Measuring the Human Response to Alarms

Tony Atkinson, Principle Consultant, ABB Consulting, Pavilion 9, Bylands Way, Belasis Technology Park, Billingham, United Kingdom TS23 4EB
tony.atkinson@gb.abb.com

Alarms are an important defence in our process safety hierarchy of control. They provide diverse layers and human flexibility and adaptability in a field dominated by automated and engineering controls. However, in order to function, they rely on the human element in the loop to be available, informed, competent and willing to take the sometimes difficult decisions. LOPA studies often claim the reliability of the human response as 0.1, without any real justification or backup for this claim to be 10 times safer as a result.

The process industry needs a valid mechanism to reliably measure and predict the human response to process alarms in terms of speed and accuracy. Currently there are few practical tools available to measure the likelihood of an effective and timely response to an alarm. Those that exist are difficult to implement in the real world of the control room without impacting on normal operations.

The one exception, the 'Alarm Usefulness Questionnaire' tool has been used to measure the operator experience of alarms in the process industries for nearly two decades. The tool has not previously been subject to an analysis of reliability. Krippendorff's alpha coefficient was used to analyse the results of multiple operators rating the same alarms in a high fidelity simulator. The results demonstrate a lack of agreement as to the usefulness of the alarms, throwing doubt on the reliability of the questionnaire as a measurement tool.

Lack of reliability has significant implications for the use of the tool in professional practice. However, the tool may have a future following a retrospective pilot study and necessary modifications and improvements.

Keywords: Human Factors, Alarms,

Introduction

At 0949hrs on April 20th 2010 on the bridge of the Deepwater Horizon drilling platform 1600m above the Macondo well in the Gulf of Mexico, multiple alarms triggered simultaneously indicating a massive incursion of explosive gas. The job of the operator in responding to those alarms is to evaluate the situation and to sound the general alarm to prepare to evacuate the platform. On the day, the operator on duty (who had only 18 months experience) was overwhelmed. The general alarm was never sounded (USNC, 2011 p.339). 11 people died in the explosion and subsequent fire. The electrical and mechanical components of the alarm worked as designed. The human component failed.

Alarms are an important and ubiquitous part of modern process industry operation. They alert the process operator to abnormal conditions on the process plant, thus enabling him or her to take timely and effective action. The consequence of failing to respond to an alarm typically has an economic (Hollifield and Habibi, 2010 p. 19), environmental or safety consequence (EEMUA, 2013 appendix 1).

Despite their general acceptance as a useful tool for managing abnormal occurrences in the process industry, alarms have been implicated in a number of process industry events that have had global implications, including the fire and explosion at Buncefield (HSE, 2011), the explosion at the Texaco refinery at Milford Haven (HSE, 1997) and explosion at the refinery in Texas City in 2005 (CSB, 2007).

Following these and other events, alarms are now taken seriously at the highest level of regulation of the process industry. In 2015, the implementation of the 'Seveso III' directive of the European Union will mandate that every member state enacts legislation that includes requirements for the management of alarms and alarm systems (EU, 2012, Annex 3). In the United States, CFRs (Codes of Federal Regulations) require strict management of alarm systems monitoring high hazard pipelines transporting natural gas (DOT, 2012a) or hazardous liquids (DOT, 2012b).

These legislative requirements are supported by guidance and standards documents, such as EEMUA 191 (2013) and ISA 18.02 (2009), now embodied in the recent IEC 62682 standard. These are incorporated into the regulatory framework as effectively mandatory for high hazard industries, either as being regarded as 'current best practice' (EEMUA, 2013) or 'Recognised and Generally Accepted Good Engineering Practice' (RAGAGEP) in the case of ISA 18.02 (ISA, 2009; Hollifield & Habibi, 2010 p. 16).

Like any safety related or other critical system, the performance of alarm systems is measured continuously for purposes of verification and demonstration of performance, and all the standards and guidance discussed above mandate measurement and propose benchmark values or acceptability targets (EEMUA, 2013 table 19, ISA, 2009 fig. 14). Much of this measurement has concentrated on numeric performance metrics, including the number of alarms per hour (EEMUA, 2013 p.96), the number of alarms following an incident or abnormality (p.97) or the number of standing alarms (p.98). However, unlike automated functions such as sequence or continuous control or automated shutdown, alarms are only valuable in conjunction with a human response (Hollender, 2010, p.378).

