Cognition, Technology & Work (2002) 4:1–8

Cognition
Technology &
Work

The Creation of a New Minor Event Coding System

B. Wallace[1], A. Ross[1], J. B. Davies[1], L. Wright[1] and M. White[2]

[1]*University of Strathclyde, CIRAS, Glasgow, UK;* [2]*University of Exeter, Exeter, UK*

Abstract: The present study began with an assessment of the reliability and usefulness of an existing minor event coding system in a British 'high-consequence' industry. It was discovered that despite the fact that the system produced replicable data, when tested in a reliability trial the causal inferences it was producing failed to meet the normal criteria for statistical reliability. It was therefore felt necessary to create a new model of the human factors component of action in this industry, from which a model of human factors *error* in the same industry could be inferred. A set of codes (to facilitate statistical analysis) were deduced from this last, which were then tested in a new reliability trial. The results from this trial were very encouraging, and after a six-month pilot study in which it demonstrated its usefulness as a trend and patterning tool, the system is now being phased in within this industry.

Keywords: Error analysis; Error detection; Human factors; 'Near-miss reporting'

1. INTRODUCTION

The extent to which 'minor error events' function as indicators or predictors of more major error events has become increasingly evident over the last 20 years. The inquiries into such disasters as the Challenger space shuttle, Hillsborough, and the *Herald of Free Enterprise* found that all these events had been preceded by relevant 'near misses' (Lucas 1991). It has therefore become apparent that efficient and reliable 'near-miss'/minor event reporting and analysis systems may be important tools in terms of accident prevention. It is generally agreed that by describing areas and functions in which minor events take place (usually in the form of 'codes' to facilitate analysis), such systems will be capable of identifying 'weak spots' from which more major events may develop. This issue is of particular importance to the nuclear industry, which faces a unique level of public scrutiny as to its safety record.

It is obvious that in order for meaningful data to be derived from such systems, they must be shown to be reliable in terms of consistent analyses (or 'coding') of events. That is to say, not only must different users of the system be able to infer similar causal analyses from the same event, but the same coder must also be able to code the same event in the same way after a period of time has elapsed.

The current study was initiated by the UK nuclear industry (both British Energy and BNFL/Magnox Generation) in order to assess its own minor event causal analysis system with regard to precisely this issue of reliability. It was prompted by a growing suspicion within the industry that the existing system was not functioning in an optimal fashion, and that changes might have to be made to its structure. This turned out to be the case, and it became obvious that the assumption that the 'old' system was producing the reliable data necessary for the accurate targeting of resources in terms of safety issues was invalid.

As discussed below, it was also mooted that any new system would function more effectively if it was a 'total system': that is, if the information was gathered in such a way that it would 'feed' the database in an optimal fashion. Given that this was the case, interest was also expressed in studying the efficiency of accident investigation procedures in the nuclear industry. It was vitally important, therefore, that adequate reliability data could be produced, for not only would effective work in improving safety at the 'macro' level, but accurate safety improvements at the 'micro' level would also be made easier if any new system were shown to be reliable. If two people, approaching the same problem, produce vastly different estimates of what went wrong and why, it will be logically impossible to target resources in the correct area, and the safety of an organisation will not, therefore, improve. As Groeneweg writes: 'Reliability is a necessary condition for validity' (Groeneweg 1994, p. 217), validity being defined as positive evidence that a system improves safety.

The researchers involved were therefore given an unusual opportunity, in that permission was now given by the industry to create a new 'model' of an event involving human factors, and to develop a new causal analysis system from this model. This new system was then tested for reliability in the same way as the previous system had been.

The present paper therefore consists of three main sections. Firstly it describes the testing and analysis of the existing system, with the implications of these results for other minor event reporting systems. It then goes on to describe the new error event model which has been developed and the new coding system which has been derived from this model. Finally it describes the testing and analysis of the 'new' coding system, and the implications of this analysis for other high-consequence industries.

