

RELIABLE QUALITATIVE DATA FOR SAFETY AND RISK MANAGEMENT

Dr Alastair Ross* and Matthew Plunkett
*Human Factors Analysts Ltd. Graham Hills Building, University of Strathclyde, 40 George Street, GLASGOW, G4 1QE; tel: 0141 548 4503; fax: 0141 548 4508; email: alastair.j.ross@strath.ac.uk

Textual or 'qualitative' reports on technical, operational and human factors failures, are rich in information about the motives and intentions of staff, yet difficult to analyse and make a case with, hence their potential as decision-making aids goes unrealised. In this paper we argue that textual accounts are quite acceptable as evidence in safety management, provided they are dealt with systematically.

ABSTRACT

Many industries collect data on technical, operational and human factors failures, which include short textual or 'qualitative' reports. These reports, whilst often rich in information about the motives and intentions of staff, offer a particular analysis problem, and consequently they often accumulate in filing cabinets with their potential as decision-making aids unrealised. In this paper we will argue that textual (or verbal) accounts are quite acceptable as evidence in safety management, provided they are dealt with rigorously and systematically. The watchword for what should or should not be accepted into risk models or safety databases is reliability. *Reliable* data generated from 'the things people say' should not be seen as inherently 'less worthy' than reliable data from any other source. In fact, once reliability is established, frequency output from coded discourse can be treated much as frequencies from any engineered system, and are just as amenable to statistical manipulation. The result of taking the time to analyse such data fully is an integrated risk management system where the 'soft' versus 'hard' data distinction is replaced by a 'reliable' versus 'unreliable' data model, this paradigm offering maximum benefit from data collected.

INTRODUCTION

There are two main arguments against the inclusion of data gathered through textual or verbal reports from staff in risk models. One is that this 'natural discourse' is 'more time consuming and demanding to collect, analyse and make a case with' than 'event' data[1]. Harvey, Turnquist and Agostinelli[3] talk of difficulties in quantifying data gathered through 'unstructured' methods, noting '. . . it is usually a laborious task to train coders to make useful discriminations and thereby to secure reliable judgments" (p 41). This argument, that analysing qualitative data is sometimes difficult, is one we would concur with.

The second argument against using such data is perhaps more fundamental, and that is that there is an epistemic or ontological difference between qualitative and quantitative

data whereby the former, for example, results from a staff interview are just 'things people tell us', whereas 'official' accident statistics, for example, are 'objectively true'. This may lead to the former 'softer' data being completely excluded from risk models, or being 'bolted on', perhaps to support a 'fact' generated from numerical data. This is generally known as 'complimentary triangulation' which involves using qualitative data to add "breadth or depth to our analysis".[4] So we have the 'real' numbers, and 'illustration' in terms of what people tell us (which is subjective).

We will argue in this paper that this second position is hard to sustain, and that, therefore, the fact that doing reliable work with qualitative data is 'laborious' should not deflect us from this work where appropriate. The paper contains an example from the UK rail industry showing how data gathered from textual reports can be an important decision-making aid in the realm of safety because reports in people's own words can add much to the understanding of the workings of a complex socio-technical system. Importantly, however, the data must be dealt with *systematically*. Meaningful analysis of qualitative data relies on reliable categorisation[3] which is discussed below.

THE SUBVERSION OF QUALITATIVE DATA

The typical view is that "qualitative data may . . . be useful in *supplementing* and *illustrating*" (*emphasis added*) quantitative data obtained from, for example, incident investigation.[7] This shows what Brannen calls "pre-eminence of the quantitative over the qualitative" (p.24).[6] Saludadez and Garcia[8] argue that the dichotomous relationship between qualitative and quantitative research is maintained in order to establish a differential in status by turning the quantitative method into something it is not (i.e. objective and value-free, see also[9]).

Hammersley[10] argues that "the distinction between qualitative and quantitative is of limited use and indeed, carries some danger" (p.39). The main danger is that the manager gathers 'objective' numerical data and then either ignored qualitative aspects, or simply *picks and chooses* things people say which seem to back up a particular interpretation. We would argue such an approach is unscientific (we would not agree with picking and choosing numerical data to make a case so why accept this with people's accounts?). We feel it is vital to try and gather qualitative data systematically as well and use it to build the picture. Perhaps the following example will help illustrate this.