As Bransby and Jenkinson (1997) state, 'the alarm system is installed to make things happen inside the operator's head, not inside some computer system'. Without the human to interpret and action the alarm, nothing happens. This is recognised in the EEMUA guidance with a recommended 'operator usability metric' (EEMUA, 2013 p.98). The EEMUA (2013, appendix

8) recommended tool for measuring alarm system usability from the perspective of the process operator is the Alarm Usefulness Questionnaire (AUQ).

The AUQ has been a part of the recommended measurement suite for alarm systems for some time. The usefulness questionnaire tool first appears in the UK Health and Safety Executive 'The Management of Alarm Systems' report in an appendix titled 'How many alarms are useful?' (Bransby and Jenkinson, 1997, appendix 6, p. 177). There is no information as to how the questionnaire was developed and no attempt at establishing reliability or validity. There is no direct suggestion at this point in its history that the questionnaire be considered a standard measurement tool for the operator experience of alarms or alarm systems. However a method of calculating an overall usefulness 'score' for alarm system usefulness is proposed based on weighted scores for each category of alarm (p.178). The weightings used for calculating the score are acknowledged to be 'fairly arbitrary' but adequate for comparison purposes.

The usefulness questionnaire then appears in all three editions of the EEMUA publication 191, 'Alarm Systems. A guide to design, management and procurement' (EEMUA, 1999, 2007, 2013). This guidance is based substantially on the original HSE report and can be considered an evolution of many of the concepts and findings (Bransby and Jenkinson 1997 p.71).

The EEMUA guidance is significant as 'the most recognised guide to alarm systems' and 'de facto best practice' (Rothenberg, 2009 p.11; Hollender, 2010 p.382). It also has the endorsement of the UK Health and Safety Executive (HSE) writing in the foreword for all three versions, with the second edition stating "Inspectors carrying out assessment and inspection activities may look, when necessary, for evidence that the principles and recommendations in the EEMUA 191 guide (or an equally effective equivalent) are being, or have been, applied to alarm system design and management" (EEMUA, 2007).

The first edition of EEMUA 191 added a target value for the weighted usefulness score generated from the AUQ of less than 2.0 as a 'suggestion' (EEMUA, 1999 p.114), acknowledging that it is an 'empirical value based on industrial experience rather than fundamental theory'. The caveat about the arbitrary nature of the weightings given in Bransby and Jenkinson was not repeated. Subsequent editions of the guide (2007, 2013) have maintained the target of 2.0 as a recommended metric for alarm system performance, losing the warning concerning the empirical nature of the measurements in the rewritten third edition (2013).

It would appear that the AUQ has drifted into the canon of alarm system measurements over the years 'under the radar' and is now firmly established as a measurement tool. Warnings and caveats that accompanied its original publication have been eroded and lost over the years. 'Thousands' of copies of the EEMUA guidance have been sold since publication in 1999 (EEMUA, 2013 p xvi), and the original Bransby & Jenkinson research report remains downloadable from the UK HSE website.

Given that the tool has persisted in the accepted 'best practice' literature (and regulatory environment) for over 15 years, it is surprising that there has been little attempt to check reliability and validity of the tool in practice.

Reliability Measurement

The AUQ tool requires individual operators to characterise each alarm received into one of five categories. That is, they have to rate the experience of responding to the alarm as requiring an action, a check, noting the alarm, little use or an actual nuisance. An important question to ask of any tool that relies on a person rating a particular subjective experience is 'would two people experiencing the same phenomenon rate the experience similarly?' This concept of 'inter-rater reliability' is most common within the social and medical sciences as a part of the process of validation of an existing or proposed measurement or classification system (Freelon, 2010). However, Freelon points out that the concept applies to any area where multiple trained raters make subjective judgements. Whenever coding is divided amongst a population of raters, establishing the degree of reliability is necessary (Lombard et al, 2002). Neuendorf (2002 p.142) considers that establishing inter-rater reliability is 'essential … for valid and useful research' where humans are involved in coding of data. Without reliability, a measure cannot achieve validity, though reliability does not guarantee or infer validity (Neuendorf, 2002 p.141). Given its position in the 'best practice' literature and regulatory environment, the AUQ would appear to be a good candidate for a formal reliability study.

