SUBJECTIVITY IN DATA EXTRACTION: A STUDY BASED ON CONSTRUCTION HAZARD IDENTIFICATION

Simon D Smith[1], Philip Beausang[2], David Moriarty[2] and Jennifer M Campbell[3]

[1] Senior Lecturer in Construction & Project Management, School of Engineering, University of Edinburgh, King's Buildings, Edinburgh, EH9 3JL, UK

[2] BEng Student, School of Engineering, University of Edinburgh, King's Buildings, Edinburgh, EH9 3JL, UK

[3] PhD Student, School of Engineering, University of Edinburgh, King's Buildings, Edinburgh, EH9 3JL, UK

The importance of effective mitigation of construction safety risks can be readily appreciated. However, the methods that are available that can optimise this process are not obvious and no standard procedure is laid down in the construction industry. An IT tool has been developed that allows mitigation measures to be identified for new risks. These measures are based on historically successful measures from a range of projects. Where more than one suitable mitigation technique is available the tool uses Case Based Reasoning to suggest the most appropriate. The data for this tool is based upon that contained in Method Statements, the industry standard format for task and safety planning and one of the more difficult aspects of the research is how to effectively extract the relevant data from these statements. Unfortunately the problem is compounded by the fact that the person extracting the data may be doing so in a subjective manner thus leading to issues with consistency and accuracy. A brief overview of the research work up to the present time will be presented, including an overview of the source of data used in this study, namely Method Statements. A discussion of the problem of subjectivity in data extraction will be provided. The paper will then present a subjectivity filter developed as a partial solution to this problem, together with initial results and findings when applying this filter to a real data extraction problem. These initial results have shown that it is possible to normalise data extraction when this exercise is carried out by multiple extractors.

Keywords: health and safety, information management, information technology, knowledge-based systems, risk.

INTRODUCTION

The safety of employees and public on construction projects has never had such a high profile. The statistics of fatalities and accidents on projects are pored over by the industry and media to the extent that there is often a perception that the UK's construction industry is the most dangerous in the country, with a greater chance of fatality than any other. In actuality this is not the case - the Health & Safety Commission's annual Statistics of Fatal Injuries suggests that the extractive & mining industry and the agricultural industry are much more dangerous in terms of the fatality rate per employee (Health and Safety Commission 2007).

[1] simon.smith@ed.ac.uk

Smith, S D, Beausang, P, Moriarty, D and Campbell, J M (2008) Subjectivity in data extraction: a study based on construction hazard identification. *In:* Dainty, A (Ed) *Procs 24th Annual ARCOM Conference,* 1-3 September 2008, Cardiff, UK, Association of Researchers in Construction Management, 1065-1073.

Nevertheless, the construction industry has, over the last decade or so, made real progress to improving its own act and reducing fatalities. This process has encompassed two editions of the Construction (Design and Management) Regulations and the report of the Construction Industry Task Force (aka the Egan Report). Whilst the statistics vary from year to year, in 2005/06 the rate of fatal injury (3 per 100,000 workers) was the lowest ever recorded. The fact that the rate rose again in 2006/07 (to approximately 3.7 per 100,000 employees) serves to remind the industry that it can never get complacent and assume that the job of minimising fatalities is done.

The role of academia in the general trend of safety improvements should not be overlooked. A great deal of research has been undertaken by many authors in recent years and this trend is likely to continue. At the University of Edinburgh, work on considering the management of safety hazards in the construction and transportation management industries has been undertaken since 1999 and its findings disseminated regularly at this conference and elsewhere. The most recent project in this continuing programme of research has specifically considered the safety risks within transportation maintenance. This work, sponsored by Carillion plc and supported by the EPSRC, has attempted to provide a process whereby the identification of safety hazards is assisted and, further, to provide suggestions on the mitigation of the risks that these hazards may pose. Campbell and Smith (2007) provides a good overview of this project and Campbell et al (2007) considers, more specifically, the details of the problem as applied to the rail maintenance industry.

A general theme to this research is that previous historical knowledge on managing safety hazards should be utilised in the management of future situations. This leads to issues on the identification of this historical knowledge and then developing methods for its capture. This paper presents the problem of extracting safety knowledge from Method Statements, and the subjectivity issues that it presents.

The overall aim of the paper is to understand the extent of subjectivity in data extraction and to propose a possible solution to using data that exhibits variable subjectivity. This will be done by providing a very brief overview of the project and an explanation in more detail of the subjectivity problem. The paper will then present a 'subjectivity filter' as a means of attempting to standardise and objectify subjective knowledge to allow it to be used is the management of future construction activities.