THE TRAIN COMPANY EXAMPLE

We were recently asked to look at data from a major Train Operating Company (TOC) who had 169 Signals Passed At Danger (SPADs) over a recent 5 year period. A breakdown of these by type of traction (train unit) showed that, when train miles were taken into account, a certain type of unit (the Class 156, a modern Diesel Multiple Unit) was involved in a relatively high incidence of SPADs.

With these figures in mind, an analysis of SPAD data from this period was undertaken in order to shed light on the relatively high incidence of 156 SPADs. One or two

predictors were identified. However, it proved difficult to find predictors for Class 156 incidents. Problems signals were less of a factor for these incidents, driver history (previous SPAD) did not seem to be a factor and the incidents were not predicted by seasonal aspects (e.g. leaf fall) or environmental factors (line gradient).

Finally, focus groups with drivers were commissioned to shed light on the 'problem' with the 156. However, when responses were coded to make sense of the data, it could be seen that drivers seemed at a loss to explain this. Note the fact that there was a problem with the set was taken for granted at this stage based on the prior incident frequencies. Driver responses included:

"I don't understand why it's the 156. It's just maybe been unlucky it's been a 156 it's been on the job"
"I don't see any technical reason for it (being) a 156"

In fact, there was some consensus that the Class 156 was a good set to drive under difficult (e.g. slippy) line conditions:

"(the 156) they've got a good brake, because you can feel it driving . . . some of them (the 158) are disc brakes. Now that's a lot softer than the 156s – you can feel it (tread brake on the 156) gripping"

Then some drivers appeared to discuss the routes the 156 was used for (compared with other DMUs like the Class 158), and conclude that they were 'more boring'. For example:

"as I say they're used more on the quite monotonous routes"

At this point the possibility was raised that the effect was not due to *the set itself*, or the drivers or any of the other aspects guessed at above. Further discussion proved illuminating, and some consensus emerged from drivers that the Class 156 was used on routes they considered more tedious than other routes where different sets tended to be employed, and that they found it more difficult to pay attention when on these routes (and coincidentally when driving a 156).

Further analysis of the company databases showed 'objective' aspects to this, like numbers of signals on the routes and patterns to the jobs/shifts. The incidents involving 156s were found to be more likely to involve starting against a red signal than running up to (and sliding past) one. This backed up the notion that attention/distraction was a factor. But this important psychological aspect predicting SPADs in a 156 would not have emerged unless drivers were consulted. More particularly, it did not emerge when drivers were told 'there is a problem with the 156' and asked to explain why.

There were a number of implications of this project for the company. Firstly and most obviously the explanation avoided the need to make any (potentially expensive) alterations to the set itself. There had also been a suspicion that the company 'Defensive Driving' programme was somehow failing in that SPADs continued to occur. However this was more targeted at overshooting red signals rather than starting against them

when distracted. The programme was simply not suited to the underlying causes of '156 type' SPADs.

Lastly, the finding of course allowed the supposed cause of the incidents to be addressed. Shifts involving 156 sets were scrutinised, and despatch a communication procedures at stations on the lines involved were reviewed. Data should be available soon to see how effective this has been.

'OBJECTIVITY' AND FREQUENCY DATA

It is, of course, still the case that, 'objectively', Class 156 incidents were more frequent than those involving other types of train. But to classify incidents in this way (i.e. using a taxonomy of train types) is an arbitrary decision. In this case, one might just as well have classified the driver's favourite TV programme to shed light on the events. Once the frequencies have been added up and the statistical association made, however, it is sometimes easy to overlook the fact that a *subjective* taxonomic (classificatory) decision underpins what our database holds, and what we then analyse.

'Incident' itself is simply a *taxonomic classification* safety managers might employ: a bit like 'communication failure' or 'system alignment/isolation not verified', only less specific. Incidents are, because they are a category of events, not independent of the interpretative process.