1.1. The 'Old' System

The system as it stood at the beginning of this study (which had no official name, but which was usually referred to as the 'Observed Cause and Root Cause Analysis System', or, colloquially, the 'obs and root' system) processed data in two main stages. Firstly reports were submitted by members of the workforce through initial report forms in which the substantive information was described by the reportee in his/her own words (Stage One). Secondly, the forms were read by experienced 'event investigators', who assigned 'codes' on the basis of the salient causal information they perceived to be present as well as information emerging from their subsequent investigations (Stage Two).

The coding system itself was highly proliferated and contained 196 codes in total. These were, however, distributed between 17 'supercategories' with such titles as 'Management Methods' and 'Design Configuration and Analysis', which were themselves broken down into between 5 and 20 more minor categories. It should be noted that there were differences in procedure and terminology between plants, and that there were also a wide variety of terminologies for the minor event systems themselves. However, the basic category coding system remained the same across all the plants. All events were considered for remedial and preventive action at an appropriate level in each plant.

1.2. The Interrater Reliability Trial

It is clear that the usefulness of *any* data analysis system of this type, that is, based on an individual's subjective judgement, is limited by the reliability with which the system can be employed. It is also clear that the salient aspect of a minor event coding system in terms of major event prediction is the reliability with which individuals can turn reports into codes. In terms of the system described above, there needed in particular to be confidence that different judges would code the same event in the same way in the majority of cases.

Since unreliable coding is, therefore, a major source of error variance, an interrater reliability trial was conducted to estimate the extent of such variance.

In this trial, 28 previously coded events were randomly selected from the existing files. Three experienced coders from within the industry were asked to read each event report and to assign causal codes in the usual way. However, the event reports (which had already been processed using the system) had also been given an 'original' coding in a 'real-world' situation. In order to give a further test of reliability, this 'original' coding was used as a 'fourth' coding of the event and compared with the other three.

1.3. Data Analysis

As three coders took part and the 'original' coding was also used, there were six possible paired comparisons for which reliability could be assessed. The comparisons relevant to the issue of coding are given in Table 1.

Table 1. Interrater reliability trial of 'old' system: reliability for each pair of coders (4 coders)

Pair	1	2	3	4	5	6	Average
Index of concordance	42%	45%	41%	42%	41%	39%	42%

Reliability was calculated from the Index of Concordance (that is, the Number of Disagreements divided by the Number of Agreements + Disagreements) (described in Martin and Bateson 1993).

Although there are no *generally* agreed criteria for acceptance of data from the Index of Concordance (Caro et al 1979), the researchers followed the National Center on Child Abuse and Neglect's position in a similar study, in positing 75% as an acceptable level of agreement: clearly the system under discussion here did not meet this criterion (NCCAN 1998).

1.4. Study of Codes over Time

As well as the reliability trial described above, data from the 'old' system were also analysed over a 22-month time period, on the presupposition that these data would be valid in terms of aiding the allocation of resources. It should be noted that, despite the lack of reliability as regards to coding, the minor event database showed strikingly similar distributions of events over categories during this time period. That is, it was discovered that clusters of root causes were continually assigned for certain categories of codes not only across time (i.e., any given time period chosen for analysis would show the same distribution of codes), but

also across plants. Moreover, within the reliability trial itself, these same patterns of root causes appeared. *But*, given the unreliability of the coding system this *could not* be the result of reliability of analysis, but instead indicates that the system was producing replicable and *apparently* reliable data from its own structure and demand characteristics. This finding may be of help in the creation of other minor event coding systems, in that it indicates that even if replicable patterns of data are produced from these systems this does not necessarily mean that the coding procedures themselves are therefore reliable. Needless to say, if the coded data are unreliable, then any predictions made in terms of predicting major events from minor events from these data will also be unreliable.

1.5. Changes in the Codes

Given the results detailed above, it was necessary to study the underlying logic of the coding system in an attempt to ascertain the reason for the low reliability.

The most obvious reason for the lack of reliability was that the codes had evolved over time, rather than having been positioned within a pre-existing logical structure. Consequently no logical hierarchy was apparent in the database, leading to a situation where some coding categories were overused while others were not used at all. Moreover, it became apparent that there was considerable overlap between categories, in that many of them were not mutually exclusive. This was partly as a result of the large number of codes that had been created: the confusion this caused, of course, also led to reduced reliability.