PROJECT BACKGROUND

Previous research suggests that a significant quantity of hazards remain unidentified on construction and transportation projects. Carter and Smith (2006) found the maximum hazard identification level for a construction project in the nuclear sector to be 89.9%; on a project within the rail sector the level was 72.8% and a construction project in the light rail sector only 66.5% of potential hazards were actually identified. These unidentified hazards pose significant risks for the simple reason that non-identified hazards will lead to non-controlled risks. The authors suggest that poor levels of hazard identification can be attributed to certain barriers that exist in the industry: knowledge and information barriers; and process and procedural barriers.

In order to overcome these barriers knowledge management is critical. Knowledge needs to be pooled together into one central source so that information can be shared within the industry to even the smallest of organisations with the most basic of resources. Secondly, this central source must also tap in to tacit knowledge, that is the information that lives inside peoples' heads, and finally it must incorporate as much hidden knowledge contained within processes and procedures as possible

Research at the University of Edinburgh has therefore focussed on creating a library of previous knowledge which can then be used to manage future scenarios. Different ways of utilising this library have been developed, the most recent by using Case Based Reasoning algorithms to select the most appropriate historical case, but in all situations this 'library' needs to be populated in the first instance. It is therefore proposed that the most useful and readily available source of data, knowledge and information for these libraries is within the method statements that have been historically prepared.

SOURCE OF DATA: METHOD STATEMENTS

Method statements describe how a given work task is to be undertaken. They are often paper-based and generally include some form of hazard identification and/or risk management documentation such as a Risk Assessment or COSHH[2] related information. Method statements are used by a variety of workers as a recipe or safe system of work with copies stored at the work task location and other storage facilities such as main or satellite offices, site offices, remote / sub-contracted storage facilities etc.

A series of visits over a two month period allowed method statements from a real transportation project to be collected from a satellite site office in Larkhall, Scotland (UK). This £35m railway construction project was the first new branch line to open in Scotland for 25 years was funded by the Scottish Executive. In total 57 separate method statements were collected and from these data were extracted to form a case library of historical hazard management processes. Unfortunately, the method statements are not prepared for the purpose to which they are utilised in this research: consistency of format is not important, abbreviations and acronyms were plentiful and interpretation of the original authors' intentions sometimes very difficult.

Data to be extracted

Each method statement considers one work task and each was analysed and data extracted manually - automatic processes not possible for the wide variety of documents present. Apart from general task information, such as scope of work and resources used the main data fields extracted can be summarised as:

- All the hazards associated with the work task. The data extractors were also able to add to this list those hazards they felt should be identified but which were not.
- Each hazard could lead to one or more 'harms' via a 'hazardous event' and these were also identified.
- A 'likelihood' number, which is a measure of the probability of harm arising from a hazardous event. This number is determined from consulting both the risk assessments contained in the method statements and the extractors own engineering judgement.
- The control measure put in to place to mitigate against the risks associated with the hazardous events.

[2] COSHH = Control of Substances Hazardous to Health

It was felt from the start that the high level of interpretation of historical method statements required would lead to problems in the normalisation of results - in other words, the authors assumed there would be bias and subjectivity amongst the data extractors. As the next section will demonstrate, this proved to be the case and thus the rest of this paper will concentrate on the issue of subjectivity in data extraction and whether it can be overcome.

INITIAL DATA EXTRACTION TEST

In order to determine the level of subjectivity and breadth of bias amongst operators extracting data, 16 method statements were analysed by three separate extractors, persons A, B and C. Whilst each extracted information based on all the data fields summarised above, the likelihood number was of greatest concern as this required a deal of assumption and judgement by the individual to determine how likely an identified harm would arise from an identified hazard. Figure 1 shows the results of this initial exercise.

