals Passed at Danger' on UK railways illustrates this approach [745]. This investigation was conducted by W.S. Atkins in order to monitor the efficacy of revised investigation regulations following the Ladbroke Gove accident. These revisions required that HMRI inspectors investigate each major SPAD incident in addition to any enquiry conducted by the railway companies involved. The Atkins report was partly intended to compare the results of the HMRI investigations with those of the rail operators. They were also intended to interpret their findings in the light of the HMRI's own separate analysis of the railway companys' investigation techniques [351]. The Atkins report examined the conclusions reached from each of these different enquiries in order to identify the 'value added' by having HMRI inspectors perform their own independent analysis of each high-severity SPAD in addition to company investigations. Of the 146 SPADS investigated by both the HMRI and the companies, 13 were selected for further analysis by the Atkins report. It is difficult to identify the precise criteria that were used to inform the selection of this subset. However, a six stage methodology was used to guide the monitoring process.

1. *Data collection.*
   The project began by reviewing the 146 incidents investigated by both the HMRI and industry investigators since October 1999. The results were collated to compare the root causes identified by one or the other or both investigations. The analysis also attempted to identify root causes that might have been overlooked in both previous investigations.

2. *Review company investigations.*
   The analysis then used the collated data to identify causal patterns that might not have been identified by the previous investigations. The Atkins report attempted 'within the limits of the relatively small sample' to identify differences in the effectiveness of investigations between different regional zones. This review also provided insights about the consistency and thoroughness of company investigations compared to those of the HMRI.

3. *Review HMRI investigations.*
   The HMRI reports were also critically reviewed to identify the relative strengths and weaknesses of their analysis. As mentioned, the intention was to identify ways in which the HMRI investigations might 'add value' to company reporting procedures.

4. *Independent analysis.*
   A small sample of 13 SPADs were then analysed in greater detail to identify any root causes not identified by either the HMRI or the company reports. This was intended to determine whether these investigations considered 'the full list of possible causes' and then proposed 'suitable measures' to prevent any recurrence [745]. This stage of the analysis was intended to uncover any correlation between the root causes that were missed and the type or category of SPAD's being investigated. It was also hoped that this analysis might improve our understanding of the reasons why any root causes might have been overlooked.

5. *Collation of root cause data*
   The results of each of the previous stages were collated to summarise all of the root causes identified since October 1999. This data then informed a series of more detailed statistical analyses to identify trends and patters across the SPAD incidents.

6. *Proposals and recommendations*
   The final report identified the strengths, weaknesses, trends and Zone differences observed in both investigation processes. Proposals were also made for ways in which the benefits of the subsequent HMRI investigation might be achieved without the need for a second investigation. Finally, the Atkins report proposed selection criteria that can be used by the HMRI to 'confirm that the Industry investigations are achieving consistency in depth of analysis and conclusions reached' [745].

The results of this process showed considerable agreement between the HMRI and the company investigations. However, the report's writers stress that their comparison focusses on 'the quality of the investigations rather than the results and recommendations' [745]. This is significant because it might be argued that the report, therefore, overlooks the benefits of more direct intervention following the HMRI report. The comparison did, however, illustrate some of the problems that arise when comparing different reporting systems. For example, the independent review concluded that although there were differences in the root causes found by the HMRI and the company enquiries, both were 'equally valid'

The third stage of the methodology, described above, was based around a subjective comparison of the quality of the investigations conducted by the HMRI and the railway companys. The score 3 represented a 'robust report', 2 was assigned for a 'good report' and 1 was used if the report was poor and had 'significant shortcomings in either analysis or conclusions' [745]. Of 228 investigations, only 10 (4.4%) were subjectively classified as 'poor'. No incidents were received a score of 1 for both investigations. In 93 SPADS, the HMRI score was equal or worse than the company score. In 21 (18.4%) incidents the HMRI were assessed as 'adding value' to the industry investigation. The analysis also identified weaknesses in both the HMRI and the company investigations. These can be summarised as follows:

- *Incident Selection.*
  The criteria that were used to select the incidents that were to be subject to both company and HMRI investigations was skewed towards shunting incidents. These were argued to be of relatively 'low consequence' and their over-representation was perceived to indicate an imbalance in the SPAD severity classification scheme that informed the selection process.

- *Terminology.*
  Company and HMRI investigations used different terminology. For example, HMRI and Railway Safety reports used SASSPAD for 'Starting Against Signal SPAD' while in other contexts the same acronym was taken to mean 'Start Away From Station SPAD'.

- *Special Exercises.*
  Special exercises, for example top gather information about hand-signalers, increase awareness of particular safety issues and help to elicit reports about certain classes of adverse events. Atkins' review argued that 'care must be taken' to ensure these initiatives do not compromise 'mainstream' SPAD investigations.

- *National Issues.*
  Although the HMRI had identified issues across many different regional operating zones, some issues had 'slipped through the net in spite of clear evidence of a trend being available' [745]. These included situations where the driver had been forced to apply power even though the signal was at red.

- *Balance between Human Factors and Infrastructure Issues.*
  It was argued that in both sets of investigations, there was a tendency to see human 'error' as

a cause without exploring further into the infrastructure issues that may have made the error more likely.

- *Organisational Issues.*
  The review concluded that both the HMRI and the company reports tended to under-emphasise organisation issues unless they were of an extremely serious nature. They found that 'it is only on rare occasions that we have seen focussed recommendations addressing supervisory or management weakness' [745].

- *Follow-up of Issues Raised.*
  Atkins expressed a concern that in some cases the HMRI had identified significant problems without initiating follow-up actions. They admitted that this concern was not, however, supported by direct evidence.

- *Miscommunication.*
  The review found that incidents involving communication failures were typically blamed on the driver even though other personnel, including signalers, may have contributed to the adverse event.

- *Aspect Sequence.*
  Both the company and the HMRI reports failed to give sufficient consideration to the aspects of the signals that the drivers had encountered immediately before the SPAD incident. In some industry reports 'the cautionary aspects which might have had a significant bearing on the incident have been completely ignored' [745].

- *Technical Complexity.*
  The HMRI occasionally misunderstood or oversimplified the issues involved in complex investigations. In other cases, they used inappropriate language or obfuscated the issues in a way that created an unhelpful image of the Inspectorate.

- *Compliance and Effectiveness.*
  The report argued that the HMRI should be more forceful in challenging previous industry practices especially when it was obvious that remedial action might be costly. They argued that 'lack of adverse comment in certain cases might be construed as acquiescence' [745].

This list illustrates the range and number of insights that can be drawn from relatively focussed monitoring techniques. It should be emphasised, however, that these observations were based on the subjective analysis of the Atkins staff. Similarly, their work was not supported by a formal methodology that might have supported the monitoring of other reporting systems. Such caveats need not, however, undermine the importance of their findings. For example, they were able to identify fundamental differences between the ways in which different company's interpreted causal information. East Anglia operated two different classification systems. In some reports, immediate causes were distinguished from underlying causes. In other reports, a three-tier hierarchy distinguished immediate causes from basic causes and root causes. Great Western A two-level hierarchy: Immediate Cause; Underlying Cause London North Eastern either operated a three-level hierarchy involving immediate, basic and root causes or a complex two-level hierarchy. In this approach, immediate causes were distinguished from 'underlying causes/personal factors or underlying causes/job factors'. Southern region operated an alternate two-tier hierarchy that distinguished conclusions from underlying causes or a three-level hierarchy involving immediate, basic and root causes. This led the Atkins review to suggest a simplified structure distinguishing three different levels. The first level describes primary events and special circumstantial factors. At the second level, basic causes are identified. These include human factors and ergonomic issues. They also include infrastructure problems and procedures or instructions. The final level documents the underlying organisational, managerial and supervisory causes. Such proposals and the diverse approaches operated within each zone indicate the importance of monitoring the operation of similar incident reporting systems across national industries. Experience within the aviation industry has shown that these different classification schemes act as significant barriers to the exchange of important safety-related information [310].

The Atkins review of the SPAD investigation processes illustrates the monitoring of reporting at two levels. Firstly, the report is itself an attempt to monitor the integration of, and value added by, the duplicate company and HMRI investigations. Secondly, the HMRI is itself an indirect means of monitoring the company investigation procedures. Hence there is a sense in which the Atkins report monitors the monitoring system. This can be seen in the review of the criteria that the HMRI used to determine which SPADS to investigate. The Atkins report chose the sample of 228 incidents because these adverse events were the ones where the HMRI had already chosen to 'duplicate' the company investigation. The report raised a number of concerns about the criteria that the HMRI used to select their sample for further investigation. Atkins described some of the incidents that the HMRI monitored as 'low risk, low value'. These included SPADs in which the trains may have travelled relatively large distances beyond the red signal but at relatively low speed and with little risk, for instance within a depot. In many of these cases, the HMRI were able to add little value to the company reports.

The report into the HMRI monitoring also raised more general questions about the limitations of incident sampling techniques. Investigating a sample of all SPADs incidents 'inevitably casts doubt on any statistical data' and especially 'the extraction of cause data'. The Atkins investigators were concerned that a widespread causal factor might go unnoticed if the resulting SPADs do not meet the selection criteria for the HMRI monitoring. They cite the example of several incidents in which the SPADs follow shortly after a train starts on a yellow signal. This type of incident is likely to result in a short distance SPAD as the signal changes to red. Hence, they will not be investigated by the HMRI. The Atkins team suspect that these incidents may be much more prevalent than either HMRI or the industry currently believe. Thee is a certain irony in the sampling criticisms made in this review. Many of these adverse comments might also be levied at this meta-level review because part of their analysis depends on a subset of the HMRI sample. Leaving aside this caveat, the Atkins report goes on to identify a revised set of criteria that might be used to guide the HMRI's decision to launch an investigations alongside a company enquiry. For instance, the HMRI should investigate an overrun which results in injuries or fatalities to either passengers or staff. They should also investigate an overrun which results in damage to the infrastructure, or damage to traction units or rolling stock. These detailed criteria are extended to include broader categories such as incidents that meet criteria defined in occasional special studies and 'a random selection of SPADs'. These requirements are intended to address the problems that can arise when particular definitions inadvertently exclude incidents involving certain causal factors, such as those described above.

In addition to suggesting alternative conditions that might be used to guide the HMRI sampling of SPAD incidents, the Atkins report also describes clear objectives for any future monitoring by the HMRI. The main aim of this activity should not be to 'duplicate' industry investigation. Instead, the HMRI should establish what happened, identify the implications of what happened and determine the effectiveness of any proposed recommendations. The HMRI need to know what happened in order to evaluate the risks of any repetition. They need to know the the implications of what happened in order to assess the potential consequences of any recurrence. They need to assess the effectiveness of any recommendations to ensure that industry proposals will address all the 'components' of the SPAD. In particular, the Atkins report identifies situations in which disciplinary action has been taken against the driver while infrastructure weaknesses were overlooked. They conclude their review of the existing monitoring functions by observing that:

> "Our investigations have revealed a wide variation in the quality of root cause analysis being undertaken. Used properly and intelligently it is a powerful tool for extracting all the implications from an incident. Used mechanistically, and we have seen a number of examples of this, it can lead to some root causes which, whilst they may expose valid weaknesses, bear little relevance to the SPAD and, when corrected, will have negligible effect in preventing a recurrence. Too often these irrelevant issues are being pursued at the expense of more serious ones which are being ignored." [745]

As mentioned, a range of different criteria can be used to identify the sample incidents that can be used to monitor many reporting systems. The Atkins study drew on the sample of SPAD incidents that were investigated by both the HMRI and by company investigators. The HMRI, in turn, used a number of complex criteria to determine which of the company investigations they would also look into. For example, one aspect of the criteria focussed on the length of the overshoot that resulted from the SPAD. The Atkins report was not the only example of meta-level monitoring to be triggered by the Ladbroke Grove accident. This also influenced the HSE to serve two 'Improvement Notices' on Railtrack, the infrastructure provider. These required that they produce a plan, with fixed implementation dates, to reduce the risk of a future SPAD at the 22 signals with the worst safety record across the network.

The signals were chosen because they have a record of multiple SPAD incidents. SPADS are divided into four categories. Category A incidents occur when a train passes a signal at danger without authority, other than those SPADs defined in Categories B, C and D. Category B SPADs occur when a train passes a signal at danger without authority because a stop aspect or indication was not displayed with sufficient time for the driver to stop safely at the signal. Category C SPADs are occasions where a train passes a signal at danger without authority because a stop aspect or indication was not displayed with sufficient time for the driver to stop safely at the signal, because it was returned to danger in an emergency in compliance with rules and regulations. Category D SPADs are those occasions when vehicles without any traction unit attached, or any train which is unattended runs away past a signal at danger [358]. The 22 signals involved in the review were the site of Category A SPADs. These incidents are also classified according to eight severity ratings. A level one SPAD is the least severe with an overrun of no more than 25 yards and no damage or casualties. Severity rating eight is the worst and is used for incidents resulting in fatalities. The review of Railtrack's actions was concerned with 'the identification, understanding and mitigation of the causation factors increasing the likelihood of signals being passed at danger' [358]. The consequences of SPADs at each signal was, therefore, of secondary importance. This contrasts with the previous Atkins report that focusses more on the high-consequence incident that were investigated both by the HMRI and by individual companys.

The HSE enforcement notices were based partly on a statistical analysis of previous SPAD incidents and partly also on a recognition that previous recommendations had failed to prevent the recurrence of adverse events. The statistical analysis showed that for signals that had been passed at danger three or more times, there was only a 4% probability that the SPADs are entirely due to random causes [358]. Where there were four or more SPADs, the probability becomes close to zero. The results of this analysis triggered an enquiry to determine whether investigators had identified all of the infrastructure elements or environmental factors that make SPADs more likely to occur at these signals. The enquiry was jointly conducted by the HMRI and by W.S. Atkins. The focus of the investigation was on Railtrack's response to Improvement Notice I/RJS/991007/2 requiring action to mitigate the risk of signals being passed at danger at the 22 signals that had been passed at danger most often between 1990 and 1998. The previous incident reports for each of the signals were examined. Discussions were held with Railtrack Headquarters and with each of Railtrack's seven Zones. Meetings were also held with representatives of Train Operating Companies and some infrastructure maintenance contractors. All of the signals were viewed from a train cab and discussed with drivers. Reviews were also held with HMRI and Atkins representatives. The methodology used in this investigation is instructive because the report describes further constraints on their terms of reference that affected the manner in which they examined incidents involving the 22 signals. For instance, the HMRI was already conducting a separate inspection of the operating company's systems for driver management in order to assess whether they were adequately addressing the causes of human 'error'. The review, therefore, focussed more on the infrastructure issues that were under Railtrack's direct control than the driver-related factors that were largely the responsibility of the operating companies. A further two of the signals, SN63 and SN109, on the exit from Paddington Station were excluded from the review. They were close to the site of the Ladbroke Grove accident and hence were covered by a separate remedial plan. The meta-level review of Railtrack's actions on the 22 signals also revealed some of the dangers of monitoring incident reporting systems. For example, some of the signals were again passed 'at danger' after the investigation was completed and Railtrack's remedial actions had been approved by the HMRI. The report addressed potential criticism of their monitoring by arguing that 'further improvements may only come from improved management of driver competence and fitness, but Railtrack is also investigating whether there are further improvements which can be made to the infrastructure' [358].

The monitoring activity for each signal was intended to determine whether Railtrack had taken effective steps to understand the probable cause of previous SPAD incidents. The HMRI/Atkins review resulted in requests for Railtrack to carry out further risk assessments on some signals. Two of the signals required remedial actions that could not be completed before the review was published. It was concluded that the remaining signals had received adequate attention from Railtrack. Sufficient attempts had been made to identify likely causal factors and to apply appropriate measures to reduce risk. The report argued that 'in most cases it was also clearly evident that there had been close co-operation with the Train Operating Companies both in identifying and applying the risk mitigation and in ensuring that drivers were well briefed' [358]. Railtrack and the operating companies also revealed 'encouraging' evidence of an improved understanding about the factors that were likely to increase the risk of signals being passed at danger and of measures that might mitigate those factors. These more positive remarks were balanced against a small number of exceptions. At Manchester Piccadilly, a blanket speed restriction had been implemented on HMRI advice. This reduced the risk of future SPADs but the underlying causes of the original incidents had not been adequately identified. Other incidents had taken a significant amount of time to address so that effective remedial action were not implemented until long after the original incident had been reported.