Possibly more important, however, was the fact that the existing system did not produce sufficient human factors data. Causes were generally classified as (a) technical and plant factors, and (b) 'work practice' human factors at the level of the man–machine interface. While it was understandable that this would be the case (most of the coders coming from an engineering background) it nevertheless fitted in with the findings of others as to the difficulty of obtaining human factors data in large organisations (Wilpert and Fahlbruch 1996; Reason 1990). It was clear, therefore, that any new system would have to be able to identify human factors when they were present, and that, as well as the obvious 'work practice' human factors, it would have to be able to identify what Reason (1990) calls 'resident pathogens' or 'latent failures'; i.e., human factors issues deep within the organisation, which, when (and only when) triggered by events at the man–machine interface, would help to cause an error event.

It should be noted in passing that all these issues deal with 'Stage Two' of the system, that is, with the way in which the reports were turned into codes. However, it was also suspected that problems were arising from variability in the way in which the reports themselves were constructed. The reliability trial was partly formulated in such a way as to deal with this issue.

2. THE HUMAN FACTORS MODEL

It was decided that only by creating a model of action in terms of human factors could a similar model (in terms of human *error* during this action) be inferred. The human factors action model is shown in Fig. 1.

The human error (during an action) model inferred from this model is shown in Fig. 2. It should be noticed that this human error model is the inverse of the human factors model, and that it was from this human error model that a coding hierarchy was inferred.

It will be noticed that the fundamental logical structure

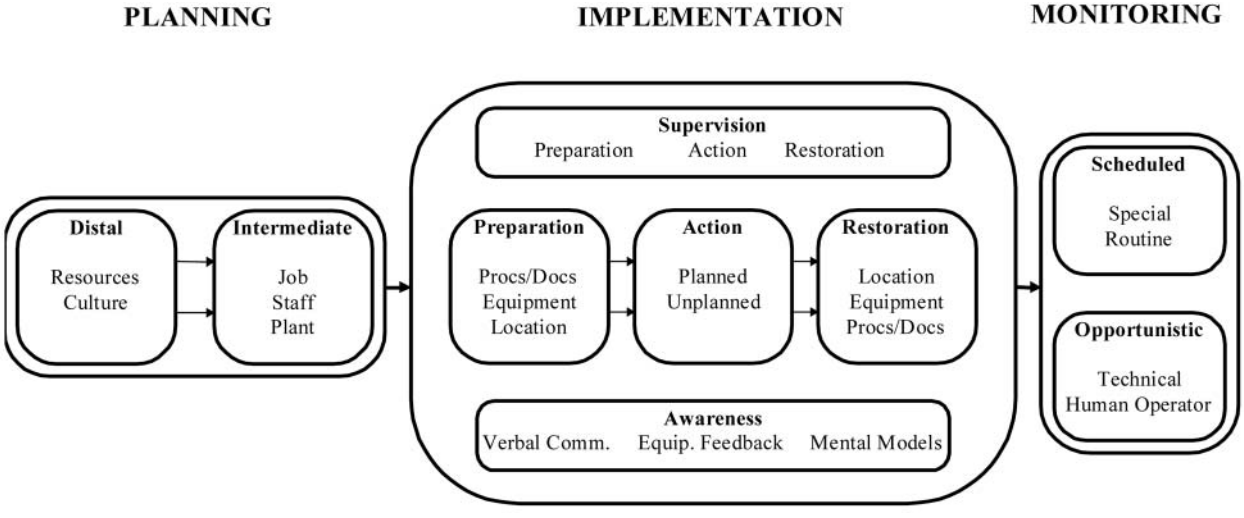

Fig. 1. Human factors in a high-consequence industry.