There will, of course, be strictly codified definitions of many types of incident (plant events, near misses, minor events, abnormal events, 'code reds', health and safety events, dangerous occurrences etc.) which are specific to different industries. However the definition of an incident (in essence something that violates the parameters we set around acceptable operations) is *arbitrary and socially defined.* 'Incident' is exactly a subjective category (or 'code') based on the level beyond which *consensus* states operations are unacceptable.

Put simply, what is an incident is always what we agree is one. We can *redefine* an event as a 'non-incident' simply by raising the acceptability threshold, and so, whilst in practical terms there may be 'actual' values on whatever parameter we are using (for example, temperature) the 'actual' rate of incidents, because it depends on arbitrary thresholds and interpretations of events, remains subjective. If we move the threshold, we alter the frequency.

Note that this does not mean we are arguing that 'incidents' are somehow illusory. Because we can *agree* on them (i.e. they can be measured *reliably*), then actual levels can be used to differentiate between systems and time periods. Once established, the written 'rules' for classifying incidents are 'objective' if consistently (i.e. reliably) applied.

But the implications of removing the 'special' status often conferred upon incident frequency data and other numerical data is to open up other forms of rigorous processing and analysis. The systematic analysis of events subsequent to coding them as incidents (or not) can be seen as analogous to the systematic analysis of verbal reports or texts prior to applying taxonomies designed to identify important aspects.

There is one important proviso. The central axiom on which use of any data should rest is that data are shown to be reliable. A brief discussion of what this means in event classification now takes place. We are conscious that the sceptic will imagine such discussions of reliability of coding to apply in the main to the interpretation of textual data. However, as we have argued above, frequency counts of incidents or accidents, for example, rest on similar interpretation and reliability should be the watchword for these data too.

RELIABILITY

Work in safety management often involves the *classification* of events (e.g. accidents, root causes, contributing variables) using coding schemes or 'taxonomies' which are models of features of a system (for example, human behaviours, organisational/environmental factors, or cognitive factors).[1] Codes applied can then be examined to help avoid unwanted events in the future.

The most important aspect of taxonomic work is whether independent users of a coding scheme, taxonomy or similar diagnostic technique can *agree* on discrete events to be coded.[20] This measure has been termed 'inter — judge'[21], 'inter — observer'[20] or 'inter — rater'[22] *reliability.* Tests of this criterion have usually been simply called *reliability studies.*[23] (The related concept of 'intra — rater' reliability refers to a comparison between the judgments made by the same judge about the same data *on different occasions*).

A basic principle of coding reliability is that agreement refers to the ability to discriminate *for individual subjects, events or cases.*[21, 24, 25] This type of agreement (consensus on individual codes) is a pre- requisite for the validity and pragmatic usefulness of a coding device.[26, 27] Ross, Davies and Wallace[28] outline why using consistency in overall patterns of coding as a test of agreement is flawed. For example, they found a correlation of .837 between two patterns of coded 'root cause' data (i.e. the coding was consistent) for which the raw agreement on individual cases was .42 (i.e. the coding was not reliable). So coding schemes can actually produce highly *consistent* data in the absence of independent *agreement* on discrete events. Indeed, in the most extreme case, *consistent* or reliable coding can be demonstrated in the absence of any *agreement* at all. [This distinction is discussed in detail by James, Demaree and Wolf].[25]

THE CIRAS EXAMPLE

Wallace, Ross and Davies[29] describe a methodology developed for use in CIRAS (Confidential Incident Reporting and Analysis System), the confidential database set up for the UK railways by the University of Strathclyde. The specific approach here was

[1]These have been described respectively as the 'external mode of malfunction'[16] or 'external error mode'[17]; the 'general failure type'[18] or 'external performance shaping factor'[16]; and the 'psychological failure mechanism'[19] or 'internal performance shaping factor'.[16]

developed towards the end of the "roll-out" phase of the CIRAS project, between 1999 and 2001, as a way of obtaining maximum value from the incident reports that were coming into the system at that time.

The initial reports received by CIRAS were raised by staff, and were followed up by telephone interviews with reporters. Thus reports were in the words of reporters and covered a broad range of safety- critical concerns. The quality of information came from the fact that reporters were encouraged to tell it as they saw it, in their own words, and were not subjected to a 'forced choice' method to quantify matters at this stage.