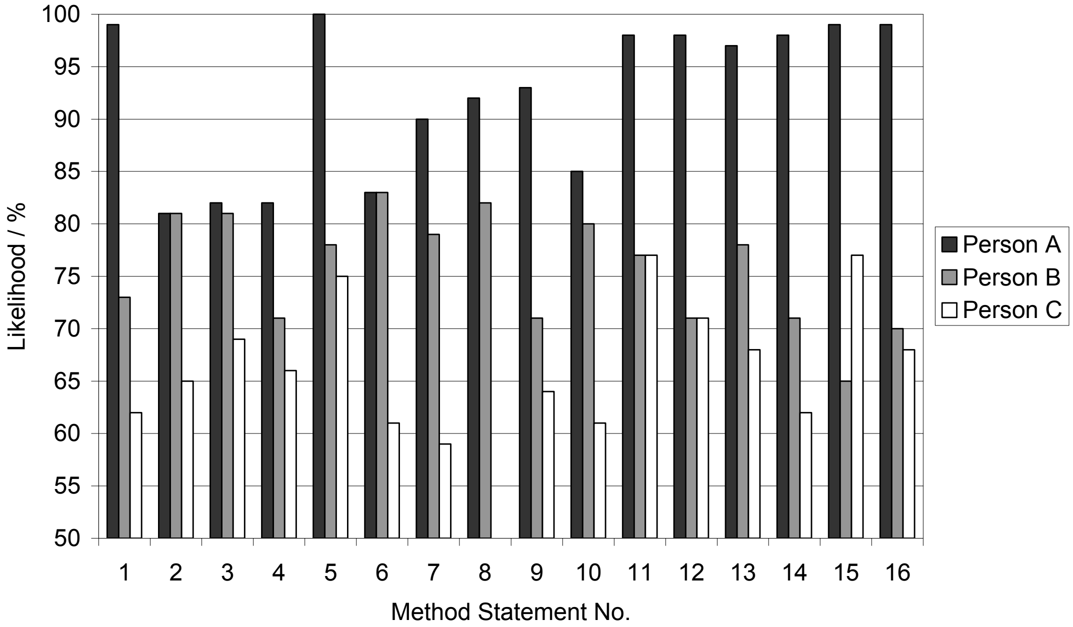

Figure 1: Comparison of likelihood judgements from three different data extractors

Overall it can be seen that Person A consistently judged risk to be higher than B or C. As each considered the same method statements the range of likelihoods, at its greatest being more than 30%, was surprising: whilst difference were expected, the level of subjectivity experienced could significantly compromise the quality of the data library - its formation being the main reason for the extraction of the data in the first place.

SUBJECTIVITY IN DATA EXTRACTION

Subjectivity in qualitative data is not, of course, a new phenomenon nor is it one which has not been investigated before. In the context of risk assessment it should also be expected. "Risk is socially constructed" as Slovic (1999) observes, and risk assessments will be a blended mix of science and judgement, constituted from many factors, including psychological and cultural. This is before the professional and educational experiences of the different operators are taken into account.

In the construction arena, Lingard and Rowlinson (2005) argue that risk decisions are, nevertheless, value judgements about harm and consequence within society. As the public will be affected by such decisions there is no justification in leaving the public out - yet to include them in the decision processes would most likely lead to greater levels of subjectivity. They further argue for a multi-disciplinary approach to risk and that workers should not accept the hegemony of a technical approach as this may lead to inequitable and unsatisfactory outcomes. In other words, not only are the subjective variations between operators expected but are essential if ideal approaches to dealing with the risks are to be formulated.

Unfortunately this does not necessarily help the present problem - whilst variable and subjective outcomes to risk opinions are the norm, the situation is incompatible with a system that relies on quantitative values of historical risk in forming judgements about future risks. If the work described in the early part of this paper is to be taken forward it is important that a method of filtering out the subjectivity be investigated.

DEVELOPMENT OF A SUBJECTIVITY FILTER

The suggestion that subjectivity will be filtered out of the process cannot in reality occur. As discussed above, subjectivity in risk assessment is inherent, and this research project, being conducted in the arena of an engineering and built environment research unit, would be expected to be subject to technical bias.

It is widely accepted that combining multiple expert opinions leads to increased accuracy in forecasting, be it judgemental or statistical (Clemen 1989). This author further observes: "simple combination methods often work reasonably well relative to more complex combinations". It is also important that any proposed 'filter' be simple enough for a construction organisation to accept it as well as implement it.

It is therefore proposed to diminish subjective assessment in method statement data extraction by use of a weighted averaging technique. The weights would be self-derived, based on knowledge components that were felt to be important to the task, and the result would be a weighted average allowing the high risk hazards to be identified and controlled in an appropriate fashion.

Demonstration using case example

In the initial example above three operators were used - these were all students who had either completed or were about to complete the same degree programme at the same university. All three could therefore be considered to have very similar educational backgrounds and career aspirations. In order to consider the filter proposed above it was considered necessary to broaden the background of the operators. Therefore four operators were used, two from the initial test (persons A and B) and two new operators, one being a radiography student and one a health care worker (persons Y and Z). Thus the pool of four operators were all familiar with the concept of risk yet two had little to no knowledge about construction related tasks.

Just one method statement was selected initially, though for a highly complex work task and one that contained many different hazards - 'Removal of bridge deck'. Step one was to determine the likelihood of harm arising from nine separate generic hazards and figure 2 shows the results from this exercise.