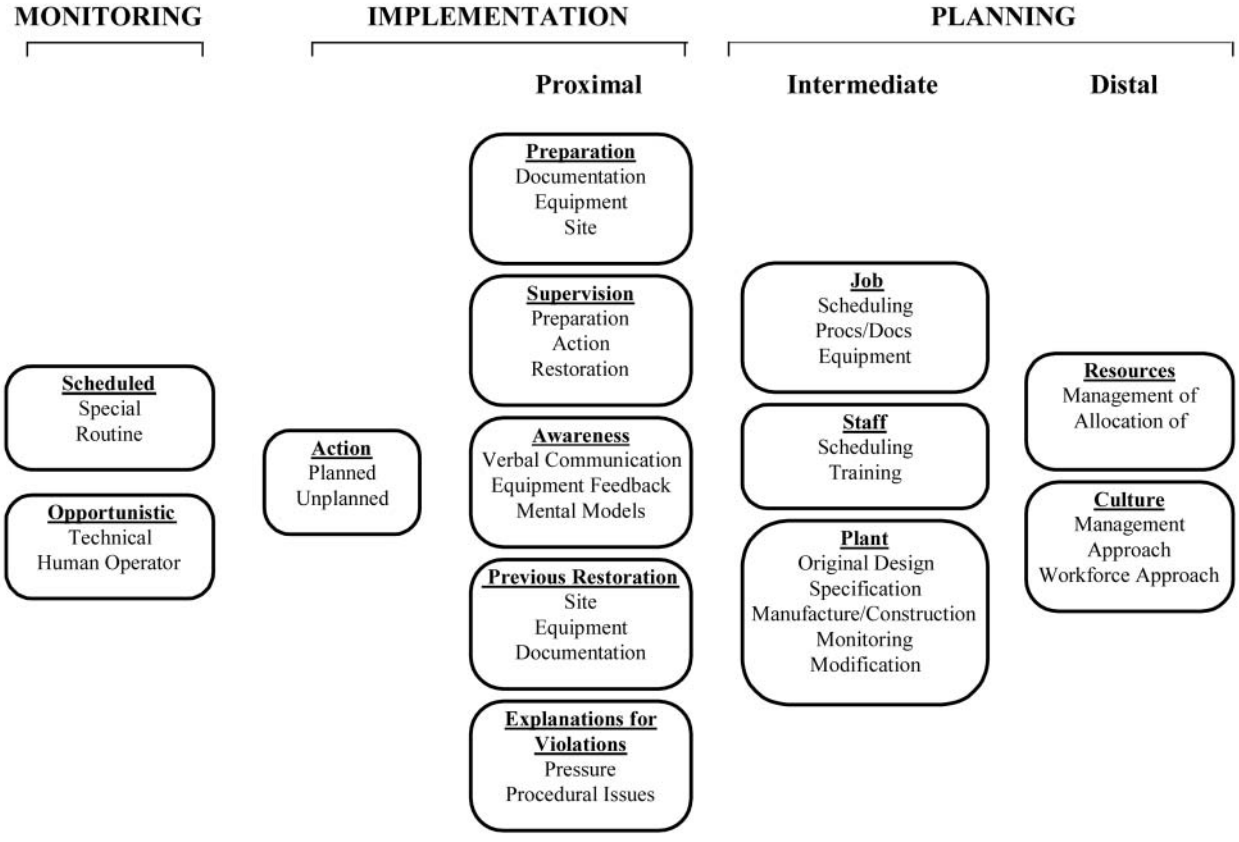

Fig. 2. Human error in a high-consequence industry.

of this model is concerned with the temporal aspect of the activity concerned, which is broken down into the basic structure of 'beginning, middle and end'. For example, in terms of the *gestalt* of all human factors in any particular plant, the model breaks down into Planning (beginning), Implementation (middle) and Monitoring (end: i.e., the safety monitoring of the event *after* it has taken place). In the same way, within Implementation itself it is posited that every action can be broken down into Preparation, the Action itself and Restoration (again, corresponding to the 'beginning, middle and end' of an action). This has the advantage of fitting in with 'common-sense' concepts of the structure of a minor event, while at the same time creating a rigid logical structure from which a system may be inferred.