However, this meant a coding procedure had to be developed to allow for frequency data to be generated from the texts. In the analysis procedure described by Wallace et al., data are audio recorded, transcribed in full, converted to rich text files and then imported into the QSR NVivo qualitative analysis software. The taxonomy used for analysis has been reported elsewhere.[29, 30] The software, importantly, allows for the interpretative process to be audited at all stages, from initial thoughts through to final decisions, allowing for the decision making process to be transparent. Repeated reliability trials are then conducted involving discussion of unresolved interpretation until satisfactory reliability is established.[29] The whole interview transcripts are studied by means of the taxonomy which is built onto the software, that is there is no attempt to selectively quote from these as might be common in a standard 'qualitative' analysis.

Frequency data generated from the CIRAS system were analysed in SPSS version 11.0. In the database produced, a row represented a safety incident and a column represented a 'reliability tested' coding variable. Codes were either applied (present = 1) or not applied (not present = 0) for each safety case. So, in a given selection of safety events (or time period), frequencies of codes would rise and fall depending on the number of safety events which, on interpretation of the transcripts, were coded 'present' for each variable. A given issue would be prominent when reports had been analysed, and coders had reliably ascertained that it was present in a significant proportion of them.

The setting up of a database like this is possible wherever textual accounts exist, be they in confidential reports or in reports gathered through other channels. Information obtained in this way, however, will not be particularly useful unless one has a baseline criterion for deciding how 'bad' is 'bad enough to investigate': that is, what the parameters are for issues in reports within which we are willing to operate. But because we have reliable (and now numerical) data which are amenable to statistical analysis, setting these should not in itself present too much of a problem.

CIRAS CONTROL CHARTS

Once we can express our system characteristic in texts as a numerical measurement, we have statistics available to help with decision making. The principle of control charts, for example, is that there are two components to process variation. What is usually called common or system variation is the normal variation inherent in all processes. What we are usually interested in is the 'special' variation outwith the parameters of normal variation, that may be due to some problem or extraordinary occurrence in the

system. Control charts allow for the distinction between these two types of variation to be made easily.

For example, using CIRAS data we generated a model of front- line fatigue related issues in reports from staff based on a sample of 40 weeks of reports (n = 551). This model was then used to set probabilities for future weeks in relation to the number of reports expected and the number of those we might expect to be about front-line fatigue related issues. From this model the probability of any given frequency (e.g. 3 fatigue reports in 12 issues in a week) being 'normal' was set. Control charts were then set for subsequent weeks, with the warning line set at p = .05. If the probability of the data in any given week was higher than p = .05 (given the assumption of 'normal' modelled variation) then we would get no warning, but if the probability of the data obtained given the model we have was lower than p = .05, we would be advised something may be worth investigating. Figure 1 shows the control chart for the 64 weeks from June 2000.

It can be seen from Figure 1 that on one occasion fatigue reports go beyond the warning limits and that these limits are closely approached on another four occasions. The warning occurred in week 44 (beginning April 9th 2001). This was Easter week, and two weeks after British summer time began (with associated changes in shift patterns). Any factor reported can be examined in this way to identify periods were reporting rates require further investigation. It must be stressed that the controlled data above is reliable data produced by an established methodology and then analysed using recognised statistical techniques to analyse patterns in the data and apply controls to it to aid decision making. 'Normal' variation in reports does not then lead to a 'throwing of hands in the air' i.e. if mathematical distribution is taken into account, then we only react to reports which are significantly different from what we might normally expect.

OTHER STATISTICS

Finally, it remains to point out that the methodology we employ (identifying aspects of safety texts reliably to produce binary data) means a number of statistical techniques can apply provided reliability can be demonstrated.[30] Another example of statistical manipulation of data of this type is cluster analysis. This has an advantage in

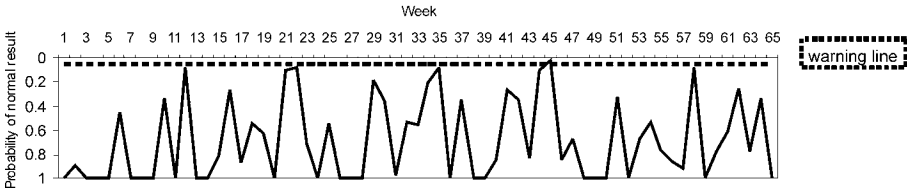

Figure 1. Control chart for fatigue issues raised in CIRAS reports from June 2001

that there are methods of clustering designed specifically for mixed variable types (i.e. for the integration of coded '1/0' data and numerical data from interval or ratio scales).