This section has described how a range of different sampling techniques can be used to focus monitoring activities on particular types of incidents. Resources can be concentrated on incidents that receive particular attention within the reporting system, such as the SPADs that were investigated by both the HMRI and individual companys. Alternatively risk-based criteria might be used, including the severity assessments that informed the selection of the 22 worst SPAD signals. The Atkins review of the HMRI and company investigation processes also identified the limitations of sampling techniques. The selection of particular incidents can systematically exclude other incidents. If an investigation were extended to include these other adverse events then the monitoring might identify potential weaknesses in the underlying reporting syste, It is for this reason that the Atkins report was careful to propose detailed conditions that might trigger HMRI investigations in addition to the initial company investigation. A number of further limitations affect incident sampling techniques. In particular, there is little point identifying detailed sampling criteria if insufficient incidents are being submitted in the first place or if it is clear that systematic biases affect the reports that are being contributed about adverse events. These caveats are not a significant problem for SPADs where it is highly likely that any incident will be noticed by signalers and other railway staff. In other domain, however, these issues encourage regulators and safety managers to seek alternative means of monitoring their reporting systems.

15.4.3 Sentinel systems

Monitoring is not simply used to assess the health of a particular reporting system. It can also help managers assessing the potential strengths and weaknesses of alternative reporting techniques. For example, there must be some means for comparing the results of different causal analysis techniques or of different form designs. The importance of such comparisons is illustrated by the Safety Case requirements that govern UK railways [352]. These documents are intended to persuade regulatory authorities of the adequacy of an operator's safety management plan. The Safety Case must provide evidence to show that they will:

- " encourage staff and others to report accidents, incidents and other events that have or could have affected health and safety;
- ensure fair and equitable treatment of those whose actions are examined as a result of an investigation;
- investigate incidents and accidents (including potential or actual instances of ill health) to determine immediate and root causes;

- provide investigators with suitable competence, seniority and authority to undertake impartial effective investigations;
- provide resources to investigate, categorise and implement action to prevent repetition;
- match the investigation response to its potential severity;
- co-operate effectively with other duty holders to ensure that all lessons are shared, learned from and acted on;
- review findings from investigations (both internal and external) periodically to ensure that technical and managerial inadequacies are corrected;
- feed back to staff and others the results of investigations; and provide procedures for all aspects of an investigation." [352]

At present, most safety cases simply assert that particular steps will be taken in the event of an incident or accident. As we have seen, however, the HMRI, Lord Cullen and W.S. Atkins have criticised the adequacy of existing reporting systems. Monitoring techniques provide a means of addressing these criticisms by providing evidence that a proposed approaches provides greater benefits than the potential alternatives. It is important to stress, however, that such comparisons should not jeopardise the operation of an existing reporting system. There is a clear concern that any short-term trials might lose confidence in a successful scheme or lose data about failures that might later prove to be essential in preventing future accidents. Sentinel systems provide a means of obtaining pre-hoc information about a reporting system without forcing changes throughout large-scale reporting schemes.

Chapter 13.5 briefly described the main features of Sentinel reporting systems [264]. Rather than running a national or regional reporting system, Sentinel schemes elicit information from a small sample of 'representative' units. This approach has numerous benefits. For example, it avoids the costs with larger scale national systems. Sentinel systems also focus training and 'awareness raising' resources on the selected units so that participation rates can be raised above those normally associated with mass schemes. Mass reporting systems are often referred to as 'passive' because they do not actively encourage each member of staff to contribute to the system. In many cases, these higher levels of participation would overwhelm the analytical resources of the scheme. In contrast, smaller-scale Sentinel applications are referred to as 'active' because they seek to encourage closer participation in many different aspects of the reporting process. The smaller scale of Sentinel systems also enables additional resources to support the causal analysis and identification of recommendations from the smaller number of adverse events that are identified by Sentinel systems. It can, therefore, be argued that these schemes provide more reliable insights that mass reporting systems. There are also some limitations. Unless the sample institutions are carefully chosen then it is likely that some incidents will go unreported. It is for this reason that Sentinel systems are often used to complement rather than replace larger scale reporting schemes. This parallel approach has obvious benefits for the comparison of different techniques. The continued operation of a mass system enables safety managers to collect incident data using established techniques. The introduction of a limited number of small scale trials enables comparisons to be made with these existing approaches.

A number of problems complicate the use of Sentinel systems to guide the evaluation of alternative reporting techniques. The additional resources that are typically associated with these schemes can prevent accurate comparisons being made with the lower levels of investment that are possible in mass systems. Sentinel systems elicit more information than mass schemes almost irrespective f the particular techniques that are being used. There are also problems associated with longitudinal trials. Sentinel systems usually involve the introduction of new techniques. The novelty factor involved in learning to implement these innovations can increase motivation and involvement beyond the levels that are associated with the routine operation of mass reporting systems. Any observed advantages might decline if the techniques used in a Sentinel system were extended throughout a national scheme over a prolonged period of time. These objections limit the conclusions that can be drawn from 'direct' comparisons between the results obtained by Sentinel systems and those provided by larger-scale systems. In contrast, these techniques have been used to help validate the results obtained by more passive approaches. For example, the distribution of information obtained from a Sentinel system can be compared to assess the coverage of a national scheme. Sentinel systems can be used to uncover incidents that are under-represented in larger schemes. Hence the focussed application of active reporting techniques help to validate rather than directly compare particular incident reporting systems. Alternatively, different approaches can be trialed in separate Sentinel systems. For example, additional resources might be used to promote the application of PRISMA within one company while another is encouraged to use Tripod. Again, however, direct comparisons can be complicated by the different operating characteristics and safety records of the firms that are involved in the study.

It has been difficult to find any well-documented report of the application of Sentinel reporting within the railway industry. Pasquini, Rizzo and Save have used a variant of this approach to support the analysis of SPADs on Italian railways [666]. In this case, they specifically developed a reconstruction and causal analysis technique that was intended to support the investigation of these incidents. The first stage involved the production of video recordings during physical reconstructions at the site of the SPAD. Focus groups then discussed the film together with relevant technical document and testamonies. These discussions were then used to generate a matrix diagram similar to the MES diagrams introduced in Chapter 10.4. The actors in this case included the train driver and an on-board signal repeating system. This safety device is similar to the Automated Warning System described in Chapter 4.3. The validation of the investigation technique involved the cooperation and training of an investigative team, including two drivers. Instead of analysing recent SPAD incidents, as would have been the case in a full Sentinel trial, the comparison of the new techniques with existing approaches was based on a post hoc analysis of three SPADs on Italian railways between April 1997 and November 1998. In consequence, their study focussed on differences in the analysis of these incidents rather than on any improvements in the elicitation of information about adverse events. In particular, they argued that the new methodology helped to identify latent problems with the way in which the signal repeating system was operated. Drivers saw the warnings as a nuisance that were to be dismissed as soon as possible rather than as valuable safety information. These findings were contrasted with the insights provided by the existing reporting systems which focussed more on inattention and lack of concentration.

As mentioned, there are few examples of Sentinel systems being used to support the analysis of railway reporting schemes. Arguably the best documented example is provided by a research project that was funded by the FDA . At the start of this study, the intention was that the Sentinel system would act as a supplement to a mass reporting system. Towards the end of the study, the costs of operating the national system led many involved in the evaluation to argue that this approach could replace the existing scheme. The study address many of the problems mentioned in previous paragraphs by recruiting 23 different facilities for a twelve month period. They secured the support of Study Coordinators who were, typically, risk managers for hospitals and directors of nursing for nursing homes. These individuals participated in orientation and training sessions. These covered the purpose of the Sentinel system, project background and goals, comparison of voluntary and mandatory reporting procedures, project plans and confidentiality procedures. The initial work on the project identified 'large gaps in the knowledge of facility clinical staff regarding the importance of reporting adverse medical device events' [264]. Participating facilities were also provided with video materials for clinical staff training. These encouraged staff to follow their facility's internal procedures for reporting of adverse events. The investigators also contacted the Study Coordinators in each of the participating facilities to gather information about each facility's reporting procedures. Information about these procedures was made available to staff after the videos, mentioned above, had been screened.

The architecture used in the Sentinel evaluation involved Study Coordinators sending incident information to a central group of analysts. The analysts then telephoned the Coordinator to acknowledge receipt of the submission and to confirm any additional information that may have been obtained in the interval after the submission. Coordinators were also encouraged to contact centre staff with more general questions related to their work as Risk Managers. They requested the names of contacts at the FDA, specific information about device tracking regulations, how to use software for filing reports and also whether they were required to report certain events to FDA and manufacturers. Once a report was received by the project staff, they were reviewed with a nurse and a specialist in medical informatics. This preliminary analysis was used to determine whether follow-up requests were required to elicit further information about the adverse event.

The Sentinel project received 315 reports from 23 units between October 1997 and November 1998. 286 reports were submitted through the post in special envelopes provided to the study, 3 were sent by fax and 26 were reported by telephone The telephone reports were particularly instructive because Study Coordinators could tell the analysts about particular problems that they had experienced in completing the paper-based forms. The investigators argued that 'there is reason to believe that the level of DEVICENET reporting activity was far above the average for hospitals in the MEDWATCH system' [264]. It was estimated that if the 5,500 hospitals in the US national reporting system contributed at the same level as the Sentinel facilities then they would receive more than 100,000 contributions each year. The actual total for 'health care facilities' in 1998 was only 5,000. All of the submissions came from hospitals. More than half of all reports came from one large hospital, and a second large hospital contributed another 15% of the reports. It is instructive that even though additional resources were focussed on the participating institutions there were no reports from the six nursing homes. This was explained by the observation that nursing homes are extremely tightly regulated. There was, therefore, strong management concern that negative information might come to the attention of authorities.

As mentioned, the relatively small scale of Sentinel systems enables safety managers to encourage submissions about events that might otherwise overwhelm a national system. In the FDA study, it was determined that only 14% percent of all reports described events that could to have been submitted under the existing mandatory schemes. 56% described events that fell under the Sentinel system's voluntary guidelines. The remaining 30% fell into a gray area; 'it was not clear whether they were mandatory or voluntary' [264]. It was argued, however, that few of these events would have otherwise been reported. Many of these reports related to incidents in which it was difficult to assess whether the patient had suffered a serious injury. The additional resources devoted to the Sentinel system also supported a number of analyses that are not normally performed in larger scale reporting systems. Two senior nurse-analysts reviewed all of the submissions and classified them as ery urgent, urgent, routine monitoring or well-known problem, or not important. Approximately one-third of the reports (113) were assessed as being urgent or somewhat urgent. However, only 19 of these incidents were clearly assessed as falling within the mandatory reporting system. About half of the mandatory events (51%) needed only routine monitoring. For example, incidents were assigned to this group if any problems were already well-documented and if the regulator had already taken action to address them. In contrast, 30% of the 175 voluntary reports were rated as very urgent (2) or somewhat urgent (50). Of the 95 reports that did not directly fall under either the voluntary or mandatory regulations, 44% were either very urgent (2) or somewhat urgent (40). The Sentinel study also examined the way in which Study Coordinators classified each incident. The results of this analysis were significant because these classification represent the primary means of pattern matching in the mass reporting system, for example using the automated retrieval tools described in Chapter 13.5. The investigators felt that about a third and a quarter of the codes were incorrect [264].

Sentinel based reporting systems are not a panacea. The lack of submissions from nursing homes illustrates that this approach cannot guarantee the participation of all potential user groups. There are further concerns. For instance, the types of facilities that are recruited to many Sentinel studies may already exhibit a high degree of awareness about safety-related issues. If this is not the case at the start of the study then the additional resources that are allocated to the promotion of health and safety can quickly alter the behaviours of many of the operators and work groups that participate in the study. Hence the types of incident information that is provided by a Sentinel system will rapidly become atypical of the adverse events that affect the rest of the user communittee. This can be interpreted as a variant of the Hawthorne effect introduced in Chapter 4.3. Users will alter their normal working behaviour if they know that their behaviour is being directly or indirectly monitored. A number of potential solutions have been proposed for this problem. In particular, observational techniques can be used to identify particular behaviours that may support or weaken the operation of an incident reporting system.

15.4.4 Observational Studies

As we have seen, Sentinel systems focus additional resources to support incident reporting in a small number of 'select' institutions. This very process of selection and the additional support can help to ensure that the sample facilities no longer resemble other units within the same industry. Hence the information that they provide many not be representative of the adverse events that occur at other sites. It can also be difficult to interpret the insights derived from focus groups and interviews. Operators can express views that are not reflected in their subsequent behaviour. For example, they may strongly support the operation of a voluntary incident reporting system but fail to contribute to a scheme even when they witness an adverse event. Techniques that rely upon the statistical analysis of incident reports suffer from similar limitations. It can be difficult to identify the reasons why particular types of incidents are not reported or why certain groups of operators are reluctant to participate. The following paragraphs describe how workplace studies and other observational techniques from the field of sociology can be used to address these criticisms.

There have been many notable attempts to use techniques from the field of sociology to provide insights into the working lives of railway staff. McKenna has investigated the strategies that railway personnel have used to maintain their standard of living during times when the railways were contracting [532]. Salaman looks at the way in which drivers' attitudes changes towards their occupation and their colleagues [722]. There have also been studies of union responses to changes in management structure [63]. Heath, Hindmarsh and Luff, However, point out that relatively few of these studies focus on the everyday working activities of railway personnel [341]. There are some exceptions. For instance, Gamst has conducted a detailed study of the work and attitudes of US locomotive engineers [286]. Even this study has, however, focussed on workers' opinions and preoccupations rather than the manner in which they accomplish their everyday tasks. There has, however, recently been a move towards applying many of these sociological techniques to provide more direct insights into working behaviour. Heath and his co-authors are amongst the leading figures in this area. Others include John Hughes and his colleagues in air traffic management [376] and Berg in the field of medical safety [78].

In contrast to many previous design techniques, these studies do not focus narrowly on the operation of high-technology systems. In contrast, they consider human-human as well as human-machine interaction. There is also a concern to consider the way in which diverse communication media, including physical artifacts such as pencil and paper, are used to coordinate and inform group activities. Many of the proponents of this approach have written about 'rescuing' the study of technology from cognitive science which concentrates too narrowly on the psychological characteristics of individual users. In practical terms, these 'workplace studies' involve participants joining the groups that they are observing for prolonged periods. They will often follow the same shift patterns as the individuals that they are studying. This is important because it helps the observer to build up a mass of background information that may be necessary to understand the significance of the events that they witness. It can also provide some indication of the prolonged impact that stress, fatigue and other workplace factors can have upon operators and managers.

During these periods of observation, investigators compile field-notes. They can also use audio and video recordings. Clearly, however, the conspicuous compilation of these records can remind workers that their actions are being observed. The nature of these records depends partly on the context in which the study is taking place and partly also on the forms of analysis that will be used after information has been elicited. For example, conversation analysis provides important insights from studying the vocabulary and structure of workers' conversations. This technique is only feasible if transcripts can be reconstructed from field-notes or other recordings. Other forms of analysis require less direct records, such as the construction of social networks to model the way in which different working groups interact [694]. They may, however, require longer periods of observation to ground any potential conclusions in observed behaviour.

There are clear ethical problems in exploiting these techniques to monitor incident reporting systems. For example, observers may witness adverse events and 'near misses' that are not notified by any of the operators who were involved. Other observers have seen users struggled to operate computer equipment that they themselves were familiar with. This creates a considerable dilemma in many safety-critical contexts. If the observer decides not to act then there can be adverse consequences. Conversely, an ill-advised decision to intervene can exacerbate rather than resolve a potential incident.

Brevity prevents a complete introduction to the many different approaches that have been developed to support observational and workplace studies. In passing, however, it is important to stress that most of these technique specifically avoid the generation of hypotheses before the study is conducted. Such concepts should emerge during the observation as more information is gathered about the workers and the context of their daily lives. This guiding principle helps to ensure that analysts do not selectively filter their observations to support pre-formed hypotheses. It is also important to stress that some 'ethnographers' deliberately reject the idea that observational techniques should be used to support particular theories [305]. This argument stems from the discussion that was introduced in Chapter 10.4. Causal asymmetries complicate the task of explaining what actually led to an observed behaviour. Experiments attempt to identify causal relationships by recreating two or more identical situations in which a causal factor is systematically varied to determine whether or not it will have the prediced outcome. This leads to problems because it can be difficult to ensure that all relevant causal factors have been controlled between the different conditions. Experiments may also have limited 'ecological validity'. This prevents conclusions from being generalised beyond the laboratory into the real world. For example, a study may focus on an analysts ability to use Management Oversight and Risk Trees (MORT) or a similar technique in a silent room without the interruptions that would punctuate their work in an office. In consequence, observational techniques cannot easily be used to provide objective, quantitative comparisons between different reporting systems. They can, however, provide rich insights into the way in which different systems can influence reporting behaviour in complex working domains.