It should be noted that the Proximal and Action layers in these models are roughly analogous to Reason's 'Active Errors' (i.e., 'associated with the performance of the 'front-line' operators of a complex system') and the Intermediate and Distal layers are roughly equivalent to Reason's 'Latent Errors' (which are conceptualised as being 'remote in both time and space from the direct control interface' (Reason 1990, p. 173)). However, in line with the 'beginning, middle and end' distinction posited above, the 'Action' layer is the only section of the model which describes the event *as it happened* (all the other layers deal with events and causes either anteceding or preceding the event). Therefore, the Action layer functions as a *description* of the event, whereas the Proximal, Intermediate and Distal layers analyse the event in terms of both latent and active *causes*.

It can also be seen that the 'Proximal' 'Intermediate' and 'Distal' layers of the model are differentiated by their degree of 'abstraction' from the Action layer. Thus, Proximal causes are those which impact directly on the action itself and broadly relate to the implementation or 'man–machine interface' stage of the event. Intermediate causes can broadly be conceived as organisational: although they vary in proximity to the event, most *precede* the proximal causes of the event. Distal causes (for example, resource issues) are also positioned as such because of common-sense notions of the temporal precedence of cause in relation to effect, although it should be noted that not all the Distal 'codes' fit into this paradigm. The 'Workforce Approach' codes, for example, are positioned at the Distal level. However, they are placed at this level because they are further 'removed from the event' (that is, more 'abstract') than causes at the Proximal and Intermediate levels, not because they are necessarily temporally precedent to them.

It would, of course, also have been possible to adapt one of the existing human factors models (such as that used in the development of the TRIPOD model (the system developed by the Centre for Safety Research at Leiden University and James Reason at Manchester University and used mainly in the oil industry) instead of developing a new model as described above. However this was not done for three reasons.

Firstly, nuclear industry management made clear that they wished for a system that was nuclear industry specific and that would deal with issues unique to organisations with a high public safety profile.

Secondly, it was hoped that by creating a model by considering organisational/managerial features to begin with, and only after they had been fully described to move on to 'man–machine interface' issues, a more accurate description of these features could be obtained. Therefore the first stage in the creation of the model was to show all organisational/human factors (*not* human error factors), whereas all other models begin with the error process itself. Only after a 'human factors model' had been created was an error model inferred from it.

Thirdly, a desire was expressed for a system from which a rigorous and reasonably proliferated set of codes could be produced that could be structured hierarchically in such a way that they could cope with 'real-world' natural language incident reports, and this was found to be difficult to achieve with existing systems. (Note: the Tripod model uses questionnaire data gathered from at least 40 personnel members of the industry in question as a validation tool. Moreover, TRIPOD presupposes an 'open' system whereby managers can gain access to information not contained in reports: neither of these options are available to managers in the British nuclear industry. The TRIPOD model is not, therefore, designed to deal with highly specific problems faced here (Groeneweg 1994)).

3. THE HUMAN FACTORS CODES

The next stage of the process of developing the new system was to develop a set of codes corresponding to each 'box' in the model. The set of codes for each of the boxes in Fig. 2 are written as 'logic diagrams': with codes *becoming more specific from left to right*. For example, the Supervision 'box' (which can be seen in the models above at the Proximal layer) is used to create the codes shown in Fig. 3.

The 'new' system takes over the 'Two-Stage' approach to data analysis from the old system: i.e., data are still collated into reports, which are then coded. The difference lies in the logic of the coding system. The coder is now required to code the reports by constructing 'sentences' using a hierarchical coding frame. Thus, in the example shown, code '**xyz**' implies that a problem has been identified with *supervision* (**x****) during *action* (**xy***) in terms of *feedback*