Cluster analysis allows for accidents or other safety events (cases) to be grouped together on the basis of data generated from the things people say about them or other aspects (variate traces or 'variables')[2]. Clusters generated can be revealing interms of the dimensions that accidents have in common (Byrne[31] calls this 'taxonomy as social exploration'). Random partitioning of the dataset can be used as a check on the robustness of the findings. The interesting point is that we can now examine how safety cases (a specific example would be accidents) cluster around variate traces which include *the things people say about them.* The variate traces emerge from the data analysis, illuminating the accidents or safety cases in terms of what binds them together (or not). This establishes a taxonomy of safety cases which is truly integrative.

A final example of the kind of manipulation that is possible with data of this type is Log linear Modelling (LLM). This can be seen as an extension of $\chi 2$ analysis, and allows for the analysis of more than two categorical variables. It can also be seen as a form of regression in that an equation can be formed to predict the expected log frequency within each cell in a contingency table. The aim of LLM is to be able to predict the expected frequency in each cell of the contingency table. LLM tells us what relationships there are, if any, between the variables. One model will predict the cell totals perfectly and is called the 'saturated' model. However, it may be possible to account for the data using a simpler model. The purpose of Log Linear analysis is to find the unsaturated model that gives the best fit to the observed data, usually by removing factors one at a time based on which has the least effect on the accuracy of the prediction the model makes until a final model is obtained which is the simplest model which does not differ significantly from the actual observed frequencies.

CONCLUSION

In short, we have argued that 'qualitative' data can be processed in such a way as to be built into models of risk. The only real contradiction is between methods which are reliable and robust and those which are not. Of course, when we are dealing with numbers we tend to think of reliability as a comparison between our data and some 'gold standard'. With textual data we tend to make the assumption that all interpretations are *fallible*[32] thus reliability is the degree of convergence of two (or more) fallible sources. However, reliability remains the watchword. And with reliable data, predictions can be made: and dynamic models created such that all available data sources are utilised in the pursuit of safety goals.

[2]The more common term 'variable' will sometimes be used for clarity, nevertheless we concur with Byrne[31] who uses the term 'variate trace' and argues coherently against the notion of variables as 'real' objects which interact outwith the matters of concern themselves (in this context safety cases). Variate trace is a more appropriate term, as we conceive of these data as being derived from a socially agreed taxonomy.