The Ladbroke Grove rail inquiry provides considerable into the potential application of observational and workplace studies to monitor the operation of incident reporting systems. It expressed considerable concern over examples of poor communication and record keeping during the analysis of SPAD incidents. The report argued that 'it is essential that an organisation has a system to record what it has learned, and a process to pass those lessons on to its employees' [195]. Railtrack procedure RT/D/P006 specified that the HQ Production Directorate should monitor and record the implementation of each recommendation. A record could only be closed once the corresponding recommendation had been fully implemented. The Formal Inquiries Process Manager was responsible for following up those recommendations that were directed to Railtrack Headquarters. Although his job description 'clearly envisaged that the progress of recommendations would be tracked after their allocation to individuals, (the Formal Inquiries Process Manager) told the Inquiry that he was given guidance by his managers to the effect that he was not responsible for ensuring that a recommendation was acted upon, but simply that someone had accepted responsibility for it' [195]. The Inquiry concluded that no-one assumed responsibility for monitoring the implementation of recommendations. Cullen also observed that had it not been for the accident and the associated investigation then the shortcomings for tracking incident recommendations might not have been discovered. Faced with this analysis, a new recommendation 'clearing house' was established to collate, prioritise and monitor the implementation of any proposed changes. An important responsibility of the new organisation was to report directly to the Board of the infrastructure company every four weeks.

This analysis indicates the potential application of workplace studies. It emphasises the way in which everyday practice can, over time, depart from published procedures and guidelines. In this example, the responsibilities of the Formal Inquiries Process Manager changed from those documented in the job description and from the intention behind procedure RT/D/P006. This does not imply that such a departure would necessarily have been identified had an observational study been conducted. However, this is precisely the type of working practice that can be observed by these techniques. Hammersely and Atkinson refer to the ways in which the production and use of documents, such as the recommendation reports, are 'socially organised activities' [305]. Ethnographers must, therefore, question the way in which a document is written and distributed. They must also consider what is the purpose and intention behind a document. Ethnographers should also compare the actual use of document against the stated intentions that justify its creation. Differences between observed practice and intended use cannot easily be elicited using monitoring techniques such as focus groups, interviews and questionnaires.

It is important to emphasise that observation techniques can be used to monitor incident reporting systems in situations that extend well beyond the workplace. This is particularly important within the rail industry. Members of the public are often involved in adverse events as well as those who work directly for operating and infrastructure companies. Ethnographic techniques have been widely used to study 'deviant' behaviour. For instance, Popkin et al have recently used this approach to observe patterns of behaviour and control structures within the gangs in many Chicago Public housing developments [684]. Such work provides insights into the relationships between drug use, vandalism, trespass and violence. Many of these activities can have an impact on rail safety. For instance, Smithsimon has conducted a prolonged study of graffiti writers. He argues that ethnographic techniques provide one of the few effective techniques that can be used to gain insights about the behaviours of these individuals and groups. He stresses that 'running from the cops, using the right language, wearing the right clothes: like other ethnographic studies, the right signals and actions, even by an outsider, help gain access to graffiti writers' [747]. Only in this way are 'respondents' willing to provide information about their work and discuss the law-breaking that is a prerequisite for many of their activities. This participation is essential to gain the confidence of individuals who often 'hide behind' the image of a street-wise 'outlaw graffiti artist'. The insights provided by Smithsimon's work can be illustrated by the following except:

> "Hasp and some friends of his offered to show us other, illegal graffiti on the walls lining the adjacent railroad tracks. But he refused to go onto the tracks while a truck belonging to the railroad was parked between the Phactory and the rail yard. John, the photographer, suggested that the railroad employee in the truck probably would not care if we walked down the tracks, but Hasp explained that the Phun Phactory has had repeated problems with the railroad and the transit authority... Before we could learn more about the tension between graffiti proponents and opponents, the truck moved and we traipsed down the tracks, looking at murals. As I spoke with Seac and Ker, they pointed out well-done murals along the walls... 'Get away from there!' yelled an angry voice. Someone on the bridge glared down at us, then dashed away. Hasp and the other writers told us it was someone from the MTA's vandal squad, which focuses on pursuing graffiti artists. We started heading toward the Phactory to get off railroad property. We were about three blocks from the street the Phun Phactory was on, and the row of warehouse walls and razor wire fences along the train tracks meant that if a cop were to get to that street (where the MTA truck had been parked earlier) before we did, we would have been trapped... Meanwhile, I opened my notebook and wrote down the names and descriptions of the writers I had met during the day. 'You writing this down?' asked Hasp. 'What?' I asked. 'What are you writing? This story?' he asked me. 'Oh, no. I'm just writing down everybody's names, and stuff like that'. I flashed a nearly blank page of the notebook toward him, too quickly for him to read much. 'Oh. OK. Cause I thought you were writing this down. Don't write down this', he said..." [747]

Smithsimon's work provides important insights into the behaviour of the graffiti artists who he observed. These insights go beyond the information that can be obtained from incident and accident reports. In particular, it can provide information about potentially dangerous behaviours that are not observed by railroad employees and are, therefore, not reported. Ethnographic studies can also provide insights into the attitudes and shared values that motivate individuals or groups who are involved in trespass or vandalism. For example, Smithsimon argues that 'graffiti represents people's desire to assert their presence in the world through pictures, words, and artistic interpretation' [747]. This conclusion arguably captures the strength and weakness of ethnographic techniques in this domain. It can be difficult to go from the insights that they provide to the recommendations that might avoid future incidents or mitigate the consequences of potential accidents.

Vandalism and trespass are not the only forms of 'deviant' behaviour to be investigated using observation techniques. For example, these approaches have yielded valuable insights into incidents that involve intersections between the road and rail systems. Several studies have shown the difficulty of conducting other forms of investigation into driver behaviour [640, 857]. Individuals will typically take fewer risks and are more likely to obey 'the rules of the road' if they believe that they are being observed. Experimental studies, therefore, seldom yield the violations and extreme behaviours that are witnessed in other contexts. For instance, Burnham's recent study in Alabama observed the behaviour of 862 vehicles as they approached STOP signs at railroad-highway grade crossings [117]. 18% came to a full stop, 50% made a slow rolling stop and 32% did not stop at all. These observations have been interpreted as showing that the majority of drivers do not understand the meaning of the symbols that are used to indicate railroad-highway grade crossings [117]. This is a strong conclusion; an alternative interpretation is that the majority of drivers understand but deliberately choose to ignore the warning signs. Burnham concluded that 'one of the most widely recognised and often overlooked traffic safety axioms is the principle that over use provokes abuse... for a traffic control sign, signal, or pavement marking to be of value it must not be overused' [117]. The difficulty of interpretating observations is a considerable barrier to the practical application of these techniques. Many ethnographers deliberately avoid the 'constructivist theories' that might explain such observed behaviours [305] Unfortunately, these explanations are often essential if we are to be confident in generalising insights from previous failures to predict the likelihood of future incidents and accidents.

It is important to stress that the difficulty of interpreting observed behaviour does not sacrifice the utility of workplace studies and ethnographic techniques. These approaches can often yield important insights even though the complex mechanisms that affect human behaviour are not made explicit. This is best illustrated by the way in which observational approaches can be used to analyse the effectiveness of recommendations that are proposed in the aftermath of adverse events. Again, there are problems with using focus groups or experimental techniques to evaluate these proposals. Expressed opinions may not predict actual behaviour, laboratory conditions may not control all of the factors influencing decision making and performance. Observational techniques provide more direct insights into the 'real world' benefits of potential safety devices. For instance, drivers and pedestrians have been killed and injured by incidents in which they stopped for a first train but then failed to wait for a second train to cross at a junction. The Maryland Mass Transit Administration (MMTA), therefore, tested a 'second train warning' system [524]. This was based around a sign that was illuminated shortly after the first train passed if there was another train approaching. The system was tested at for a ninety day evaluation period at one crossing in Timonium, Maryland. An independent evaluator assessed the performance of the sign by observing driver behaviour and by analysing videotapes. The study concluded that 'risky' driver behaviour decreased by 36% after the installation of the system. Such behaviours can be defined in terms of a specified minimum safe interval between the moment when a vehicle enters the junction and the time at which the second train arrives. A similar study conducted in Los Angeles also used video tape observations on a 'live site' to demonstrate a 14% reduction in risk behaviour. In this instance, 'risky' behaviours were defined to occur when a pedestrian entered the track area six seconds or less before the train entered the crossing [439].

Another major problem complicates the application of observational techniques to monitor the performance of incident reporting systems. In many applications, there are relatively few 'serious' adverse events. In consequence, it is unlikely that an ethnographic or work place study will observe such an incident or accident. These techniques can still provide insights into more frequent, less critical events. However, it is also possible to use some of the analytical techniques that are associated with workplace studies in a post hoc manner. For example, Law has used this approach to demonstrate that 'the character of explanation and cause is relevant in thinking about safety-critical socio-technical systems such as railways' [476]. His analysis of the Ladbroke Grove report and enquiry focuses on a 'rhetoric of spatiality'. Many of the questions and responses during the investigation referred to location, such as 'where does responsibility lie?' or 'Thames Trains could be prosecuted if an incident occurred where driver error was partly to blame'. This use of language reveals that the analysis of failure is understood in terms of distinct 'pigeon-holes' or 'compartments' that are associated with technical, managerial or psychological domains. He identifies other forms of spatial reference. For example, failures can be 'located' within particular subsystems. These views are criticised. Systems responses are compromised by the way in which many social systems are incomplete and unstable. Law argues that in many cases 'the world is simply too fluid and disarticulated' for failures to be located within particular systems. The techniques that Law uses are very similar to those exploited by other sociologists to directly analyse the observations derived from workplace studies. What makes his approach different is that instead of working from his own field notes, his analysis is 'grounded' in the documents produced by an investigation. He is, therefore, sensitive to the context in which such documents are produced. They cannot be analysed at face value but must be seen as publications that are intended to achieve particular objectives.

In passing, it is important to note that Law's ideas have important consequences for analysis of causation in incident and accident reports. His spatial rhetoric can be used to draw conclusions that are broadly similar to many of the other authors in this area who were introduced in Chapter 10.4. Law criticises the idea of blaming a driver or even the safety culture in an organisation because these are regional interpretations. They place responsibility in a specific pigeon-hole and assumes that blame can be confined within particular boundaries. If the safety culture is at fault then operators can be absolved? However, Law extends his analysis to identify weaknesses in the systemic view of failure. Many of these criticisms have been implicit in the previous chapters of this book, for example in the analysis of some of the findings from the NASA missions in Chapter 9.3. Law argues that the concept of systemic failure often erodes the boundaries between locations but also often implicitly relies upon the idea that there can be a single focus for particular activities. The proponents of this view, he argues, often talk about 'bringing the system together' or of 'a meeting of minds', For much of the time 'the ordering of the railway is indeed imagined and performed in terms of a system with a more or less strong centre' [476]. This view is, however, flawed. Law argues that the rail system best viewed as a system of dynamic and changing relationships that cannot easily be ordered in such a manner:

> "This is that speed and rapid change together push towards tightly-coupled systems with dense webs of self-sustaining relations. But such systems are best avoided in safety-critical locations. This is because, as we have seen, when things go wrong disruption is rapidly and unpredictably transmitted through the system. Failsafe mechanisms and the tight control of centralised management may work most of the time. But sometimes they will fail. And then they fail there is no play. No slack. Everything falls down. The conclusion is that partially connected, multiply ordered, ambiguous and not very coherent systems are usually more robust. And the corollary is that if we find that we are proposing technologies that demand tight systems then we need to stop and think. This the ultimate lesson of the Ladbroke Grove tragedy. It is that we have unwisely created a world which demands coherent systems" [476]

It is interesting that sociological approaches should at the same time yield immensely detailed observations of group and individual behaviour as well as such high-level insights into safety-critical organisations. Unfortunately the broad range of these approaches cannot overcome some of the problems that arise when attempting to use the resulting insights to improve safety. Workplace studies and sociological analyses seldom yield direct recommendations. In many ways, this is the point behind the techniques. The insights they provide inform decision making but do not automatically help to shape or focus those decisions. In contrast, statistical techniques can be tailored more directly to support particular hypotheses about reporting behaviour and the operation of reporting systems.

15.4.5 Statistical Analysis

Previous paragraphs have introduced different forms of statistical analysis that support the monitoring of incident reporting schemes. These include simple incident and reporting frequencies as well as threshold models, such as UK Railway Safety's Precursor Indicator Model, and more advanced techniques, including least squares regression used for trend analysis [698]. Several specialist textbooks provide an introduction to the particular mathematical approaches that support these techniques []. In contrast, the remainder of this section focuses on the managerial and organisational issues that must be considered when using statistical methods to support the monitoring of incident reporting systems. For instance, it is important to consider whether particular numerical values can yield 'valid' insights into the performance of the underlying systems. It is for this reason that the Transportation Safety Board of Canada do not preset accident totals for particular railways [788]. They argue that 'the track, rolling stock and personnel in an occurrence may all belong to different companies; also an occurrence may have several contributing factors'. Presenting data about one of these factors might be 'misleading' and there is a danger that misinterpretation of the data could have an unfair affect on a company's 'competitive position'. As we have seen, other organisations reject this argument and instead rely upon normalisation techniques to help make valid comparisons between different organisations. For instance, the independent review of Australian rail safety argued that without 'measurable, appropriately normalised data' it is impossible to determine whether the industry is becoming safer; whether passenger safety is improving or not and whether there are significant trends in freight and passenger train incidents and accidents [55]. Table 15.13 illustrates this point. It documents the number of fatalities associated with different modes of transport and is taken from the the Australian rail report cited above. It is difficult to make direct comparisons between these statistics because the table does not record the risk exposure associated with each mode. For example, the relatively high number of fatalities associated with road travel can be offset against the disproportionately large number of journeys or trip distances that are made each year using this form of transport. Similarly, the low number of deaths from maritime transportation cannot be correctly interpreted without information about the numbers of people involved in this industry.

	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	% Fall
Rail	60	68	76	48	61	52	39	48	34	39	-35%
Road	3197	2935	2601	2324	2172	2048	2044	2126	2031	1876	-41%
Water	69	60	92	89	77	73	62	60	57	48	-30%
Air	57	97	74	65	79	87	50	65	71	49	-14%
Other	6	3	6	2	1	0	1	2	1	2	
Total	3389	3163	2849	2528	2390	2260	2196	2301	2194	2014	-41%

Table 15.13: Transport Accident Deaths, Australia 1988-1997 by Year and Mode [55]

Unfortunately, normalisation does not offer a panacea for the problems of interpreting incident statistics. As we have seen, it can be difficult to agree upon appropriate criteria. For example, the information in Table 15.13 might be normalised to present the ratio of deaths to passenger miles. This would clearly be appropriate within mass transportation modes such as the road and rail systems. It is less certain that this normalisation would yield meaningful data for the maritime industries. Relatively small numbers of passengers are carried; fatalities are often associated with shore personnel servicing shipping. It might, therefore, be more appropriate to normalise according to the tonnage carried within each industry. This would raise further problems in the analysis of rail and road data unless a distinction was made between fatalities involving freight and passenger services.

The choice of normalisation criteria can have an important impact on the calculation of trend statistics. Different industries respond in different ways to changes in the underlying economic cycle. For instance, adverse conditions in the world economy during 2001 have arguably had a more profound impact on air transport than they have upon road transportation. Normalisation factors that ignore the impact of this down-turn upon passenger traffic might, therefore, exaggerate any associated drop in the incident rate. Similarly, as with many forms of statistical analysis, trend identification can be profoundly affected by the base date that is used in any comparison. It can be difficult to interpret the significance of the percentage reductions in Table 15.13 without understanding the reasons for selecting 1988 as the starting point.

The problems of normalisation and of trend analysis are general in the sense that they affect the monitoring of many different systems. A number of further problems specifically affect the monitoring of incident and accident statistics. Many of these stem from the causal asymmetries that have been described in previous chapters. Statistical returns often depend upon subjective analysis to determine the causal factors behind past events. Changes in aggregate data can reflect changes in interpretation rather than new forms of failure within an industry. In Table 15.13, problems arise when compiling statistics for incidents involving more than one mode of transport. As we have seen, many rail incidents involve pedestrians and road vehicles. It is difficult to interpret the statistics provided in this summary without some explanation of how such accidents would be encoded. Such fatalities might be associated with the rail system or with the road system or both.