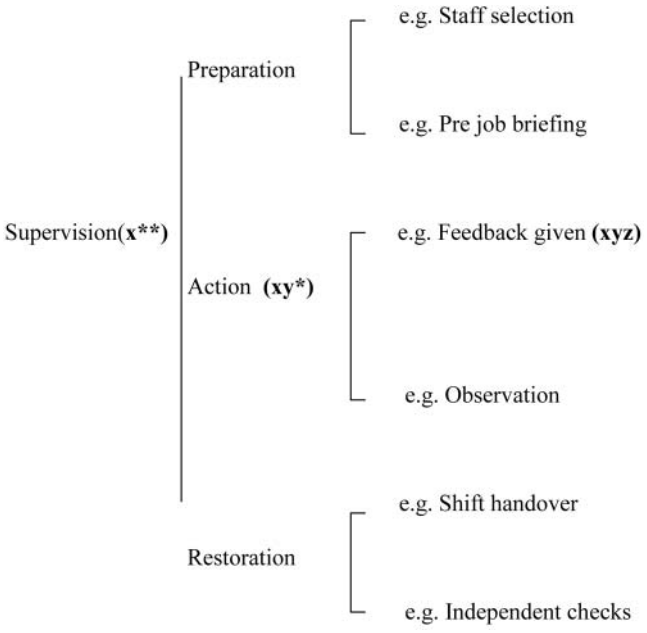

Fig. 3. The supervision codes of the 'new' system

given (**xyz**). The fact of the codes being organised in this way allows the coder to stop at any of the above stages depending on how specific the information available to him/her might be (i.e., to refrain from proceeding through the 'logic tree' to the right-hand side of the page unless he/she has the information to do so).

The codes were intended to be all-inclusive and mutually exclusive. In the example above, if a supervision problem during the action was identified (i.e., an **xy** code was given, not an **xyz**) then we assume that error *must* have concerned the feedback given or the supervisor's observation of the action. The logical structure of the system is predicated on the idea that these are the only two codes it would be possible to pick, had more information been available.

In the same way, the codes were designed in such a way that there was the minimum possible 'overlap'. In a minor event, any given cause of the event could be *either* a Supervision error during the Action or a Supervision error during the Preparation – not both.

4. THE RELIABILITY TRIAL

It was now necessary to test the 'new' coding system in order to test its reliability. Due to the increased numbers of coders available for the trial of the 'new' system two identical trials were conducted with a total of nine experienced coders taking part. However, these trials still used the basic template of the trial of the initial ('obs and root') minor event coding system (as discussed above). The

main difference was that, instead of one daylong trial, there were two day-length trials, with five coders taking part in the first trial, and four coders taking part in the second. As before, the purpose of the trials was to assess the interrater reliability of the new system. On each day the trial had two sections. The first section of the trial was modelled closely on the initial reliability trial of the 'old' system. Coders were given 12 randomly selected event reports (these were, of course, not the same reports as used in the previous trial), and were asked to code them using the proposed system. Reliability was calculated, as before, from the degree of concordance between coders. This section of the trial was designed to see whether coders could reliably obtain human factors data from the event reports in a way expressible in terms of the new system.

The second section of the interrater reliability trial dealt with 'recoding'. This was the term adopted to refer to the process by which the salient parts of the event report text (in terms of what were perceived to be the 'most important' causes of the error event) were selected and coded by one (randomly chosen) coder and then passed to another, who assigned the fragment of text a code in the usual way. Reliability was calculated by the degree of concordance between the first coding and the second coding of these selected aspects of the report.

The purpose of this procedure was twofold. Firstly, in a 'real-world' situation, coders might discuss the events they had to code, and some consensus would often be expected to arise in terms of what were the key aspects of the event in terms of error causation. In a sense, therefore, this meant that the recodes were a more accurate indication of the results which might be achieved if the system were implemented than the 'non-recoded' data from the first section of the trial.

Secondly, this procedure would give an indication as to possible sources of error in mapping the codes to the reports. The coder assigning the 'recodes' would not have to 'interpret' the event report document in order to isolate the important causative elements, but, having had this decision made for him, would simply be able to assign codes to the fragments of text provided. In other words, if reliability increased during the recodes, this would strongly suggest that any remaining difficulties coders encountered in creating causal chains expressible in terms of the proposed system would arise from difficulties in interpreting the reports, and not, therefore, from the logical structure of the codes themselves.