REFERENCES

1. Antaki, C., 1988b, Explanations, communication and social cognition, In Antaki, C. ed., *Analysing everyday explanation: A casebook of methods*, London: Sage pp 1–14.
2. Miles, M.B., 1979, Qualitative data as an attractive nuisance: The problem of analysis, *Administrative Science Quarterly*, 24, 590–601.
3. Harvey, J.H., Turnquist, D.C. and Agostinelli, G., 1988, Identifying attributions in oral and written material, In Antaki, C. Ed., *Analysing everyday explanation: A casebook of methods*, London: Sage pp 32–42.
4. Fielding, N.G. and Fielding, J.L., 1986, *Linking Data Qualitative Research Methods, Vol. 4.*, London: Sage.
5. Campbell, D.T. and Fiske, D.W., 1959, Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychol Bull*, 562: 81–105.
6. Brannen, J., 1992, *Mixing methods: qualitative and quantitative research.* Aldershot, England: Avebury.
7. Robson, C., 1993, *Real world research: A resource for social scientists and practitioner-researchers*, Oxford: Blackwell.
8. Saludadez, J.A. and Garcia, P.G., 2001, Seeing our quantitative counterparts: construction of qualitative research in a roundtable discussion forum, *Qualitative Social Research* [on-line journal], 2, 1, Online. Available HTTP http://qualitative-research.net/fqs/fqs-eng.htm Accessed 20th September 2002.
9. Laurie, H. and Sullivan, O., 1991, Combining qualitative and quantitative data in the longitudinal study of household allocations, *The Sociological Review*, 39, 1: 113–130.
10. Hammersley, M., 1992, Deconstructing the qualitative-quantitative divide, in J. Brannen (ed.) *Mixing methods: qualitative and quantitative research*, Aldershot: Avebury.
11. Flick, U., 1992, Triangulation revisited: strategy of validation or alternative?' *Journal for the Theory of Social Behaviour*, 22, 2: 175–197.
12. Flick, U., 1998, *An Introduction to Qualitative Research*, London: Thousand Oaks; New Delhi: Sage.
13. Breakwell, G.M., Hammond, S. and Fife-Schaw, C. (eds), 1995, *Research methods in psychology*, London: Sage.
14. Denzin, N.K., 1978, *The Research Act. A Theoretical Introduction to Sociological Methods*, New York: McGraw Hill.
15. Maxwell, J.A., 1998,. Designing a qualitative study, in Bickman, L. and Rog, D.J. (eds), *Handbook of Applied Social Research Methods*, London: Thousand Oaks.
16. Rasmussen, J., Pedersen, O. M., Mancini, G., Carnino, A., Griffon, M. and Gagnolet, P., 1981, *Classification System for Reporting Events Involving Human Malfunctions*, Risø National Laboratory: Roskilde, Denmark.
17. Isaac, A., Shorrock, S., Kirwin, B., Kennedy, R., Anderson, H. and Bove, T., 2000, Learning from the past to protect the future- the HERA approach, paper presented at the *24*th *European Association for Aviation Psychology Conference*, Crieff, Scotland, September 2000.
18. Reason, J., 1990, *Human Error*, Cambridge: CUP.
19. Hollnagel, E., 1998, *Cognitive Reliability and Error Analysis Method*, Oxford: Elsevier Science Ltd.

20. Caro, T.M., Roper, R., Young, M. and Dank, G.R., 1979, Inter – Observer Reliability, *Behaviour*, LXIX, 3–4: 303–315.
21. Cohen J., 1960, A coefficient of agreement for nominal scales, *Educ Psychol Meas*, 20: 37–46.
22. Posner, K.L., Sampson, P.D., Capln, R.A., Ward, R.J. and Cheney, F.W., 1990, Measuring interrater reliability among multiple raters: An example of methods for nominal data, *Stat Med*, 9: 1103–1115.
23. Grove, W.M., Andreasen, N.C., McDonald – Scott, P., Keller, M.B., and Shapiro, R.W., 1981, Reliability Studies of Psychiatric Diagnosis, *Arch Gen Psychiat*, 38: 408–413.
24. Fleiss, J.L., 1971, Measuring nominal scale agreement among many raters. *Psychol Bull*, 76: 378–381.
25. James, L.R., Demaree, R.G. and Wolf, G., 1993, r_{wg}: An Assessment of Within- Group Interrater Ageement, *J Appl Psychol*, 78, 2: 306–309.
26. Groeneweg, J., 1998, *Controlling the Controllable: the Management of Safety (4th ed.)*, Leiden: DSWO Press.
27. Stanton, N.A., and Stevenage, S.V., 1998, Learning to predict human error: issues of acceptability, reliability and validity, *Ergonomics*, 41, 11: 1737–1756.
28. Ross A.J., Wallace, B., and Davies, J.B., 2004, Measurement Issues in Taxonomic Reliability, *Safety Sci*, 42, 8, 771–778.
29. Wallace, B., Ross, A.J. and Davies, J.B., 2003, Applied hermeneutics and qualitative safety data: The CIRAS project, *Human Relations*, 56, 5: 587–607.
30. Davies, J.B. Ross A.J., Wallace, B., and Wright, L., 2003, *Safety management: A qualitative systems approach*, London: Taylor and Francis.
31. Byrne, D., 2002, *Interpreting Quantitative Data*, Sage: London.
32. Spitznagel, E.L., and Helzer, J.E., 1985, A proposed solution to the base rate problem in the kappa statistic, *Arch Gen Psychiat*, 42: 725–728.