There are a number of specific problems that complicate the compilation of incident and accident statistics. It is often difficult to know whether or not particular effects can be attributed to an adverse event. For example, the data is Table 15.13 cannot easily be interpreted without additional information about the definition of a transport-related fatality. This is important because an individual may die several days after an incident has occurred. In extreme cases, they may receive injuries that contribute to their death many months or even years after the adverse event. This attribution problem is exacerbated for occupational illnesses where individuals may also be exposed to other contributory factors within their wider environment. In the UK, this reasoning led to the Court of Appeal 'Fairchild' decision (December 11, 2001, Lord Justices Brooke, Latham and Kay). This focussed on cases brought by victims of the mesothelioma lung disease. Mesothelioma is linked to exposure to asbestos products. The defendants were all employers or operators of premises where asbestos was being used or cleared. The Appeal Court Justices refused damages to the claimants because mesothelioma 'is a single indivisible disease and a claimant cannot establish on the balance of probabilities when it was he inhaled the asbestos fibre, or fibres, which caused a mesothelial cell in his pleura to become malignant'. The impact of this ruling has been profound. It has subsequently been argued that since it is impossible to identify the individual fibre that causes the mesothelioma then any exposure to asbestos should be regarded as a possible cause. From this it follows that the level of liability should reflect the degree of exposure to any potential cause of the disease. It remains to be seen whether this line of argument will stand against the Fairchild decision.

Incident and accident statistics can only be interpreted correctly if analysts understand the criteria that guided the collation of source data. Using different definitions for reportable incidents can lead to very different statistics being presented. This can be illustrated by the way in which the Transportation Safety Board analysis the performance of its rail network using both Canadian criteria and the criteria proposed by the US Federal Railroad Administration requirements. The Canadian criteria since 1992 consider that all main-track and non-main-track accidents are reportable as long as the damage to rolling stock renders it unsafe. The Federal Railroad Administration requires a minimum dollar damage threshold of $6,300 US for all reportable accidents. This policy of using dual criteria in the collation of accident statistics enables accurate comparisons to be drawn between these two different approaches. For instance, the data compiled by the Canadian Pacific Railway for the January - August period of 1994-1996 show very different trends in main track derailments depending on the criteria used.

> "... when TSB reporting requirements were used, CPR's main line derailments were 30% higher during the January - August period of 1996 than during the corresponding period in 1994 and 1995. However, when FRA reporting criteria were used, the number of main line derailments remained unchanged throughout the period being examined. During this three year period, an average of 75% of occurrences that met TSB but not FRA accident reporting guidelines involved derailments of only one car." [784]

The maintenance of different statistics to reflect different reporting criteria is a general problem. For example, it affects many agencies and commercial organisations that implement their own local criteria but must then follow different agreed criteria when reporting to higher organisations. For instance, most European Air Traffic Management organisations must pre-process their incident statistics before submitting them to EUROCONTROL [423]. One consequence of this is that analysts must always check which criteria are being applied when interpreting meta-level trend information. There are additional complexities. For instance, reporting systems often revise their own criteria. This creates problems when analysts attempt to draw comparisons between more recent statistics and those gathered under previous reporting criteria. For instance, the Canadian Transportation Safety Board revised its guidelines in August 1992. Before this time, derailments and collisions were only reportable if casualties or dangerous goods were involved or for main-track accidents if there was property damage in excess of a monetary threshold. As we have seen, since 1992 all main-track and non-main-track accidents are reportable as long as the damage to rolling stock renders it 'unsafe'. After 1992, all crossing accidents are reportable. Prior to that year accidents at farm and private crossings were reported only if they involved a casualty/dangerous goods/derailment resulting in property damage in excess of a monetary threshold [788]. It is difficult to underestimate the consequences of such changes on the compilation of incident statistics. Some occurrence categories previously regarded as incidents were now regarded as accidents. Other types of occurrence were no longer reportable. The changes also made it difficult to calculate trends, data reported under the new definitions could not be directly compared to historical data that was gathered under the previous criteria. Where possible the Safety Board revised previous data in an attempt to adapt it to the new criteria. They did, however, emphasise that 'caution is required' when comparing statistics before and after the reporting requirement change and that 'the interpretation of the results from recent years has been clouded because of the change in TSB reporting requirements'.

The problems of gathering, of normalising and of interpreting the statistical information used to monitor reporting systems has led some safety-related organisations to develop extensive criteria to guide many different aspects of data analysis. For instance, the UK Health and Safety Executive have been involved in an initiative to 'revitalise' the use of Health and Safety targets to promote national performance. Part of this work has involved the development of a 'note' to ensure the validity of the statistics that will be used to measure progress towards these revised objectives [339]. This note lays out a number of general principles:

1. "Progress measurement will involve more than one data source and some adjustment or integration of data from the different sources will be necessary; as a rule this will only be appropriate at the global level.

2. Percentage changes over time are what matter for monitoring progress against the targets, so efforts should be focused on measuring change; estimates of absolute levels may vary as information sources evolve.

3. In assessing trends and progress over the strategy period, statistical modelling techniques will be used to limit the impact of sampling variability in the figures for individual years.

4. To support the outcome data on injuries and ill health, supplementary approaches should be explored, for example collecting data on economic, social and cultural factors.

5. The data and methods used for progress measurement will be National Statistics, so the methods will be subject to independent quality review and stakeholder consultation.

6. A report on progress will be prepared each autumn, comparing the latest data with those for the base year (1999/2000). For at least the mid- (2004/5) and end-point (2009/10) of the strategies, this report will incorporate external peer review." [340]

The note also includes specific sections describing techniques for describing the injuries target. For instance, the reporting rate for less serious injuries will be used to calculate an adjustment for the under-reporting of major injuries. It also lays out criteria for the statistical analysis of work-related ill health. Existing data sources will be refined, for example to account for 'the effects of raised awareness'. The note also promises to identify new sources of statistical data, including workplace-based surveys. Diseases with long latency periods between exposure and health outcome will be included but will be handled separately from other illnesses for the reasons described in previous paragraphs.

The note builds on International Labour Organisation recommendations by arguing that data from different sources should be integrated 'to produce an overall judgement about progress'. The note also outlines some of the high level problems that arise from this approach. The integration of data can often involve labour intensive adjustments, for instance where they may be subtle differences in the periods over which aggregate data has been compiled. Also 'where one or more of the sources involve sampling, the reliability of the resultant estimates will reduce as the level of disaggregation increases' [340]. There are further technical problems. For example, statistical measures are subject to random error. It is, therefore, usual to indicate a central estimate for any measure together with upper and lower confidence levels around this estimate. It, therefore, follows that the lower confidence limit must equals or exceed the target value before analysts can argue that the target has been achieved. Similarly, a target can be shown not to have been met only if the upper confidence limit falls short of that target. If the target value falls between the upper and lower confidence limits, no definitive statistical judgement can be made. These importance of confidence levels is often underestimated or ignored by analysts who interpret incident statistics. It can be argued that this book also falls into this trap by postponing any discussion of these issues until relatively late in our exposition. The HSE emphasise that these issues must be considered precisely because even our best measurements are subject to 'uncertainty', to 'under reporting' and to 'sampling error'. For example, a 1999 labour force surveys reported that there were 380,000 reportable injuries to workers. Of these, 343,000 were to employees. Employers, however, only completed 161,000 injury reports. Self-employed people made 1,599 non-fatal injury reports in 1998/99. This compared with 35,000 injuries estimated by the same survey. This suggests a reporting level of less than 5% for the self-employed. Similarly, the HSE report that the margin of uncertainty on disease estimates drawn from self-reported surveys is up to 30% for stress/depression/anxiety, upper limb disorders and back disorders. The margin of error is assumed to be lower for data collected using Sentinel systems, such as the FDA's trials mentioned in Chapter 13.5. However, the HSE report the lack of any agreed methodology for measuring sampling error in these systems. It may not, therefore, be possible to use statistics to determine whether or not an incident reporting system has actually met a particular target!

The statistical note, mentioned above, argues that the relatively large statistical errors associated with accident and incident reporting represent an 'unsatisfactory situation' and urge analysts to 'reduce the statistical uncertainty to as low a level as possible' [340]. Partial solutions include the use of statistical modelling across several years. Data can be taken from successive years rather simply comparing the base and final years of the sample. Overlaps between samples for successive periods can also be used [698]. The HSE argue that 'the precise statistical models will depend on the series that emerge'. A uniform decline in incidence rates would justify the use of simple linear regressions. More complex methods may, however, be needed to explain complex trends. This would be the case if an initial fall in incidents rates was not sustained. The HSE note that a recent analysis of German accident rates suggested that trends were best modelled as an exponential rather than a linear decline. They emphasise that these decisions must be 'data driven'. They must not be influenced by whether or not they provide a favourable answer [340].

Even if statistical studies obey well intentioned guidelines, such as those cited above, there can be problems in the monitoring of reporting systems. In particular, problems can arise from the degree of sophistication implied by those guidelines. The end recipients of the statistics may fail to understand or interpret the information that they are being provided with. This point can be illustrated by recent problems in the presentation of UK SPAD statistics.

> "HSE statisticians have advised that the method of presenting SPAD information should be revised to avoid potentially misleading interpretation of the standard analysis. Previously, ... the standard presentations each month have been represented using the ratio of that month's SPAD count with the average of the corresponding months of the six preceding years. These data were then plotted out month by month, together with a 'trend' line fitted to these points (technically a linear least-squares squares regression line). The main visual message of this representation is the trend line, which could lead to readers naturally assuming that this represents the trend in SPAD numbers (i.e. if the slope is up, SPADs are increasing, and vice versa). However this natural assumption is wrong. The slope of the trendline indicates whether the change in SPAD numbers (change being measured over the past six years) is getting bigger or smaller month by month, regardless of whether the change itself is upwards or downwards. A flat trend indicates a steady increase or decrease: it does not discriminate between the two. The

> previous presentation thus shifts attention from the issue of primary interest (are SPADs increasing or decreasing?), to a secondary issue (is the rate of change in SPAD numbers increasing or decreasing?)." [356]

In other words, the statistical techniques were appropriate for the data being analysed. However, the graphical presentation of those statistics was easily misinterpreted. Fortunately, there are a number of texts that discuss appropriate presentation formats for such information. For instance, Tufte's books warn analysts about a range of common biases that affect our interpretation of the graphical presentation of statistical information [790]. Rather than repeat this material, the following pages focus on a range of computer-based visualisation techniques that have been developed to support the monitoring of incidents and incident reporting systems.

15.4.6 Electronic Visualisation

The previous section has argued that many safety managers and regulators have difficulty in interpreting the statistics that are collated to support the monitoring of incident and accident reporting systems. Engineers and scientists 'need an alternative to numbers' when analysing such complex data sets [527]. Graphical visualisations can be used to address this problem. However, as the previous quotation illustrates, there are also associated problems when people fail to correctly interpret those graphical representations. Previous arguments can be illustrated by UK SPAD statistics. For example, UK SPADs are categorised according to a severity classification scheme. The following definitions introduce the term 'signal overlap'. This is the distance specifically provided after signals as a safety margin to cater for misjudgement or problems with the train braking systems. Italics are also used to represent changes from previous definitions. The associated HSE reports do not describe these changes in detail. It must be assumed that the HSE have updated the historic data to reflect the new definitions. This further emphasises the importance of providing sufficient information about the criteria used when compiling statistics so that readers can correctly interpret the impact of such changes in the definitions of particular categories:

- Category 0: Not entered
- Category 1: Overrun 0 to 25 yards, *overrun not exceeding overlap*, and no damage, injuries or deaths.
- Category 2: Overrun 26 to 200 yards, *overrun not exceeding overlap*, and no damage, injuries or deaths.
- Category 3: *Overrun greater than overlap* plus all overruns greater than 200 yards and no damage, injuries or deaths.
- Category 4: Track damage only with no casualties.
- Category 5: Derailment with no collision and no casualties.
- Category 6: Collision (with or without derailment) and no casualties.
- Category 7: Injuries to staff or passengers with no fatalities.
- Category 8: Fatalities to staff or passengers.

Table 15.14 presents an extract from the associated UK SPAD statistics. Although it is relatively easy to extract salient features from this data, it is important to remember that it only represents a very limited snapshot of previous SPAD incidents. The introduction of additional information, such as the operating companies involved in each SPAD, can add considerable complexity to the interpretation of these statistics. The visualisation of this additional detail will be addressed in subsequent pages. For now, it is sufficient to observe that the statistical information in Table 15.14 can be visualised in a number of different ways.

	94/95	95/96	96/97	97/98	98/99	99/00	00/01
0	0	0	0	0	0	0	0
1	43	38	39	35	35	38	33
2	28	33	26	28	29	31	24
3	20	24	28	29	28	27	37
4	7	4	5	4	5	2	2
5	1	1	2	2	2	2	4
6	0	0	0	0	0	0	1
7	1	1	0	1	1	1	0
8	0	0	0	0	0	0	0

Table 15.14: Percentage of Total SPAD Incidents by Severity by Year [355]

Figure 15.1 illustrated a 'conventional' visualisation for this data, based on several similar graphs in the HSE SPAD returns. As can be seen, the percentage of incidents at particular levels of severity are mapped onto the Y-axis. The X-axis is used to denote the data set from each year's returns. Lines can be plotted between individual data points to illustrate potential trends between SPAD incidents at particular levels of severity. It is important that safety managers and the administrators of incident reporting systems carefully consider the strengths and weaknesses of such visual representations. On the one hand, Figure 15.1 provides a relatively clear overview of the data in Table 15.14. On the other hand, it fails to capture the confidence intervals that was emphasised as a important part of any statistical analysis in the previous section. This can be done by introducing bars above and below the central point for each data value in Figure 15.1. There are further problems. A far greater percentage of incidents occur at severity levels 1,2 and 3 than 0, 4, 5, 6, 7 and 8. Similarly, the relatively low proportion of incidents at severity levels 0 and 4to 8 also creates considerable overlap between the 'trend lines' in Figure 15.1. It is difficult to distinguish between small differences in the proportion of incidents at these severity levels over the time period concerned. Such problems might be addressed by using two different scales on the Y-axis. A relatively large interval might be used for the large differences between values for severity levels 1,2 and 3. A more fine-grained scale might be introduced for the lower proportion of incidents at severity levels 0 and 4 to 8. It is important to emphasise that these caveats are typical of the problems that complicate the visualisation of incident reporting statistics in many different industries. The severity distribution reported by the HSE is similar to those identified by EUROCONTROL in air traffic management [423] and by maritime reporting agencies [368].

A number of further problems affect the visualisation in Figure 15.1. In particular, readers may fail to notice that the values in this graph represent percentages rather than absolute frequencies for SPAD incidents. This potential pitfall is addressed by the alternative visualisation in Figure 15.2. In this representation, the data is cumulatively mapped onto a percentage scale so that users can see the way in which different levels of severity contribute to that total proportion of SPAD incidents. This visualisation arguably 'exposes' the relatively large percentage of level 8 events in 1994-1995 more clearly than the previous representations. As with Figure 15.1, number of further criticisms can be made. It is difficult to identify the precise percentages for each severity level, especially for the categories 0 and 4 to 8. This could be addressed by introducing labels to provide the numeric values in table 15.14. These might also provide absolute numbers of SPADs in each category to avoid the confusion between proportions and frequencies, mentioned above.