Studies investigating inter-rater agreement typically require that multiple raters rate the same objects or behaviour using the same variable or instrument (Tinsley and Weiss, 1975; Mitchell, 1979). For an agreement coefficient to measure the reliability of a tool it has to be applied to a data set where two or more raters duplicate the process of observation and recording (Krippendorff, 2004). Providing raters are widely available, use identical instructions, are independent of each other, and rate an identical experience (Krippendorff, 2004), then they could be reasonably be expected to rate the experience or observation similarly. A coefficient representing the level of agreement between the raters can be calculated and quoted, which is used to evaluate the reliability of the instrument under consideration (Mitchell, 1979).

Applying this concept of raters duplicating the same process to an identical experience has problems when applied to alarms. In typical process control systems there may be as many thousands of configured alarms (Rothenberg, 2009, p. 77 figure 3.2.1). The chance of two identical alarms annunciating to two or more operators during the period of study is remote. Even where sufficient duplicates do occur, it would be difficult to demonstrate the experience as being the same. An alarm that is annunciated correctly in context may be considered useful. An identical alarm generated from a faulty instrument (Hollifield & Habibi, 2010 p. 123), or as the result of inadequate deadband (p. 110) may be considered a nuisance. Equally it would not be appropriate to introduce artificial disturbances to a high hazard process simply to control the operator experience of alarms.

In order to generate the necessary 'identical experience' for the raters to evaluate, a high fidelity closed loop simulator was employed which is capable of simulating the behaviour of the process plant. The participants (raters) used the system to undergo one of two scripted simulator scenarios where a number of process abnormalities were introduced according to a timed schedule. These abnormalities generated process alarms which were responded to by the participants in real time.

Alarms were therefore presented in the same order, with similar timing and in the same context to each participant. Each operator ('rater') categorised each alarm presented to him or her using the AUQ as the measurement tool. In order to minimise costs (use of simulator suite, development of simulator scenarios, and salary and expenses of participants) the study utilised an existing programme of 'refresher training' for qualified process operators.

Calculating the Agreement Coefficient

Rater agreement is calculated differently for different types of reliability data, nominal, ordinal, interval or ratio (Gwet, 2012). There is a case for the AUQ being either nominal or ordinal level data. This distinction is important when considering reliability. If the data is nominal in nature then any disagreement is treated in the same way in calculating the coefficient. The difference between 'Action' and 'Noted' is the same as between 'Action' and Nuisance'. If the data is ordinal in nature, the difference between adjacent categories is treated as less of a disagreement than the difference between non-adjacent categories when calculating the agreement coefficient (Gwet, 2012 p.5).

Clearly the five categories of the AUQ are named and could be considered nominal. However, can they be legitimately ranked in order of 'usefulness' and therefore be considered ordinal? Unfortunately, while 'nuisance' can be considered the extreme end on the scale of usefulness, it is not clear that 'action', 'check' and 'noted' are necessarily a scale of usefulness to the operator. However, the intent is clearly that the categories are ranked, and ordinal data can be assumed where there is an 'informal hierarchy' (Taylor and Watkinson, 2007). As the clear intention of the measure is that the data is a hierarchy of usefulness, the data has been treated as ordinal for the purposes of selecting and calculating a coefficient of agreement.

When measuring inter-rater reliability it is important to consider the possibility of agreement by chance (Taylor & Watkinson, 2007) and to correct appropriately. As Gwet points out, agreement between raters based on chance alone tells us nothing about the reliability of the measurement tool not the individual rater (Gwet, 2012 p.18). Any coefficient used to measure agreement using the AUQ should therefore correct for chance or 'lucky' agreement.