In terms of method, only one coder was common to both the trial of the 'old' system and the 'new' system (the project liaison representative for the industry): since the trials were held 18 months apart it was hoped this would negate the 'test–retest' effect. Moreover, coders had been trained 'in-house' by the industry itself in the use of the 'old' system, whereas training in the use of the 'new' system was carried out by researchers at the University of Strathclyde and took roughly three hours. Apart from these factors, experimental conditions were identical in both trials.

5. ASSESSING RELIABILITY

After initial scoring using the same method as in the first reliability trial, reliability was again calculated from the Index of Concordance (Number of Agreements divided by the Number of Agreements + Disagreements). The interrater results from day one are shown in Table 2, and those from day two are shown in Table 3.

The interrater reliability for the first day was calculated to be 56%, and that for the second day was calculated as 66%. Overall interrater reliability was therefore 61%.

'Recoding' data was also calculated (see Tables 4 and 5).

For the first day this was calculated at 72%. On the second day, it was calculated at 89%. Overall 'recoding' reliability was therefore calculated to be 81%.

It can be seen, therefore, that reliability rose from 42% to 61% without recoding, and from 61% to 81% with recoding. It is therefore posited that the rise from 42% ('old' system) to 61% ('new' system) is due to the more logical structure of the codes, and the rise from 61% to 81% is a consequence of the removal of variability arising from features of the reports (lack of clarity, etc.) rather than from the use of the system.

Table 2. Interrater reliability trial of 'new' system: day one (5 coders)

Pair	1	2	3	4	5	6	7	8	9	10	Average
Index of concordance	56%	51%	59.5%	57%	56%	68.5%	49.5%	59%	46.5%	61%	56%

Table 3. Interrater reliability trial of 'new' system: day two (4 coders)

Pair	1	2	3	4	5	6	Average
Index of concordance	64.5%	59.5%	69.5%	63%	64%	77%	66%

Table 4. Recoding data from 'new' system: day one (5 coders)

Pair	1	2	3	4	5	Average
Index of concordance	78%	86%	64%	58%	72%	72%

Table 5. Recoding data from 'new' system: day two (4 coders)

Pair	1	2	3	4	Average
Index of concordance	90%	100%	79%	85%	89%

6. DISCUSSION AND CONCLUSIONS

Given the importance of reliability and event modelling in terms of the creation of minor event coding systems, it is striking how little reference there is to these issues in the literature. Indeed, in an analysis of existing similar systems, Wagenaar and van der Schrier (1997) note that at the time of writing only TRIPOD had collected any reliability data *at all*. To the best of our knowledge the situation has not improved since 1997. As Wagenaar and van der Schrier write: 'There is no real excuse for the lack of reliability testing since it is not difficult to measure between-raters reliability' (Wagenaar and van der Schrier 1997). Reliability is especially important in systems such as the project under discussion because, as mentioned earlier, apparently reliable output can be produced from highly unreliable data, and the only way to demonstrate that this is not the case is via a reliability trial.

In terms of creating a *new* minor event coding system, we would argue strongly that only by going back to 'first principles' in terms of developing logical models of action can one develop a system with sufficient logical rigour to produce useful data. We propose, therefore, that there are three essential stages in the creation of such a system:

1. to model the salient features of an action (in terms of human factors) in the industry in question;
2. to infer a human factors error action model from this first model; and finally
3. to create logical hierarchies from this last.

In the absence of such modelling, one runs the risk of inadvertently relying on ad hoc structures without a clearly defined logical hierarchy. It was precisely the absence of such a hierarchy which caused such low reliability in the 'old' system used in the industry in question.

It also seems clear that any logical system in terms of minor events should be a *total system*. Considerable unreliability will be introduced into the 'new' system if the raw data collection, and the method by which these raw data are turned into reports, are still carried in a manner congruent with the 'old' system (in other words, if Stage One of the data processing is not compatible with Stage Two). Work is now being carried out in order to create investigation techniques such that information compatible with the Proximal, Intermediate and Distal distinctions will be spontaneously generated, and that the reports will then be written such that they will translate into the coding system with the greatest possible efficiency and accuracy. A coding system is, after all, only as good as the data it has to code.