Many further visualisations can also be used to represent the data in Table 15.14. For example, Figure 15.3 uses a form of 'radar' presentation. A separate axis is drawn for each of the data sets being considered. Each axis begins at a common origin and ends at a point such that there is a uniform distance between consecutive axes on the circumference of a circle. Lines can be drawn between the proportion of incidents at a particular severity level for each data set. If the proportion remained the same then one would expect to form a regular shape. In this case, with seven data sets we would anticipate regular heptagons. This is illustrated by SPADs at severity levels 1 and 2 whereas

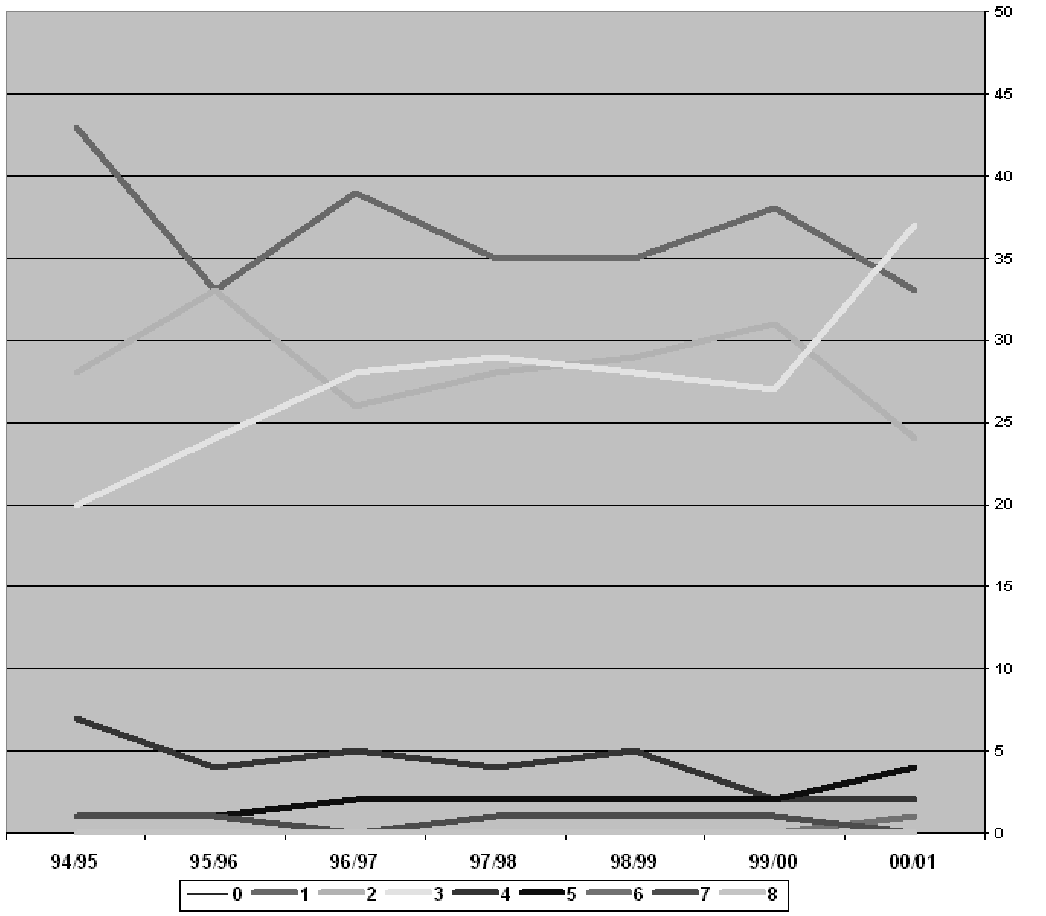

Figure 15.1: Static Conventional Visualisation of SPAD Severity by Year

a clear distortion can be observed at level 3 in 2000-2001. Again, however, the use of a uniform scale along each of the seven axes creates problems in distinguishing the relative proportion of incidents in categories 0 and 4 to 8. The key point in introducing each of these alternate representations is that their strengths and weaknesses must be matched against the particular requirements of their intended users. Figure 15.2 arguably provides a more accurate impression of the relative percentages at particular severity levels. It would, therefore, be appropriate for presenting this data to analysts who were not frequently required to monitor these statistics. In contrast, Figure 15.3 and Figure 15.1 might be used in circumstances where readers might be expected to recognise that these images did not record incident frequencies at particular severity levels .

Most of the visualisations that support the monitoring of incident reporting systems are generated by teams of statisticians who work from incident databases. They publish summary documents that are distributed to more senior management at regular intervals. This approach can create a number of limitations. For example, there may be a significant delay between the collation of incident data and the publication of the graphs and forms that are used to monitor those statistics. This prevents regulators and managers from taking prompt action in response to sudden changes. Electronic visualisation systems avoid these problems by linking graphical representations to on-line incident information systems. Users can automatically update the values in a visualisation to reflect the most

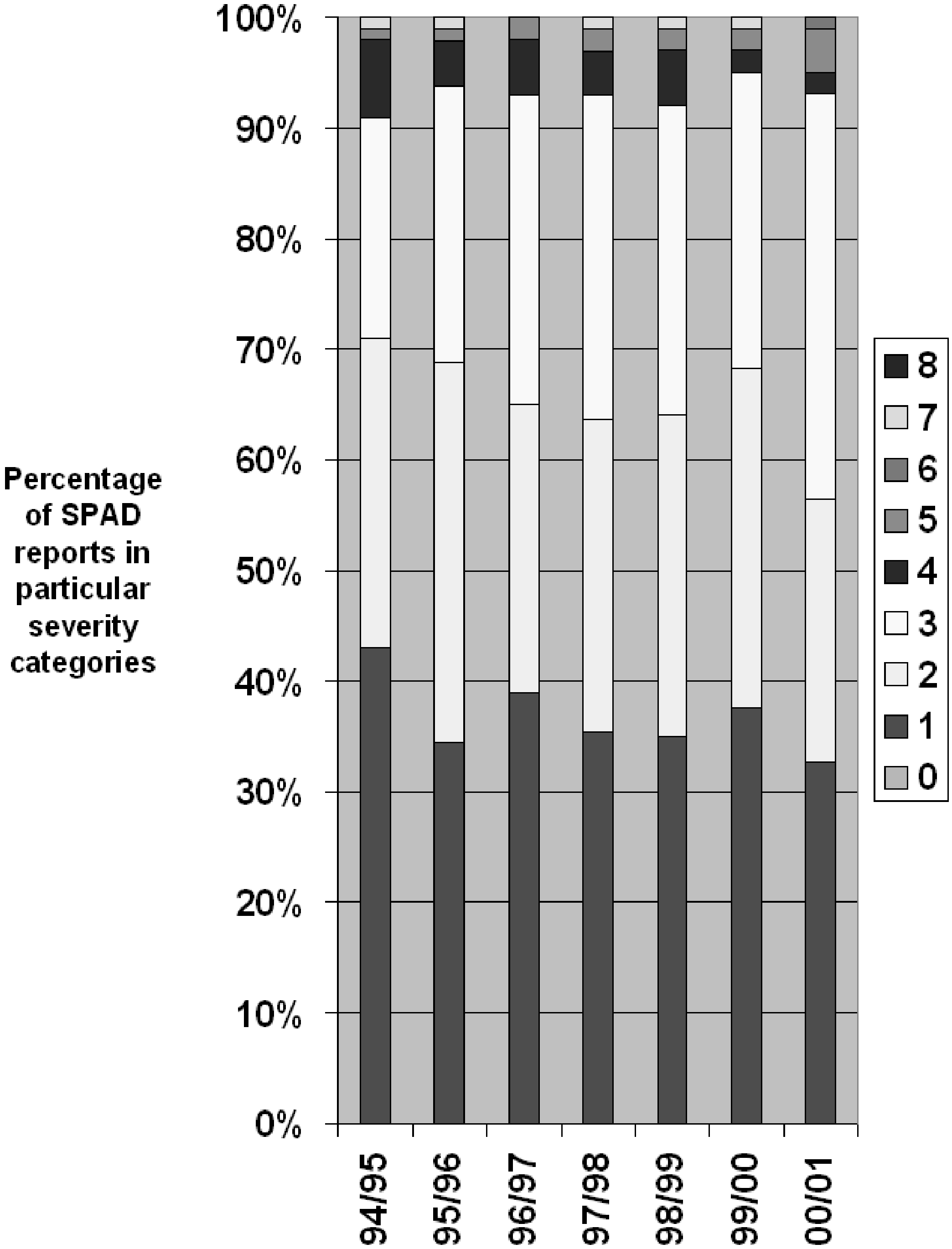

Figure 15.2: Static 'Column' Visualisation of SPAD Severity by Year

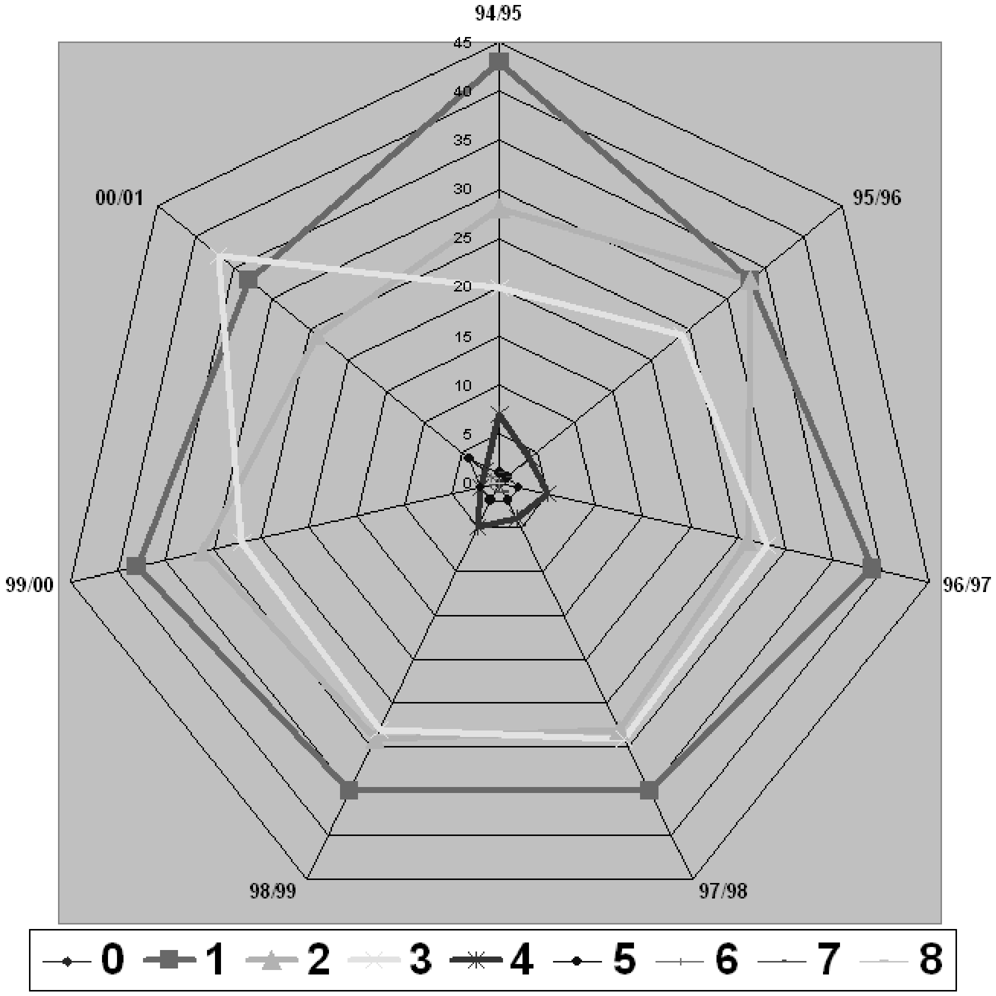

Figure 15.3: Static 'Radar' Visualisation of SPAD Severity by Year

recent changes to a database. There are further speed benefits . The use of wide area networks can rapidly disseminate monitoring information across many different contributory organisations and stake holders. For example, UK Railway Safety's Safety Management Information System (SMIS) provides on-line updates about the 250 safety events that are reported across the network each day. Over 50 organisations and 300 registered users can access updates about these incidents as they are entered into the system.

Electronic visualisation tools also offer a number of further benefits. Safety managers often find it difficult to produce new visualisations that might better help them explore alternative hypotheses about the causes and consequences of adverse events. If standard graphs and bar charts do not satisfy their requirements then managers must negotiate with their statisticians to request changes in the standard presentation formats. In contrast, many reporting systems now provide managers with direct access to incident data through the relational database technology described in Chapter 13.5. These systems help safety managers to pose interactive requests for data using the Structured Query Language (SQL) and its derivatives. Alternatively, user interface facilities can be provided to simplify query composition by supporting a more limited range of retrieval tasks. In either case, it is possible to use these information requests to drive graphical visualisations of statistics that characterise the information within the relational database. For example, Railway Safety's SMIS integrates with another PC-based reporting package (Crystal Reports). This enables users to customise the extraction of information from the safety management system to drive a large number of alternate visualisations.

As we have seen, however, a number of problems limit the utility of relational systems. It can be difficult for many safety professionals to master the syntax and semantics of SQL queries. For example, the following command for extracts SPAD data from an SQL database described by Speirs and Johnson [753]. It selects incidents where the overshoot is between 250 and 1,000 yards and the severity level was between 4 and 8 occurring between 18:00 and 24:00. The syntax has been simplified from that used in the full implementation of the reporting system!

```
select * from Incidents as i
      where Overshoot between 250 AND 1000
      AND ProvisionalSeverity between 4 and 8
      AND Time between 180000 AND 240000
```

The problems that managers have in interpreting relational commands has profound implications; 'inaccurate' queries can trigger recommendations that might not otherwise have been made. The use of pre-canned queries developed by support staff can be equally problematic. Previous studies have shown that safety managers often misunderstand the information requests that others develop for them [749]. Further criticisms can be raised because many existing electronic visualisations are simple extensions of paper-based representations. They are dominated by bar charts, scatter plots and conventional graphs. This is a missed-opportunity given the more advanced visualisation facilities that have been developed in other application domains. A further problem is that standard bar charts and graphs cannot easily be used to represent the complex 'multi-dimensional' data that is gathered about adverse events. For instance, Table 15.15 provides a brief excerpt from the SPAD data that is routinely disseminated by the HSE. It is unclear how any single visualisation might be used to capture this heterogeneous information. If such a system could be developed then users might be enabled to explore different attributes of the data. For example, it might be possible to explore whether a particular zone suffered from incidents at a particular level of severity at a similar time of the day or week. Alternatively, it might be used to determine whether the consequences of SPADs in some zones were effectively mitigated by defensive driving, leading to lower severity incidents with limited overshoots.

Previous paragraphs have criticised many incident reporting systems because they have failed to exploit the benefits offered by recent advances in information visualisation. A vast range of computer-based visualisations have been developed to help users extract statistical information from medical and other forms of scientific data [527]. They have also been used to visualise document collections and navigate the many thousands of potential results from web-based search [707]. This work has had little impact on the current generation of relational databases that support most incident reporting systems [754]. Speirs [753] has, therefore, begun to transfer more recent visualisation technology to support the presentation of UK SPAD data based on the subset of SMIS, presented in Table 15.15. The intention is not that these visualisations would replace other forms of numerical data analysis. In contrast, the aim is to provide a decision support tool that will enable safety managers and regulators to perform more interactive forms of search. Such visualisations are intended to help users 'discover' new properties of incident data that will enable them to monitor both changes in the underlying failures and the performance of the reporting system itself.

The design of the prototype visualisation was based around requirements derived from discussions with staff involved in the original, UK CIRAS confidential incident reporting system. These discussions emphasised the importance of juxtaposition within any visualisation for rail incident data. For example, it is important to compare the incidence of multiple and single SPADs within the same region. This information can be used to identify existing 'hot spots' where repeated SPADs have occurred in the past. This information can then be used to anticipate future problems where single incidents might, over time, become multiple SPADs. It is important to emphasise that one of the aims in implementing the SPAD visualisation tool was to identify the requirements for future interfaces to incident data. Before building such a prototype, it was difficult to provide railway

Date (2000)	Time	Signal No.	Location	Train Operating Company	Zone	Dist. Passed (Yards)	No. of SPADs at signal	SPAD severity	No. of SPADs by driver	HMRI Action Level
1/11	0339	B248	Narroways Junct.	EWS	GWZ	2377	8	7	2	3
1/11	1910	LD32	Liskeard	Wales & West	GWZ	12	2	3	1	2
1/11	1310	TT83	Pye Bridge	EWS	MZ	30	1	2	2	2
1/11	0430	T626	Horsham Road Xing	AMEC	SZ	200	1	3	1	3
2/11	1058	BW6	Bottesford West	Central Trains Ltd	MZ	200	1	2	1	3
4/11	1015	RETB	Halesworth	Anglia	EAZ	50	1	2	1	1
5/11	1705	St And X	Bristol Temple Meads-P5	Wales & West	GWZ	85	3	3	1	2
5/11	0030	LR534	Hathern	EWS	MZ	200	1	2	2	2
7/11	1035	HP20	Harringay Park	Silverlink	EAZ	1056	3	3	3	3
7/11	1240	E118	Taunton	Wales & West	GWZ	25	3	1	2	2
8/11	0620	R838	Newbury	Thames Trains	GWZ	3	3	1	3	3

Table 15.15: Subset of SPAD Incidents for November 2000 [355]

staff with an accurate impression of what was, and what was not possible, given current technology. The prototype has since been evaluated with support from UK Railway Safety and this process has helped to provide more detailed requirements for future versions of the visualisation. Some of the more surprising results from this consultation process are discussed in later paragraphs.

Figure 15.4 provides a screen shot from an initial version of the visualisation tool. The top left of the screen presents a map of the UK, each dot on the map represents the location of a SPAD. Colour coding can be used to distinguish multiple SPADS. The user can select each of these icons to obtain a range of more detailed information about the signal location. Figure 15.5 shows how this can include a time-line of previous events relating to the placement of the signal and also any previous SPADs. It can also include photographic information as well as plans and 3D models of the signal location.

The bottom of the screen in Figure 15.4 provides access to the more detailed information that the system holds about each SPAD. This corresponds to an extended form of the information presented in Table 15.15. By default, every SPAD incident is represented by a single dot and by a row in the panel at the bottom of the screen in Figure 15.4. However, the exact number of dots and rows will change in response to user selections. These are formed using the panel on the top right of the display. Figure 15.6 provides a more detailed image of these input widgets.