Krippendorff's alpha coefficient was selected to evaluate the level of inter-rater agreement in this study. The alpha coefficient is a measure of inter-rater agreement that can be used with ordinal data (as well as nominal, interval and ratio), it allows for chance agreement and copes with missing data (Lombard et al, 2002; Taylor & Watkinson, 2007; Neuendorf, 2002). Taylor & Watkinson consider Krippendorff's alpha to the gold standard for reliability measurement. Drawbacks to and criticisms of Krippendorff's alpha appear to be limited to the difficulty of hand calculating the index (Lombard et al, 2002). Neuendorf (2002 p.151) considers Krippendorff's alpha 'highly attractive' as a measure of agreement, but tedious to calculate and therefore 'little used'. A number of solutions to the complexity of the calculations have emerged; including a proposed simplification to the calculation of the coefficient (Gwet, 2011) and a web based calculation service (Freelon, 2010). Support for the calculation of the alpha coefficient in SPSS is now available via a macro 'KALPHA' (Hayes & Krippendorff, 2007), which was utilised for the calculations made for this study. The KALPHA macro used to calculate the Krippendorff's alpha coefficient for this study was version 3.2 and represents the most recent available download (Hayes, 2013).

$$D_o = \sum_{u=1}^{N} \frac{m_u}{n} D_u = \frac{1}{n} \sum_{u=1}^{N} \frac{1}{m_u - 1} \sum_{i=1,j=1}^{m} metric \delta^2_{c_{iu} k_{ju}}$$

Equation 1 - Krippendorff's Alpha (fortunately there's an App for that)

Applying the calculation of the alpha coefficient to a number of operators' rating of the same alarm in the same context gives a numeric value indicating the reliability of the Alarm Usefulness Questionnaire. This numeric value can then be used to judge the reliability of the questionnaire against suitable thresholds of acceptability.

Method

Participants

Participants were process operators from a process plant located in the UK. An opportunity sample representing those participants scheduled for refresher training was utilised. All participants were qualified in the process operator role, with a minimum of 3 years' experience. All had received similar levels of validated training as part of qualification, which included responding to process alarms.

A number of participants undertook both simulator scenarios, which prevented the combining of the datasets for the two scenarios (see Data analysis).

Materials

Participants used a copy of the AUQ which uses wording and layout identical to that published in the EEMUA guidance (EEMUA, 2009 p.136). No additional instructions or guidance completing the form was provided, as none is specified in the published guidance. Additional instructions were provided as to the reason for the study, the need for consent and the ability to withdraw from the study.

Equipment

The simulations were run in a dedicated simulation suite, designed to mimic the normal environment of the control room. This includes the use of similar layout, equipment and furniture to the 'live' control room.

The simulation is run on a Honeywell 'TPS' Distributed Control System (DCS). The participants interact with the simulated exercise via an identical interface to the 'live' plant, including customised keyboards/annunciator panels and mouse (see Figure 2). The simulation is 'closed loop', utilising real plant dynamics and process delays. Alarms were presented with identical audible and visual components to the 'live' system. Dynamic behaviours of the alarm system (e.g. the need to silence and positively acknowledge the alarm) are governed by the preprogramed control system workflow and were identical to the real environment.