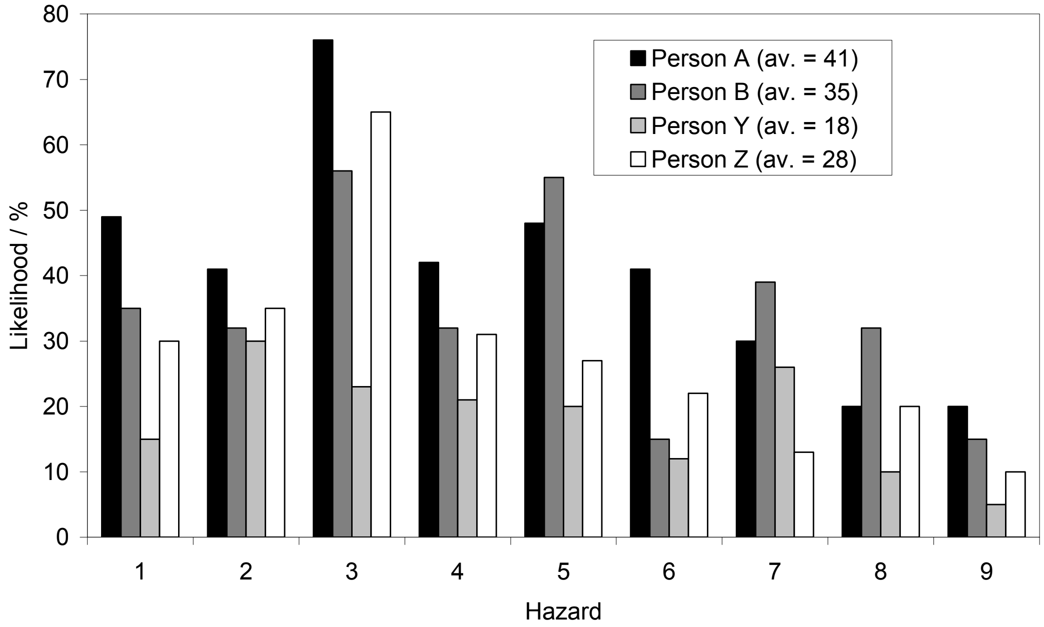

Figure 2 Range of likelihood values for single work task 'Removal of bridge deck'

As expected, the range of values is wide, though what is perhaps interesting is that A & B (engineering undergraduate students) have a combined average likelihood value of 38% and persons Y & Z, who have no engineering or construction experience, have a combined average of 23%. The sample size is too small to make definitive conclusions but it is possible the engineering students are envisaging more detailed and severe outcomes from their experience and rating risks higher. It perhaps contributes further to the argument that individual risk perception has many factors.

The next stage is to determine and apply suitable weightings to reflect the data extractors' knowledge and understanding of the task. There were two aspects to the task - first the person's construction knowledge on the hazard situations themselves; and secondly the person's knowledge of overall risk assessment procedures. Each knowledge area was assumed to have equal importance though these themselves could have weightings applied for different work tasks. Each of the extractors, A, B, Y and Z provided their own score on each of these areas on a scale of 1 to 5 where 5 is complete knowledge and zero is no knowledge at all.

Table 1 indicates both what these knowledge areas are and the individual extractors self assessment of their score. The scores are summed and the weight is determined as the proportion of the overall score of the group of extractors.

Table 1: Knowledge areas for subjectivity filter and self-weighted scores for each data extractor for single work task 'Removal of bridge deck'

		Data Extractor				
Criteria	Sub-criteria	A	B	Y	Z	Total
Construction Knowledge of Hazards	Railway Hazards	4	5	1	1	11
	Construction Hazards	5	5	2	1	13
	Identifying Hidden Services	5	4	2	1	12
	Fuel Containment	5	5	3	2	15
	Plant and Equipment	4	4	1	1	10
	Construction Terms	5	5	1	1	12
	Temporary Works	4	3	1	1	9
	Site Briefings	5	4	3	2	14
	PPE	5	4	4	4	17
Risk Assessment Knowledge	Method Statements	5	5	3	2	15
	Risk Assessments	5	5	3	5	18
	Control Measures	5	4	3	4	16
	Qualitative Assessments	5	5	2	2	14
	Quantitative Assessments	5	5	2	2	14
	Residual Risk	5	4	2	3	14
	COSHH Assessments	5	5	3	5	18
	Total	77	72	36	37	222
	Weight	0.347	0.324	0.162	0.167	

The results in table 1 clearly show data extractor Y & Z's lack of construction knowledge and thus their input to the final likelihood assessment will be low.