The visualisation tool uses a technique known as 'dynamic querying' [6] to avoid the problems associated with forming and interpreting the results of SQL statements. This technique enables users to extract information from a data set by directly manipulating common interface widgets such as sliders, lists and radio buttons. In Figure 15.6, the user can select either end of the Overshoot, Severity Category, Time of Day and Date sliders. For example, if they select the left hand icon on the

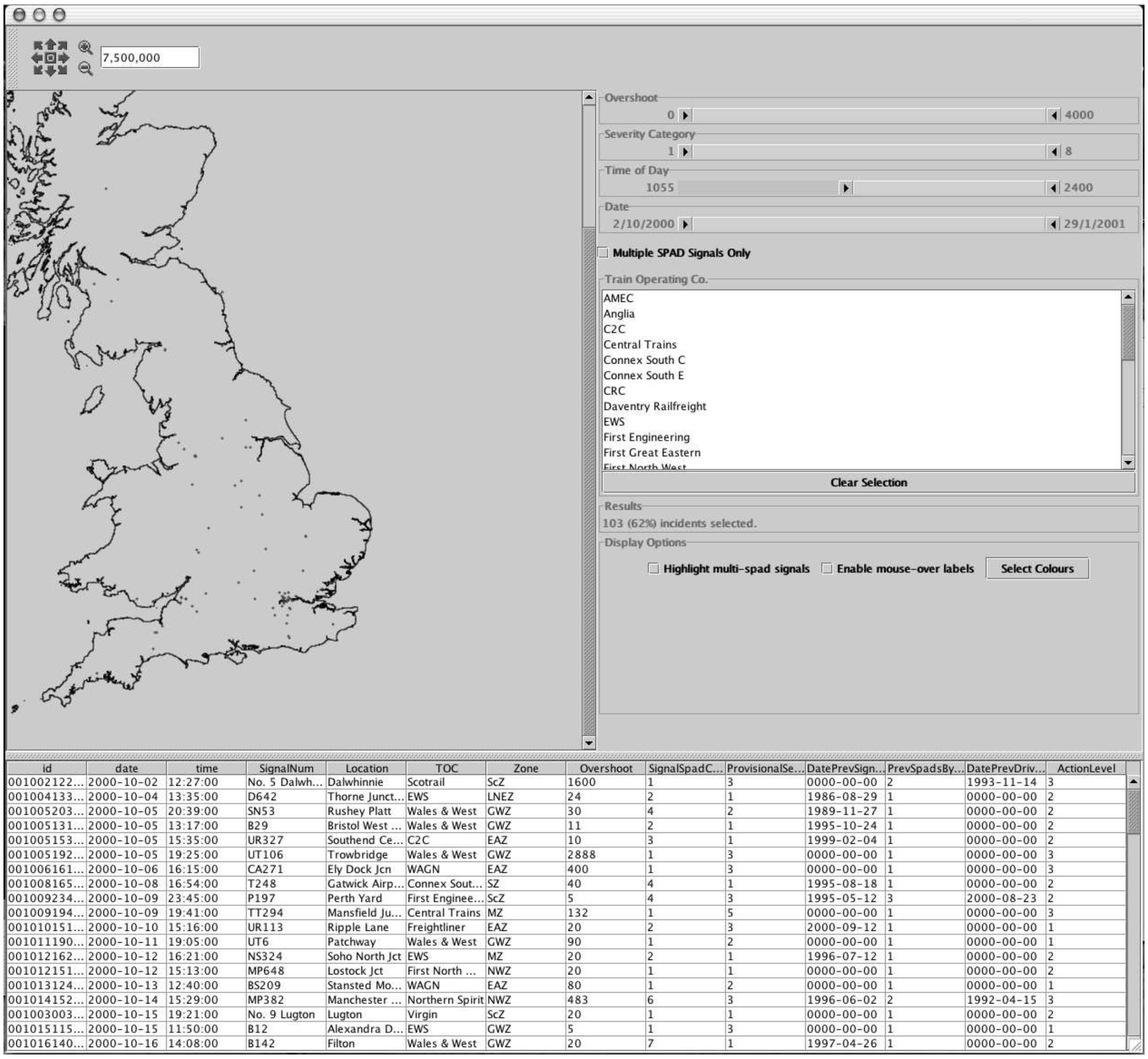

Figure 15.4: Computer-Based Visualisation of SPAD Data

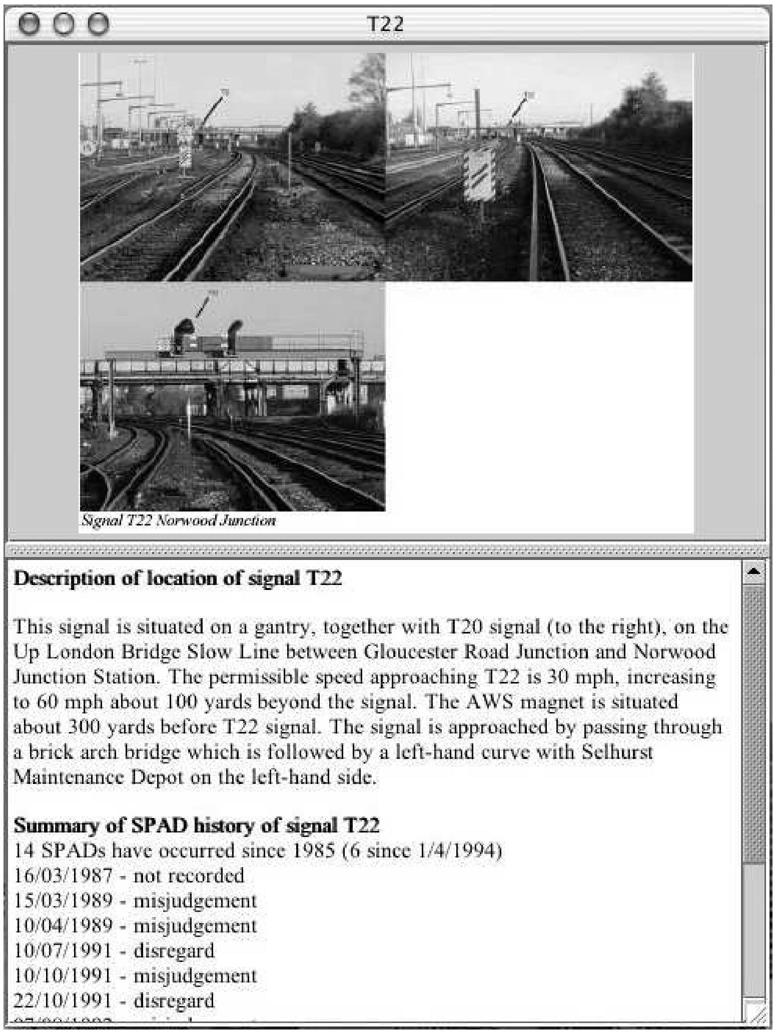

Figure 15.5: Signal Detail in SPAD Visualisation

Overshoot slider then they can alter the minimum overshoot distance for any SPAD displayed in the map view on the left. All SPAD incidents with an overshoot that is less than the value displayed on the slider will not be shown as dots on the map. Similarly, if the user selects the right end of the Severity Category slider and moved it from 8 to 3 then only SPAD incidents assessed to be in categories 1, 2 and 3 would be displayed. This form of interaction is known as 'dynamic querying' because it is, typically, built on top of a more conventional database. Each time the user makes a selection and changes the position of the slider, a new query is automatically generated by the visualisation tool and evaluated by the underlying database. It is important to stress, however, that the user is only aware of the interface components shown in Figures 15.4 and 15.6. They need not consider the underlying complexities of relational implementations. The lower portion of Figure 15.6 represents a choice or list widget. The user can filter their query to only display SPADs associated with particular train operating companies. By combining this approach with the slider mechanisms, it is possible to identify the most severe incidents involving particular operators during particular times of the day.

As mentioned, the number of dots shown on the map view in Figure 15.4 is updated in response to each query made by the user. As might be expected, the number of dots shown will be greatly reduced if the sliders are altered so that only the most severe incidents are considered. Conversely,

Figure 15.6: Dynamic Queries in the Visualisation of SPAD Incidents

the number of dots will increase greatly if the sliders are then adjusted to consider lower severity incidents. This use of dots on the map view is based on a number of previous visualisations in other domains, including epidemiology, where it is important to monitor the location of particular types of incident [753].

The visualisation, described above, created a number of practical difficulties when applied to incident reporting. For example, many incidents are clustered within a particular geographical location. In epidemiology this can correspond to particular out-breaks of a disease. In our railway case study, these 'hot spots' often corresponded with complex network characteristics and poorly sited signals. This resulted in a large number of dots representing SPADs within a small set of locations on the map. Unfortunately, this representation created problems because users had to 'zoom' in to gain the more detailed map view that was necessary to distinguish between different SPADs on signals that were close together. This was a particular problem given the high density of commuter rail operations in the South East of England and especially around London. Fortunately, other visualisation research can be used to identify potential solutions to this problem. For instance, Figure 15.7 shows how current versions of the prototype exploit a form of Fekete and Plaisant's eccentric labelling technique [248]. As the user moves their mouse over a region in which more than one SPAD has occurred the system will automatically 'pop up' an associated label with the location name of the incident. The user can then select each individual label to gain more detailed information about one of several incidents within the same area.

As mentioned, there are relatively few examples of more advanced visualisation techniques being applied to the monitoring of incident reporting systems. In consequence, the selection of appropriate techniques remains an area of active research. For example, the need to use eccentric labelling emerged as the visualisation prototype was expanded to support larger quantities of SPAD data. This illustrates the way in which the selection of appropriate techniques is driven by the problems that are created by the application domain. Given the lack of research in this area, it is unsurprising

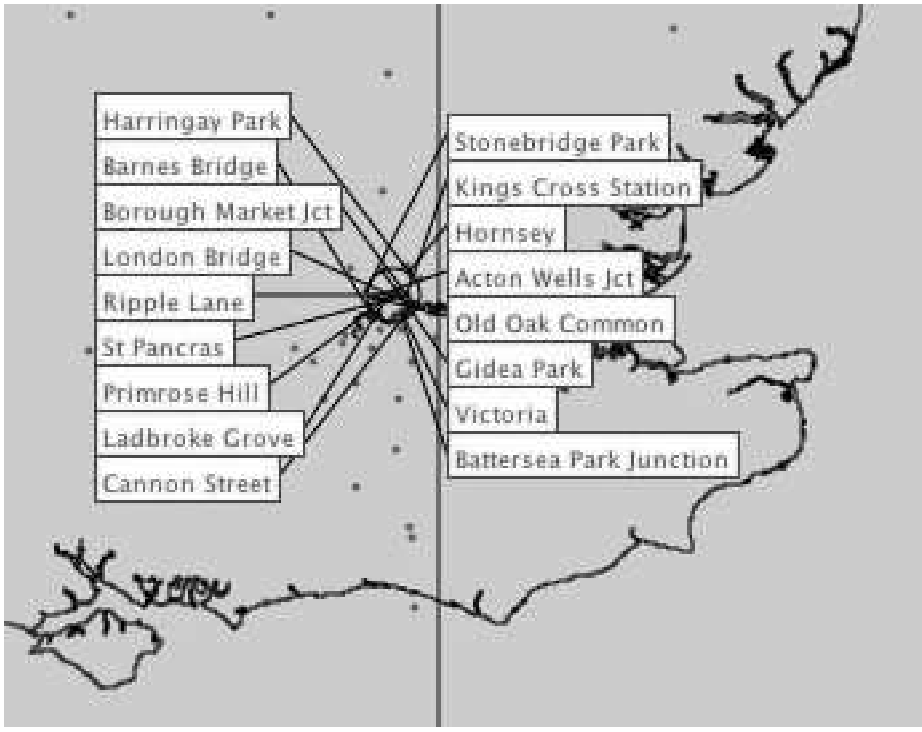

Figure 15.7: Eccentric Labelling in the Visualisation of SPAD Incidents

that the initial attempts described in this chapter should also produce some techniques that failed to support the needs of their intended users. For example, initial discussions with representatives from UK Railway Safety indicated that their safety managers would prefer to integrate the sliders and lists of the dynamic querying technique with more conventional graphs and charts. During 'walk-through' demonstrations, several of their senior managers argued that they were already familiar with the geographical distribution of events and more would be gained by integrating standard forms of statistical representation with the more innovative aspects of dynamic querying. Current research is investigating whether these views are shared by Train Operating Companys. The intention is to extend the initial prototype so that users can alter the image in the top left of the display in Figure 15.4.

Hamming argues that 'the purpose of computing is insight and not numbers' [306]. This section has argued that computer-based visualisations can be used to increase the insights that might otherwise be gained from monitoring statistical summaries of incident reporting metrics. We have not, however, presented any direct evidence to demonstrate the validity of these claims. Previous sections have considered some of the problems that arise from attempts to obtain such evidence. For instance, ethical problems restrict opportunities to introduce new information systems in 'live' reporting schemes if new visualisations can potential hide important information about adverse events. Further problems arise from the novelty of many computer-based visualisations. Users can express strong subjective satisfaction during the initial use of these systems simply because they represent a departure from existing techniques. This initial approval does not, however, imply that they will continue to be satisfied with the system over a longer period of time. The following section, therefore, describes how experimental techniques can be used to gather evidence about the utility of such visualisation tools and about incident reporting schemes in general.

15.4.7 Experimental Studies

The work on SPAD visualisation provides a useful case study of the problems that can arise in assessing or evaluating the meta-level effectiveness of incident monitoring tools. Some of these problems stem from the nature of such information systems. Computer users often find it difficult to express their requirements and needs to a designer. The field of requirements engineering within computer science has developed numerous techniques to address this problem [459]. As has been explained in the previous section, prototype implementations can provide potential users with a better idea of what is possible using existing technology. There is a danger, however, that considerable resources will be invested before safety managers and regulators confirm that the implemented system provides few benefits beyond those of existing systems! Participatory design techniques have been developed to address this problem. End users are represented in design teams and provide detailed guidance on a daily basis. The introduction of this end-user feedback can reduce the likelihood that a final implementation will fail to meet the needs of its potential users. Unfortunately, a number of problems limit the use of such validation techniques for the monitoring systems that support incident reporting schemes. In particular, most systems will only ever be used by a handful of domain experts, safety managers and regulators. These individuals usually play important safety management roles and, hence, their time is both a valuable and scarce resource. In consequence, it is often impossible to secure the level of commitment implied by participatory techniques.

The difficulty of obtaining access to the individuals who are involved in monitoring incident reporting systems creates further problems. As mentioned, key personnel seldom have the time that is necessary to conduct prolonged evaluations of prototype systems. This makes it difficult to perform the longitudinal studies that combat the biases created by the introduction of novel technology and by the Hawthorne effect, mentioned in Chapter 4.3. Long-term studies also often imply that monitoring systems will run alongside existing applications. This duplication avoids the ethical problems of experimenting with a 'live system'. It also creates additional managerial complexity and considerable expense.

Fortunately, a range of low cost evaluation techniques can be used to address the problems of gaining access to key staff. Some of these methods minimise the direct participation of end-users in the early stages of design. For instance, heuristic evaluations provide designers with rules of thumb that can be used to make inferences about potential usability problems in the final implementation of a computer-based monitoring system. There are both general heuristics [636] and heuristics that support the evaluation of particular systems. For instance, Shneiderman [740] provides a set of criteria that can be used to assess interactive visualisations:

- *overview*: the visualisation must provide the user with a high-level overview of the information that is being presented;
- *zoom*: the visualisation must enable the user to move from the higher-level overview to focus on specific items of interest;
- *filter*: the visualisation must enable the user to filter out related items of information that are not relevant to their current information requirements;
- *details on demand*: it should be possible to select a particular item or group of items and obtain additional information about the selected items;
- *relate*: it should be possible to use the visualisation to view relationships between items of information;
- *history*: it should be possible to undo the effects of a selection or filtering operation, it should also be possible to redo previous operations when necessary.
- *extract*: it should be possible to extract sub-collections from the mass of initial data so that queries can be posed on subsets of the data.

Speirs has used these criteria in the evaluation of the SPAD prototype that was described in the previous section. For example, the dynamic querying facilities provide means of rapidly undoing a filter operation by returning the sliders to their original position. However, early implementation did not enable users to *extract* and save sub-sets of the data for later analysis. It can be argued that this is unnecessary given the ease with which queries can be composed from the simple interface widgets. This argument illustrates both the strengths and weaknesses of heuristic evaluation. These guidelines are a starting point for the evaluation of a potential system. They are also subjective and open to a wide range of interpretations. This makes it likely that different designers may apply the same criteria in a range of different ways. In consequence, there is often a need to perform direct user evaluations to resolve the different claims that can be made about particular heuristics.