Figure 1 - Simulation Suite

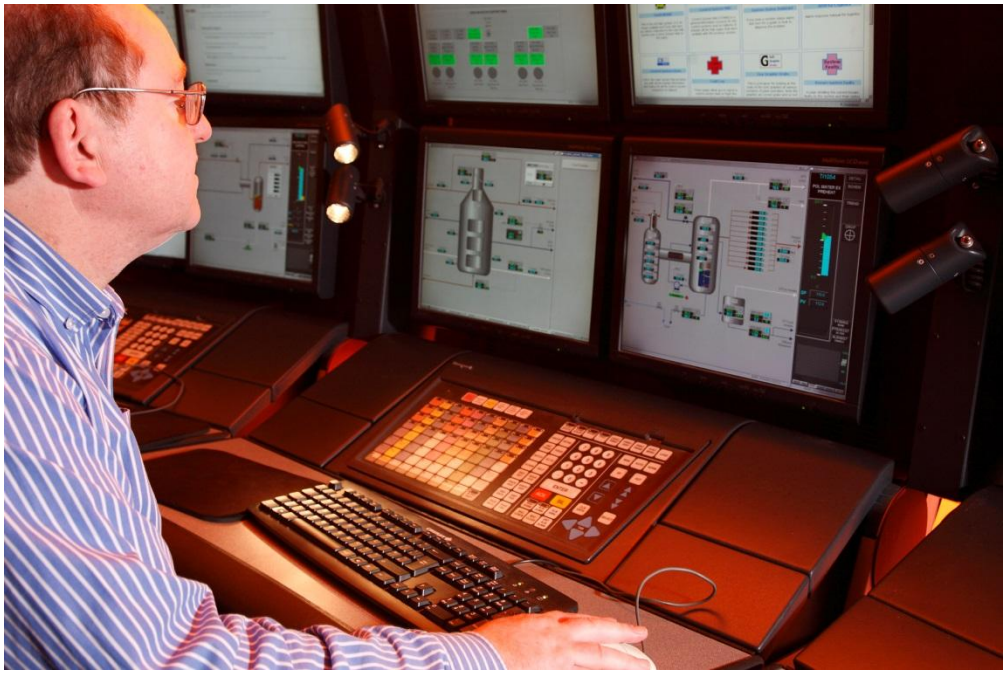

Figure 2 Detail of simulator operator interface showing keyboards, pointer and typical graphical interface.

A mix of alarm types were presented to the operator over the course of the scenario (see Table 1), representing three common alarm causes, threshold alarms, equipment failures and 'off-normal' alarms.

Experimental Procedure

After indicating consent, the participants received instructions for completing the Alarm Usefulness Questionnaire in written format. The participants then undertook one of two possible training exercises using the simulation equipment. The simulation exercise commenced with a steady state process into which simulated failures and instabilities were introduced by a tutor according to a timed script, generating process alarms. Each scenario lasted approximately 90 minutes. Alarms were presented to the operator at a rate of between 2 and 4 per 10 minute period, a rate corresponding to the industry average (Bransby and Jenkinson 1997; EEMUA, 2013).

Training Exercise	Number of Participants/returns	Total Number of Alarms	Number of Threshold Alarms	Number of Equipment Failure Alarms	Number of 'Off Normal' Alarms
ACID	13	28	10	8	10
CO	6 (7*)	23	16	1	6

Table 1 - Characterisation of Training Simulation Exercises

*One questionnaire was eliminated from the study due to ambiguity in rating 19 of 23 alarms.

The operators rated each alarm using the paper copy of the AUQ. A list of all the alarms used in the procedure was made available to the operators (in a 'real' control room a similar historical alarm record would be available to the operator). Paper copies were transcribed to a spreadsheet prior to being copied to SPSS for analysis.

Data analysis

The data analysis was performed using IBM SPSS Statistics for Windows (version 19.0.0). Krippendorff's alpha and related confidence intervals were calculated using the KALPHA macro (Hayes, 2013; Hayes & Krippendorff, 2007) version 3.2 dated 24th April 2013.

Data was structured in a format with each alarm representing a 'case' (row) and each rater represented by a column (see Table 3 - Dataset Acid Simulation & Table 4 - Dataset CO Simulation). This format is recommended as maintaining the most complete level of information for reliability studies (Gwet, 2012 p. 34 and table 2.14). Coded values from the five categories are represented by numeric values, see Table 2.

Category	Value
Action	1
Check	2
Noted	3
Little Use	4
Nuisance	5

Table 2 - Category Coding

Two separate datasets were analysed, one for each simulated scenario (Acid and CO). Although it would have been possible to combine the datasets, this is not recommended as the raters are not guaranteed to be independent of each other across the two scenarios.

		Rater												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Alarm (Case)	1	1	1	5	2	1	1	1	1	1	1	1	1	1
	2	2	4	1	3	3	1	3	2	4	4	2	4	4
	3	2	2	1	3	1	2	3	1	4	1	1	2	1
	4	2	1	1	1	1		1	1	1	1	2	1	5
	5	1	3	1	3	1	1	2	2	4	1	1	1	1
	6	1	4	1	3	1		2	2	3	1	1		1
	7	2	1	1	1	1	1	1	2	3	1	1	1	2
	8	2	1	1	1	1	2	1	1	1	1	1	1	1
	9	1	3	3	3	1	1	2	2	5	1	3	1	2
	10	1	4	4	3	1	1	2	2	5	1	3	1	2
	11	1		1	2	1		3	4	5	1	3	1	2
	12	1	1	2	1	1	1	3	3	2	1	1	2	1
	13	2	5	1	3	2	2	3	4	3	2	3	1	2
	14	1	4	2	2	1	2	1	3	3	1	2	2	2
	15	1	1	2	1	1	1	2	1	1	1	1	1	1
	16	1	2	1	3	1	2	3	1	3	2	1	1	3
	17	1	4	4	3	1	2	3	1	3	2	1	1	3
	18	2	2	2	1	2	2	2	1	3	2	1	2	1
	19	2	2	1	3	2	3	3	1	3	2	1	2	3
	20	2	2	3	3	2	3	3	1	3	2	1	2	3
	21	2	2	1	1	2	3	2	1	3	2	1	2	1
	22	2	2		3	2		3	1	3	2	1	2	3
	23	2	2	1	3	2	3	3	1	3	2	1	2	3
	24	1	1	1	1	1	1	1	2	1	1	1	1	1
	25	1	3	3	3	1	3	2	3	3	1	2	2	3
	26	1	1	1	1	1	1	1	1	1	1	1	1	1
	27	2	3	1	3	2	3	2	1	3	1	1	2	3
	28	2	3	1	3	2	3	2	1	3	1	1	2	3