Finally, the overall weighted likelihood values for each of the hazards can be determined, as is shown in table 2.

Table 2 Final overall weighted averages for likelihood of hazard causing harm for single work task 'Removal of bridge deck'

	Hazard	A	B	Y	Z	A x 0.347	B x 0.324	Y x 0.162	Z x 0.167	Weighted Average Likelihood
3	Manual handling	76	56	24	65	26	18	4	11	59
5	Explosion or collapse	48	54	19	26	17	18	3	4	42
2	Struck by falling objects	41	32	31	35	14	10	5	6	35
1	Working at height	48	34	16	29	17	11	3	5	35
4	Working near electricity	42	33	22	31	15	11	3	5	34
7	Collision/impacts	29	38	26	13	10	12	4	2	29
6	Harmful substance release	41	14	13	22	14	5	2	4	25
8	Working with lifting equipment	21	31	11	22	7	10	2	4	23
9	Slips, trips, falls	19	14	5	10	7	5	1	2	14

DISCUSSION AND CONCLUSIONS

Table 2 has shown the final likelihoods ranked in order from greatest to least likely. Ambiguity of opinion is now removed and whilst the results will still be subjective - and perhaps even incorrect - the result is far more useful to planners and managers of such future operations. What is more important, in the context of the overall research project, is that this information can be now entered in to the library of past cases, together with the associated information on the control measures that were put in to place.

A comment should be made on the actual results - remembering that the actual work task is 'removal of bridge deck' it is possibly surprising that the data extractors considered 'Working with lifting equipment' to be ranked 8th out of 9; and for 'Manual handling' to be the hazard deemed most likely to cause harm. This reflects the fact that two of the data extractors were non-experts, and two were still to complete their undergraduate studies and were yet to gain deeper experience of construction projects.

This paper set out with the overall aim of understanding the extent of subjectivity in data extraction and to propose a possible solution to using data that exhibits variable subjectivity. This has been done - figure 1 clearly demonstrating the variability of opinion amongst four separate data extractors who were considering 16 different method statements. A 'subjectivity filter' has been proposed, and this is certainly effective in removing the variability amongst data extractors opinions. Whether it actually removes subjectivity will remain a moot point - it is after all a simple weighted average of opinion - yet we will remain encouraged by Clemen's observation that simple methods may be more effective than those which are more complex (Clemen 1999).

If the results of the final case example are flawed then they are included here without reservation, if only to demonstrate the inexact science that is risk assessment, that it covers a range of understandings of the risk concept and, as Redmill (2002) observes, will be on a scale that ranges from wholly objective and measurable to completely subjective and socially constructed. This will be the case even though the context of these examples is technical and serves to remind us that risk assessment can itself pose risks - users feel protected from an analysis which can provide results to an accuracy of one decimal place. Yet they should always question the interpretation of the risks that someone else has provided on their behalf - they should focus on the nature of the hazard as they experience it themselves.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the financial support of the Engineering and Physical Sciences Research Council and Carillion plc to the background research project presented in this paper. They would also like to acknowledge the unnamed assistants at Carillion plc for the data source. Finally, the input from the unnamed 'healthcare worker' and the 'radiography student' to the final case example should be noted.

REFERENCES

Campbell, J and Smith, SD (2007) Safety, hazard and risk identification and management in infrastructure management: A project overview. *In:* Boyd, D (Ed) *Procs 23rd Annual ARCOM Conference*, 3-5 September 2007, Belfast, UK, Association of Researchers in Construction Management, 599-600.

Campbell, J; Smith, SD; Forde, MC and Ladd, R (2007) Identifying hazards in transportation construction and maintenance tasks: a case based reasoning approach using railroad data. *Transportation Research Record*, 1995, 69-75.

Carter, G and Smith, SD (2006) Safety hazard identification on construction projects. *American Society of Civil Engineers, Journal of Construction Engineering and Management*, **132**(2), 197-205.

Clemen, RT (1989) Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**(4), 559-583.

Health and Safety Commission (2007) *Statistics of Fatal Injuries 2006/07*. London: National Statistics.

Lingard, H and Rowlinson, SM (2005) *Occupational Health and Safety in Construction Project Management*. London: Taylor and Francis.

Redmill, F (2002) Some dimensions of risk not often considered by engineers. *Computing and Control Engineering Journal*, **13**(6), 268-272.

Slovic, P (1999) Trust, Emotion, Sex, Politics, and Science: Surveying the Risk-Assessment Battlefield. *Risk Analysis*, **19**(4), 689-701.