Given the findings of our trial it is likely that, at present, there is inadequate investigation of minor events and that there is a bias towards more Proximal causes as represented in the reports produced, and that the difference between the comparatively low reliability results of 61% and the figure of 81% produced when this had been taken into account reflects this. It is hoped that an Accident Investigation system based on the 'new' coding system could overcome these biases, and that with reports based on these findings coding systems should aim for a reliability of 75% or higher, as discussed earlier. When allowance had been made for the issues with the reports mentioned above, the new system clearly met this criterion.

There is one final point to make. Even though a system has been proven to be reliable, and can provide a logical and coherent hierarchy for accident investigation, reporting and analyses, the *purpose* of such a system should never be forgotten. A minor event causal analysis system must be able to predict 'weak spots' where more major events may take place, in a manner which facilitates the targeting of resources to such areas. Not only this, but it must be able to *demonstrate* that it does this, by showing an improvement over time in minor event error rates when resources are accurately targeted at these areas. In other words, the system itself must be able to demonstrate that it is not merely reliable but also valid.

It is with this in mind that the system (now termed SECAS or 'Strathclyde Event Coding and Analysis System') underwent a six-month 'pilot' study in 10 nuclear power plants (NPPs), in order to study not only whether the system was felt to be 'easy to use', but also whether it could fulfil the criteria posited above. Only if it was felt that the system had proven that it was both self-monitoring and self-validating was permission to be given for it to be 'phased in' across the industry.

References

Anastasi A (1990). Psychological testing (6th edn). Macmillan, London.

Caro TM, Roper R, Young M, Dank GR (1979). Inter-observer reliability. Behaviour 69:303–315.

Fliess JL (1981). Statistical methods for rates and proportions. Wiley, New York.

Groeneweg J (1994). Controlling the controllable. DSWO Press, Leiden.

Ives G (1991). Near miss reporting pitfalls for nuclear plants. In van der Schaaf TW, Lucas DA, Hale AR (eds). Near miss reporting as a safety tool. Butterworth-Heinemann, Oxford, pp 51–57.

Jackson DN, Messick S (1967). Problems in human assessment. McGraw-Hill, New York.

Lucas DA (1991). Organisational aspects of near miss reporting. In van der Schaaf TW, Lucas DA, Hale AR (eds). Near miss reporting as a safety tool. Butterworth-Heinemann, Oxford, pp 127–147.

Martin P, Bateson P (1993). Measuring behaviour. Cambridge University Press, Cambridge, UK.

NCCAN (1998). Study for National Center for Child Abuse and Neglect (Grant 90-CA-1550). Available online at http://www.nccd-crc.org/crc/nccan.pdf

Taylor RK, Lucas DA (1991). Signals passed at danger: near miss reporting from a railway perspective. In van der Schaaf TW, Lucas DA, Hale AR (eds). Near miss reporting as a safety tool. Butterworth-Heinemann, Oxford, pp 79–93.

Reason J. (1990). Human error. Cambridge University Press, Cambridge, UK.

Shaugnessy J, Zechmeister E (1994). Research methods in psychology. McGraw-Hill, London.

van der Schaaf TW, Lucas DA, Hale AR (1991). Near miss reporting as a safety tool. Butterworth-Heinemann, Oxford.

van der Schaaf TW (1991). Introduction. In van der Schaaf TW, Lucas DA, Hale AR (eds). Near miss reporting as a safety tool. Butterworth-Heinemann, Oxford, pp 1–9.

Wagenaar WA, van der Schrier J. (1997). Accident analysis: the goal and how to get there. Safety Science 26:26–33.

Wilpert B, Fahlbruch B (1996). Integrating human factors in event analysis in nuclear power plants (NPP). International Journal of Psychology 31:3–4, p. 147, 239.4.

Wright L, Davies JB (1997). Setting up a third-party reporting system and human factors data-base with ScotRail. In IDER conference proceedings, The Hague. Andrich International, Warminster, pp 242–248.

Correspondence and offprint requests to: B. Wallace, University of Strathclyde, CIRAS, Lord Hope Building, 141 St James Road, Glasgow G4 0LT, UK. email: brendan.wallace@strath.ac.uk