Fortunately, a range of low cost techniques can be used to provide user feedback but without the expense associated with longer-term evaluations. For instance, cooperative evaluations and 'think aloud' techniques require that analysts set potential users a series of tasks with a prototype implementation [877]. The focus on accomplishing specific tasks, such as using the system to create a particular statistical summary, avoids the need to ask leading questions, such as 'hwta do you think?'. The responses to such prompts are, typically, impossible to interpret as they can be biased by a range of subjective factors including a concern not to appear ignorant about information technology. By focusing on whether or not the user can perform particular tasks, the intention is to determine whether the system will meet their needs. Subjective satisfaction can be assessed as part of subsequent validation activities. The participants in a cooperative evaluation then attempt to perform the tasks as best they can. If there is any confusion or they do not know what actions to perform then they should express their uncertainty by 'thinking aloud'. The designer can then either provide appropriate feedback or allow the potential user additional time to work on the problem. In either case, there break-downs are noted and become the focus for subsequent re-design. As mentioned, these techniques are relatively low cost because they do not require the prolonged participation of senior staff and domain experts. Feedback is provided in a relatively informal setting and evaluations can be conducted in a relatively short period of time.

Unfortunately, a number of further limitations affect the use of cooperative evaluations to monitor the effectiveness of monitoring tools. In particular, it can be difficult to generalise beyond the results provided for particular users operating a particular version of a prototype implementation. The fact that one safety manager successfully accomplished a task does not imply that others will achieve similar successes. Further problems arise when validation tasks must derive comparative measures for the relative utility and usability of rival designs. It can be difficult to use the introspections derived from 'think alouds' to show that one design is better than another. It is for this reason that several groups have attempted to perform experimental evaluations of monitoring tools [754]. These techniques rely upon established methodologies, often derived from experimental psychology [687]. They rely upon the analysts' ability to distinguish the change, or independent variable, that is linked to a change in the measurement of a dependent variable. For example, an experiment might be conducted to establish the hypothesis that a new monitoring system reduces the time taken to perform a range of key tasks. The dependent variable would be the two versions of the monitoring system. The independent, measured variable would be the timings taken over the range of tasks. The method chosen to conduct such an experiment must be carefully considered to ensure that only the dependent variable is altered between the two conditions. For example, if experienced staff were used with the new system and novice staff with the old then one might expect better performance with the newer system than with the old.

Speirs has used this approach to evaluate the utility of his visualisation tool as a means of monitoring incident reporting data [754]. This study illustrates the complexity of conducting experiments within this area:

1. it is unclear how to develop appropriate tasks for users to perform during the evaluation Traditionally, many validation activities have focussed on well-specified tasks such as finding the answers to particular questions. For instance, a user of the SPAD visualiser might be asked to find out the distance of overshoot associated with an incident in a particular location. This style of experimental evaluation cannot easily be used to support the validation of incident reporting systems. Safety managers are seldom faced with such specific information requests. Greater challenges come from the need to identify patterns and trends within a complex data set. This implies that any evaluation will have to focus on less directed forms of interaction;

2. fatigue can complicate experimental evaluation. The longer that a user interacts with a tool then the more tired they can become, especially if they are being asked to use new and unfamiliar systems. One consequence of this is that if an evaluation compares two systems then many users' will perform less well with the second system that they meet. Fatigue becomes more of an influence than any potential design improvements. Counter-balancing can be used to address this problem. This implies that half of the user group will meet the old system first and the other half will meet the new version first;

3. learning effects with new systems complicate experimental evaluation. Counter-balancing cannot reduce the problems created by learning effects. For instance, it may take some time before novice users of a new system can build up the same level of expertise that they have achieved with the existing implementation. This effect can often be observed when users' performance with a new system slowly improves as they attempt successive tasks. This problem can be addressed by designing a series of training tasks to ensure that users are happy with both of the systems that are being compared. The users' performance with these tasks is not measured as part of the experimental evaluation and users are only encouraged to progress once they feel happy that they are able to perform them unassisted;

4. learning effects with particular tasks complicate experimental evaluation. Learning effects not only complicate the comparison of alternative monitoring systems. They also effect the tasks and the data sets that are presented to the user during the evaluation. For example, if users were asked to perform the same tasks with two different systems then one might anticipate that a knowledge of their previous answers might influence subsequent responses. It is for this reason that many experimental evaluations provide different tasks for each system. This creates further problems because some questions might be 'easier' than others. Hence, it becomes necessary to ensure that counter-balancing also considers the questions that are associated with each system. Table 15.16 provides an example of the complexity that this can create. It also emphasises the point that was made earlier, experimental evaluations can require access to relatively large numbers of users in order to exploit such techniques;

Group 1	Old System	Questions A	Questions B	New system	Questions A	Questions B
Group 2	New System	Questions A	Questions B	Old system	Questions A	Questions B
Group 3	Old System	Questions B	Questions A	New system	Questions B	Questions A
Group 4	New System	Questions B	Questions A	Old system	Questions B	Questions A

Table 15.16: Counter-balancing Systems and Tasks

5. learning effects with particular data sets complicate experimental evaluation. It can be difficult to use counter-balancing as a means of reducing learning effects associated with particular data sets. For example, users may get a better 'feel' for the information that they are being asked to monitor as they interact with it over time. This effect could be addressed by partitioning the incident dataset into two or more sections and then introducing additional user groups to experience each data set with one of the experimental conditions, shown in Table 15.16. Such techniques undermine the *validity* of the evaluation. It is likely that any final implementation would have to support the entire available data set. Frequently the need to control experimental conditions can lead evaluators to impose unrealistic constraints on the use of a system. This creates problems because the results of any study are, therefore, indicative of the system that was evaluated and not of the system as it might operate in an eventual working environment;

6. what do we measure? It can be difficult to identify the measures that might be used to assess the effectiveness of a monitoring system. It is seldom the case that fine grained timings would have a significant impact upon many safety managers. It may make little different whether it takes 6 or 8 minutes on average to identify a particular trend so long as that trend can be identified. Further problems arise because measurement criteria are, typically,

multi-dimensional. In consequence, it can make little sense to compare very different systems using the same criteria. For instance, relational databases can provide relatively fast access to information in response to specific queries. In contrast, the SPAD visualisation tool supports less directed forms of search. It would be unsurprising to find that each tool performed less well when assessed against criteria that were not used to guide their initial development.

This is a partial list, many further problems complicate the design of experimental evaluations. The interested reader is directed to the summary in [127, 305]. In contrast this paragraph provides an example of the experimental techniques that were used in the validation of the initial SPAD prototype.

It was initially only possible to recruit seven subjects. This illustrates the way in which access constraints can frustrate attempts to employ the counter-balancing mentioned above. As this was a preliminary evaluation, Spiers compared the performance of the SPAD visualisation tool with the statistical presentation of SPAD data that is presently hosted on the HMRI's web cite [356]. Further studies are currently comparing the visualisation tool more directly with the existing presentation of data from Railway Safety's extended Safety Management Information System (SMIS), mentioned in previous sections. The initial comparison focussed on a number of relatively open ended tasks. Users were asked to rate their agreement with a number of statements on a scale from one to seven [754]. As we shall see, this complicated the analysis of the results from the study. It was, however, intended to provide a measure of the certainty that participants felt in the conclusions that they reached using a particular visualisation. For the purposes of counterbalancing, the questions were assigned to one of two groups and the position of questions within each group was varied. From this it follows that each participant answered all of the questions, however, the order that they answered them was varied as was the system used to generate their answer:

- Set A:
  1. Events at multiple SPAD signals constitute around 40% of all SPADs.
  2. Most SPADs have a severity category of 3 or over.
  3. Most SPADs involve an overshoot of less than 200 yards.
  4. The number of incidents at multiple SPAD signals is increasing from month to month.
  5. Incidents involving a signal that had previously been passed at danger usually also involve a driver that has been involved in a previous SPAD.
- Set B:
  1. Railtrack Midland Zone (MZ) is the zone with the lowest number of SPADs.
  2. The number of incidents is relatively stable from month to month.
  3. SPADs are more common in the morning (24.00-12.00) than in the evening (12.00-24.00).
  4. The incident with the longest overshoot distance occurred in Manchester.
  5. No SPAD has occurred north of the incident at Perth Yard (signal P197).

Some questions can be answered directly from the information provided by a particular system. For instance, the map view of the visualisation can be used to directly identify whether or not a SPAD occurred beyond signal P197. Similarly, the spreadsheet view provided by the HMRI, illustrated by Table 15.15, can be directly used to identify whether or not the longest overshoot occurred near Manchester. Other questions involved a greater degree of analysis and interpretation. For example, there is no direct means of determining whether multiple SPADs formed a particular percentage over overall incidents from the data that was presented to each user in either system.

The degree of interpretation and analysis involved in some questions, together with the seven point scale, created problems in analysing the results of the evaluation. In order to do this, ideal responses were identified for each question. These 'solutions' were based on the HMRI's interpretation of the SPAD data. The users' performance with each system was measured in terms of the absolute (ie., non-negative) divergence of their score from this ideal value. For example, a user might assign the value 1 to show that they disagree with the statement 'most SPADs involve an overshoot of less than 200 yards'. If the HMRI indicated that most SPADs did involve overshoots of less than 200 yards then the ideal score would have been 7. The user would then be assigned the value 6 for their performance on this question to indicate variance from the ideal answer. The results from this evaluation are presented in Table 15.17. As can be seen, timings were not taken during the study and users were encouraged to take as much time as they liked.

Subject	SPAD Visualisation	HMRI Site
A	9	9
B	6	9
C	6	4
D	4	10
E	13	1
F	2	5
G	5	10
Total variance from 'ideal'	45	48

Table 15.17: Initial Results from Experimental Evaluation of the SPAD Browser

As can be seen, the results showed very little difference between the performance of the users with either visualisation. Perhaps more remarkable is the variation in individual performance. For example, subjects A and B did equally well with either monitoring system. In contrast, subjects D did much better with the SPAD visualisation while E showed less variance from the 'ideal' responses when using the more conventional spreadsheet format used by the HMRI. These results led Speirs to realise that users were exploiting the graphical visualisations and the tabular format for different purposes. The sliders of dynamic querying techniques were used to filter the data set while the tabular or spreadsheet view was used to rapidly scan for particular numeric values. This justified the introduction of tabular data into the bottom half of the display illustrated in Figure 15.4.

Subsequent work identified a number of limitations with this experiment. These limitations illustrate further problems with the experimental method as a means of validating incident monitoring tools. Cooperative evaluations were conducted with more senior staff from Railway Safety. They argued that the tasks were not particularly significant for the end-users of SPAD incident data. The location information provided by the map view simply helped to reinforce the correlation between SPADs and the density of railway operations within particular zones. In contrast, they advocated the geographical presentation of information about suicides as well as trespass and vandalism incidents. They argued that these incidents did not relate so directly to the flow of traffic but did possibly cluster around particular geographical regions, for instance with particular problems of social deprivation and unemployment [754].

The previous paragraphs have explained how ethical issues can prevent investigators from evaluating new reporting techniques on 'live' systems. Initial 'teething' troubles can lose safety-related information. Dissatisfaction with prototype tools and techniques might jeopardise the future success of an existing scheme. Laboratory techniques, typically, avoid these problems by examining the operation of a new system in under carefully controlled conditions with simulated tasks. Some attempts have, however, been made to conduct experimental comparisons with 'live' systems. These evaluations have abandoned some of the controls that are normally associated with laboratory assessments in order to increase the experimental validity of their results. Novel techniques are compared to existing approaches using real operators reporting actual incidents. This approach was used to evaluate initial versions of the ATSB's INDICATE program [46]. INDICATE provides company's with a framework for eliciting, documenting and monitoring safety-related incidents. It is also supported by a range of software tools that can be tailored to the individual requirements of participating organisations. The INDICATE program has also been extended to support organisations in the aviation, road, rail and maritime industries. Initially, however, the evaluation focussed on the aviation domain. An Australian regional airline agreed to use INDICATE in one of its operational bases while another section of the same organisation was used as a control group over an eight month trial period The length of the evaluation reflects the intention to recreate 'realistic' reporting behaviours during the study. Experiments that run over a couple of hours can often suffer from biases that create 'atypical' reporting behaviour. This style of evaluation has much in common with the Sentinel studies, described in previous paragraphs. Resources are focussed on an initial trial of a new reporting technique. There are some important differences. In particular, the use of a control group is directly taken from experimental evaluation techniques and helps to provide some basis for comparing the results obtained by the introduction of a new scheme. As we shall see, however, it can be difficult to make accurate comparisons between these two different groups.

Five evaluation criteria were used to determine whether INDICATE had a positive effect: airline safety culture; staff risk perception of safety hazards; willingness of staff to report safety hazards; action taken on identified safety hazards and staff comments about safety management within the airline. 48 safety reporting forms were submitted by the INDICATE group, 9 were submitted by the Non-INDICATE group. Analysts argued that this difference 'may be a direct result of an attitude change within the INDICATE base as a result of the safety program (e.g. a more positive attitude to reporting safety issues, increased staff confidence that safety problems would be addressed, more awareness amongst staff of operational hazards, and improved staff commitment to improving company safety)' [46]. A questionnaire was used to provide feedback about the other evaluation criteria. A reliability analysis was conducted to establish that the questions elicited consistent responses from individual participants. Such studies are importance because doubts can arise if the same person provides radically different answers to the same questions within a short period of time or if individuals in the same organisation express radically different views about the same topic. Under such circumstances, it may be more important to understand the differences within a group than between two different groups. After the trial, the INDICATE group showed 'a significant improvement in their mean safety culture score, while the Non-INDICATE base results showed a poorer safety culture'. Various statistical measures, including T-tests and ANOVAs, were used to support the finding. It was argued that there was a '99.9% certainty that the safety culture improvement, demonstrated in the INDICATE base, was due to the implementation of the safety program and not some chance factor' [46]. Such arguments are, however, complicated by the problems of conducting 'experimental' evaluations on reporting systems. At the start of the study, staff in both centres indicated that safety was managed 'in a positive manner' [46]. However, the INDICATE site achieved a slightly better initial score than the control group. This difference makes it difficult to interpret the results, cited above. Any subsequent change in attitudes or reporting behaviour can be explained in terms of the initial differences between the two groups. The initial score of the INDICATE group may reflect a pre-existing increase in awareness about safety issues. The subsequent improvement might, therefore, be part of this previous trend rather than a 'direct result' from introducing the new reporting programme.

There was a smaller difference between the two groups in the risk perception questions from the initial questionnaire. At the end of the evaluation period, however, the INDICATE group showed a significant decrease in the risks that it associated with particular hazards. The non-INDICATE group exhibited a much smaller reduction. These results were not expected. The analysts argued that they might have been due to chance factors that did not affect the other metrics. Alternatively, the reduction in the level of risk perceived by the INDICATE group might show that the program had provided staff with a clearer idea of the hazards facing their industry. The reduction in risk perceived by the non-INDICATE group was argued not to be statistically significant. Again, however, these arguments were complicated by the difficulty of identifying the direct influence of the INDICATE program.

It was argued that comments from staff in the INDICATE group revealed that there was: 'better provision of safety training to new staff; more management praise for safe working; better company feedback regarding safety performance; and an increased frequency of safety audits' [46]. These comments were also interpreted to show that staff in the INDICATE group were more confident in their safety management systems. Staff in the non-INDICATE group were 'generally more negative' about communication from management and the reporting of safety incidents. The qualitative assessment criteria were not, however, explicitly documented. This is important because it can be difficult to agree upon the best means of extracting such conclusions from the informal comments of individual workers. For instance, respondent validation can be used to ensure that staff within the INDICATE group agreed with the summary of their comments. Similarly, staff within the non-INDICATE group might have been asked to check the interpretation of their comments. This method can be difficult to apply when comparisons are made between two groups. Unless respondents have access to the comments of their colleagues, they cannot judge whether their responses were 'more negative' than those of another group. Alternatively, independent assessors can be used to summarise respondent comments without knowing the experimental context in which they were obtained. The form of 'blind' reviewing can, however, be difficult if analysts lack the information that is necessary to interpret particular comments. In either case, the key point is that some form of additional validation is often required to support the interpretation of qualitative responses.

A slight variation on the use of experimental techniques in 'real world' settings is to conduct limited studies to support the gradual introduction of a regional or national system. For example, in October 1998 the FRA awarded a 3-year contract to design, develop and test a Toll-Free Emergency Notification System (ENS). This was intended to centralise the reporting of problems at highway-rail intersections [242]. The ENS System was initially installed along limited areas of track within the State of Texas. Early in the project, a number of liability issues were identified and special legislation was required to authorise particular organisations to manage such a facility. The initial study was extended from Texas into Connecticut and then to areas of Pennsylvania. The Pennsylvania program modified the Texas system so that it could be operated in an existing 911 emergency centre. It began by supporting eight selected railroads but the longer-term objective was to 'continue refinement, based on operating experience with the demonstration system, so that a system suitable for statewide usage by short-line railroads is realised' [242].

The ENS project illustrates the way in which a Sentinel-style approach can be integrated into a form of iterative development. The ENS project also considered a number of issues that complicate the use of experimental techniques with 'live' reporting systems. For example, the sample had to broad enough to make it likely that incidents would be reported. It was equally important, however, that the size of the study was not so large that it overwhelmed the available resources. Signs had to be deployed at all public railway crossings along the eight chosen railroads. They also had to be deployed a private crossings that were considered active enough to create a significant risk of a potential incident. Farm field crossings that were used two or three times per year did not warrant a sign. This may have prevented the reporting of some incidents but also helped to focus the allocation of finite resources. The ENS evaluation illustrates further problems that complicate the use of experimental methods on 'live' systems. In particular, there was a concern that sufficient data should be available about the safety record of the existing system. Without this data, accurate comparisons could not have been made following the introduction of the new ENS application. Unfortunately, the under-reporting of adverse events makes it very difficult to obtain accurate data. There can also be a range of more prosaic problems. The ENS analysts also had to ensure that they obtained accurate information about the location of the existing crossings.

A similar approach to that adopted in the ENS study was also used to examine a range of techniques that were intended to reduce the use of train horns at railway crossings in the United States [237]. This, in turn, was intended to reduce environmental problems associated with the use of train horns to warn drivers and pedestrians of an approaching train. This study is interesting because it mixes elements of several different monitoring techniques. An experimental method is used in that the study attempted to control conditions around a number of road-rail crossings so that comparisons could be made between driver behaviour with different protection mechanisms. The techniques also borrowed from the observational approach mentioned in previous sections because video taping was used to record the behaviour of 'real' drivers as they approached the crossing. The results of this analysis not only provided insights into the effectiveness of safety measures that were intended to address previous incidents. The video analysis also provided information about the reliability of reporting systems because it helped to identify a range of 'near miss' incidents in which accidents were narrowly avoided but which would not otherwise have been notified to regulatory organisations or the operating companies.

Previous paragraphs have looked at the application of experimental techniques to support the meta-level monitoring of incident reporting systems. In other words, we have concentrated on the evaluation of innovative reporting techniques rather that on the performance of an individual reporting system. The same evaluation methods can also be applied more directly to anticipate whether the nature of incidents will change as a result of revised operating procedures. The study involved an agreement between Spokane County, the Washington State Utilities and Transportation Commission and the Burlington Northern Sante Fe Railroad. It was based around four phases. This helped to ensure that the behaviours, which were witnessed during the study period, provided accurate insights into longer-term driver performance. The first phase provided a 'control' or 'baseline period'. During this time there was neither a median barrier nor a whistle ban. The second phase of the experiment introduced a barrier but did not enforce a whistle ban. The third phase involved the introduction of median barriers and a whistle ban. Each of the first three phases lasted 115 days each. The final phase monitored driver behaviour for one week in each of the following three months and one week each quarter for the year after the original study. During this time, the median barriers remained in place together with the whistle ban.

The analysts focussed their attention on incidents in which vehicles and pedestrians continued to cross even though the crossing had been activated. An incident was also defined to have occurred if a pedestrian or vehicle collided with the gate or if they went around a gate after it had been activated. They also argued that most attention should be devoted to those incidents where there was a train present. Their study identified numerous cases in which the gate activated without a train being present, for instance through gate malfunction. Table 15.18 summarises incident frequencies for each of the four phases in the study. There was a sharp decline in the incident rate between phase 1 and 2. The percentage of gate activations in which there was an incident fell from 34% to 1.2% after the introduction of the barriers. There was little change after the introduction of the horn ban between phases 2 and 3. There was a relatively small increase in incidents during the final phase.

	Phase 1	Phase 2	Phase 3	Phase 4
No. of Gate Activations				
Train present	4,556	4,924	5,003	680
No train present	31	155	117	18
No. of Incidents				
Train present	1,565	61	66	14
No train present	419	5	9	7

Table 15.18: Number of Gate Activations and Incidents, With and Without a Train Present

This mix of experimental techniques in 'real' contexts with more observational techniques illustrates a range of further practical problems. Residents reported frequent soundings of the train horn even during phases three and four when the bans was in place. A dosimeter placed at this location expressly to monitor possible whistle soundings revealed 231 noise spikes in Phase 3 alone. As in previous studies, it was also important to ensure that accurate comparisons could be made between the results for each of the different conditions. In this case, it was necessary to ensure that there were no differences in the number of times that the crossing gates were activated between the different phases of the project. In Phase 1, there were 4,587 activations in all at an average of 39.9 per day. In Phases 2 and 3, the averages were slightly higher 44.4 and 44.8 respectively. The shorter periods of Phase 4, yielded an average of 38.8. Unfortunately, the relative similarity in the number of gate activations did not characterise the statistics for car/automotive traffic. The average annual daily traffic for the first phase was estimated at 3,831 cars per day. This was significantly higher than the 1,918 cars in phase 2 and the 1,991 in phase 3. This statistic was not calculated for the more limited observations in period 4.

These problems illustrate the difficulty of accounting for the many different variables that might influence the incident rate between a number of 'experimental conditions'. It is difficult to know how to interpret the results from such a study. It might be argued that the relatively high number of automobile journeys observed in the first period invalidates the use of the incident data for the control of baseline phase. It is important to emphasise that those involved in conducting the experiment would almost certainly have been unaware of the potential imbalance as they conducted the study. It would only have become apparent during the detailed analysis of the video data. Given such objections, analysis would have been forced to remove the barriers and the whistle ban to recreate the conditions under which the first phase was conducted. However, there would then be the problem of ensuring compliance with staff who had already become familiar with the ban. The ethical implications of removing barriers must also be considered. The FDA analysts, therefore, simply present the data and identify the potential flaw in the evaluation. This pragmatic stance enables individuals to form their own judgements about the validity of the trial.

In passing, it is important to note that this study identified a number of important insights about driver behaviour at railway crossings. Many of these insights could not be obtained through more conventional reporting systems, For instance, many cars and pedestrians crossed after the gates had been activated but in situations where it was clear that no train was present. For instance, the seven incidents in phase four of table 15.18 represent seven cars crossing one after the other when the gates had failed. The ways in which such events might have led to more serious violations were not considered as part of the study. However, such observations illustrate the way in which monitoring techniques can produce a broad range of insights into the limitations of reporting systems. Until the study took place, nobody reported these ‘successful violations’ in which users were forced to cross an activated crossing in order to cope with a gate failure.

There are further examples of experimental techniques being used to support the monitoring of incident reporting systems. For instance, the FRA’s Volpe Center often performs empirical studies to validate the recommendations that are made in the aftermath of adverse events. Their Transportation Technology Center (TTC) was specifically designed to test all types of rail equipment and vehicles in a variety of weather and terrain conditions. Their facilities allow a limited form of replication in which collisions can be recreated. During these tests certain factors are kept constant while others are systematically varied to support particular hypotheses, in the manner described in previous sections. For instance, identical rolling stock can be operated at different speeds along the same track. These full-scale crash simulation are increasingly used to validate computer models, such as the finite element simulations that are needed to describe the nonlinear properties of material behaviour in rail collisions. These models take the place of destructive testing [852]. There is a particular sense in which these activities help to validate the products of the FRA’s reporting systems. The Volpe center conducts crash testing and similar experimental studies to help determine which adverse events pose the greatest concern for the future safety of the railways. The results of their work should identify those conditions that are known to pose the greatest threat from accident reports. They may also validate the potential for future accidents by identifying problems that have not yet resulted in injuries or fatalities. The results of these experiments help to determine the potential consequences of events described in ‘near miss’ incident reports. This integration of experimental testing and the reporting of adverse events illustrates the way in which many safety-critical organisations are addressing the problems created by causal asymmetries. In the aftermath of an adverse event, we often cannot be sure which of several possible causal ‘paths’ actually led to an observed outcome. Full-scale simulations and computer models can be used to recreate the circumstances leading to an incident or accident. This increases confidence that a causal explanation can account for the observed event .

15.5 Summary

This chapter has identified several different forms of monitoring that can be used to assess the utility of an incident reporting system. For example, accident rates can be used to assess the impact that the recommendations from these schemes have upon the safety of application processes. In particular, analysts can look for evidence that training and operational practices have been directly improved by the insights from reporting systems. They can also demonstrate the effectiveness of such schemes by assessing the contribution that they make to the calculation of future failure rates and consequence assessments during subsequent risk assessments. These validation activities focus directly on the impact that a reporting system has upon the safety of underlying applications.

In contrast to these outcome measures, other forms of assessment focus on the processes that are used to derive recommendations from incident reports. The success of a reporting system can be assessed in terms of the number of reports that are elicited from staff and management. Alternatively, analysts might focus on the efficiency of a scheme by monitoring the costs associated with analysing each report. In particular, process metrics can be devised to calculate the percentage of contributions that result in particular safety measures being introduced into the 'target' organisation.

Incident reporting systems offer a number of additional benefits beyond the specific recommendations that are made in the aftermath of an adverse event. In particular, participation in these schemes can also have a more general effect in raising awareness about safety-related issues. A third set of monitoring techniques, therefore, focus less on outcomes or on process metrics and, instead, focus on acceptance measures. Analysts can assess the effectiveness of a reporting system in terms of the contribution that they make to a wider 'safety culture'. They can also determine whether potential contributors are satisfied that a system is both confidential and unbiased.

Numerous problems complicate the monitoring of incident reporting systems using these three different approaches. Outcome measures are difficult to gather and hard to interpret. In many industries, there are significant concerns about the accuracy of accident statistics. Incident reporting systems also, often, form part of a wider range of measures that are intended to improve the safety of application processes. This makes it difficult to demonstrate that any changes in accident rates are directly due to the introduction of a reporting scheme. At a lower level, recommendations to change employee training and operational practices in the aftermath of an adverse event can be identified as specific benefits from a reporting system. Often, however, these changes can introduce new failure modes. The impact of such changes can, therefore, only be assessed over a relatively prolonged timescale.

Similar problems frustrate the use of process metrics to support the monitoring of incident reporting systems. For example, a paradox of many reporting systems is that the introduction of such schemes will typically increase rather than reduce the number of reported failures. This has led some safety managers to focus on the criticality of reported incidents rather than on submission rates as a metric to measure the impact of such schemes. If a reporting system is having a positive effect then the number of submissions should remain high but the potential consequences of any adverse events should decline as necessary interventions are made in response to previous reports. This approach is difficult to apply effectively because it requires a relatively sophisticated means of measuring the potential consequences of an adverse event. Investigators frequently show considerable disagreement over the potential outcome from the same adverse event. Other process measures suffer from similar problems of subjective interpretation. Attempts to monitor the performance of individual investigators are complicated by the need to determine a 'gold standard' for causal analysis and the generation of recommendations. Attempts to measure the proportion of submissions that lead to safety interventions can be rendered ineffective when a high number of relatively unimportant recommendations might be considered to have a greater impact than a smaller number of far-reaching innovations.

It can also be difficult to use acceptance measures to support the monitoring of reporting systems. It is far from easy to agree upon a set of metrics that can provide adequate feedback about the impact of such schemes on the safety culture or climate within heterogenous organisations. Similarly, different individuals within the same working group can have radically different opinions about the probity or 'trustworthiness' of a reporting system. In such circumstances, monitoring systems can help to identify the diversity of views but provide little help in encouraging greater confidence.

The second half of this chapter has reviewed a range of monitoring techniques that are intended to address some of the problems that complicate the use of outcome, process and acceptance metrics. For example, public hearings, focus groups, working parties and standing committees all provide means of monitoring incident reporting systems. Focus groups can help regulators and safety managers to assess attitudes towards these schemes within sections of the workforce. In contrast, standing committees provide a more sustained framework that can be used to coordinate a range of monitoring activities over longer periods of time.

Subsequent sections considered ways in which incident sampling can be used to focus monitoring activities. One approach is to select a random sample of reports so that they can be followed as they are analysed and recommendations are implemented. Alternatively, monitoring activities can be focussed on the response to particular types of report. Resources might be allocated to see if there are any problems in the way in which reports from particular user groups are handled. For instance, the Ladbroke Grove enquiry focussed attention on the way in which SPAD reports were handled within the UK rail industry. Such sampling techniques create considerable methodological problems. If employees become aware that attention is being paid to particular incidents then it can become more likely that these adverse events will be reported and analysed in greater detail than might otherwise be the case. This illustrates another example of the Hawthorne effect, introduced in Chapter 4.3. Individuals will alter their patterns of behaviour if they know that their actions are being observed.

The Hawthorne effect that complicates the interpretation of insights gained from incident samples is actively exploited in Sentinel monitoring systems. This approach deliberately sensitises particular groups or organisations so that they are more likely to report adverse events. These groups are given additional training and resources that could not be provided throughout a mass reporting system. The incidents that are reported by the participants in a Sentinel system can then be compared to reporting patterns throughout an industry. This approach can provide insights into the under-reporting problems that affect many schemes. It can also be used to conduct limited evaluations of additional training materials that might eventually be distributed more widely throughout a reporting scheme. As mentioned, Sentinel systems do not overcome the problems of the Hawthorne effect. Participants are, typically, aware that their reporting behaviour is being monitored. Sentinel systems also rely upon sensitising employees so that they are more aware of the adverse events that should be reported. This can have the paradoxical effect of reducing the likelihood of these events. Individuals are less likely to be involved in particular incidents if they have already been warned about them in their Sentinel training.

Observational techniques avoid many of the biases that can be introduced through incident sampling and Sentinel schemes. These approaches rely upon investigators conducting detailed studies of the reporting behaviours in particular working groups. Such techniques often reject the suppositions of theoretical frameworks that can bias the subsequent interpretation of any observations. The intention is to let particular theories about reporting behaviour develop in a bottom-up way. Theories are grounded in the observations rather than confirmed by them. Unfortunately, a number of methodological problems complicate the application of these approaches. In particular, the relatively low frequence of adverse events implies that observers may have to spend many months studying a particular group of employees before they witness any 'reporting behaviour'. Observational techniques are resource intensive. It can take several hours to analyse a single hour of video tape . There are also problems with generalising from the insights gained by monitoring a particular work group. One team's reporting behaviour can provide relatively few insights into the reporting behaviour of their colleagues in different companys, different geographical regions etc [305]. The rich qualitative data that is derived from these studies cannot easily support the statistical analyses that are often required by governmental and regulatory organisations. For instance, many industries are required to monitor their reporting systems as part of a wider assessment of their safety management schemes. In such cases, normalised accident statistics are used to provide direct outcome metrics. Unfortunately, it can be difficult to identify appropriate normalising factors that reflect the diversity of many industries. For instance, the safety of a rail operator might be assessed in terms of the number of fatalities per passenger mile. Such a normalisation could not easily be applied to freight operations. It might also provide few insights for networks in which marshalling operations accounted for the majority of fatalities.

Previous sections have reviewed a range of additional factors that complicate the use of statistical techniques to support the monitoring of incident reporting systems. In particular, it can be difficult for managers, regulators and the general public to correctly interpret the values that are derived from more sophisticated forms of statistical analysis. It is for this reason that many incident reporting systems are being supported by computer-based monitoring systems. Visualisation tools enable managers to explore and exploit a range of statistical information about their schemes. We have illustrated this argument by describing a SPAD tool that provides railway regulators and operating companies with 'dynamic querying' techniques.

The chapter concluded by arguing that a form of meta-level monitoring must be conducted to ensure that visualisation systems and similar monitoring tools actually support their intended users. In the case of the SPAD tool, mentioned above, the use of a geographical information system to provide feedback on the location of incidents was perceived to be less important that the provision of statistical information in the form of more conventional charts and graphs. Laboratory-based evaluations provide, arguably, the best developed set of methods for validating such meta-level monitoring tools. These techniques rest upon experimental situations in which it is possible to distinguish the change, or independent variable, that is linked to a measure of the dependent variable. Unfortunately, the ecological validity of laboratory-based experiments is often questioned. In other words, the controls that are necessary to isolate the dependent variable often creates situations that are a long way from those that characterise the working lives of incident investigators. For instance, they may be required to use monitoring tools in purpose-built evaluation labs. These, typically, exclude sources of distraction including their colleagues, telephones broken printers etc. Fortunately many of these problems are being addressed by hybrid techniques that combine elements of observational and experimental validation. For instance, the FRA have used video cameras to record driver behaviour at rail-road intersections. These observational techniques have been used to monitor the safety improvements provided by a range of barriers and warnings. These different measures are varied in a systematic way that borrows from the use of control groups and different experimental conditions in the laboratory studies mentioned above. Although such techniques overcome some of the objections that have been made towards both laboratory-based studies and observational techniques, they also introduce a range of ethical concerns. There is a danger that lives will be lost in control groups or experimental conditions that are deliberately deprived of certain safety features.