Table 3 - Dataset (Acid Simulation)

		Rater						
		1	2	3	4	5	6	7
Alarm (Case)	1		1	2	1	1	2	1
	2		1	2	1	3	1	1
	3		2	1	2	1	1	2
	4		3	2			2	2
	5		3	2	1	1	1	1
	6		3	2	3	4	3	2
	7	3	4	1	2	1	1	1
	8		4	1	2	2	1	1
	9		1	1	1	1	1	1
	10		3	1	2	1	3	1
	11		3	2		1	3	2
	12		3	3	2	3	1	2
	13	3	1	2	1	1	1	1
	14		3	3	3	1	2	3
	15	1	3	3	3	1	2	3
	16		3	1	1	1	1	2
	17		1	2	1	2	2	2
	18		3	3	2	2	1	2
	19		1	2	1	1	1	1
	20	3	3	2	1	2	2	2
	21		1	1	1	1	1	1
	22		2	1	1	2	4	1
	23		1	2	1	1	1	1

Table 4 - Dataset CO Simulation (note rater 1 excluded from analysis)

Results

Results of the reliability analysis are given in Table 5 - Reliability Analysis.

Simulation	Mean Alpha	SD	LL95%CI	UL95%CI
Acid	0.175	0.037	0.10	0.25
CO	0.196	0.059	0.08	0.31

Table 5 - Reliability Analysis

Interpreting the results

Krippendorff's alpha defines 1.0 as 'perfect reliability' and 0.0 as 'absence of reliability' (Hayes and Krippendorff, 2007). In interpreting the value of the alpha coefficient, Krippendorff (2004) suggests the need to balance the level of acceptability chosen with the cost of drawing invalid conclusions from the data, with .80 being required where consequences are serious in human or economic terms, and .67 where tentative conclusions are acceptable. Lombard et al (2002) suggest .80 as appropriate in most cases, with lower values being appropriate for indices known to be conservative (such as Krippendorff's alpha). Confidence intervals for the value of alpha are also quoted at the 5% (LL95%CI) and 95% (UL95%CI) level.

Discussion

This study investigated the reliability of the published 'Alarm Usefulness Questionnaire' using interrater agreement measured by Krippendorff's alpha coefficient. The study identifies a clear gap between the reliability achieved by the AUQ and the minimum level of reliability commonly quoted as acceptable. This has some significant implications for the use of the questionnaire in professional practice.

There is no way of telling how many times the AUQ tool has been used in industry, nor the purposes to which the results have been put. Bransby and Jenkinson (1997) make a number of claims for the Alarm Usefulness Questionnaire that rely on the reliability of the AUQ as a tool. In particular, they claim that comparisons can be drawn between companies and plants on the basis of the percentage of alarms in different categories (p.178). Given the lack of inter-rater agreement in the use of the tool this claim may not be appropriate. Bransby and Jenkinson also suggest (p.178) that the number of alarms in the nuisance or little use category could be eliminated in some cases. Again, given the lack of inter-rater agreement, the AUQ tool would not seem to be a suitable mechanism for screening alarms for consideration for removal. Bransby and Jenkinson also use questionnaire results to confirm a hypothesis that 'plant with computer-based alarm systems tend to generate more spurious alarms than plant with hard-wired systems' (p.178), though the questionnaire scores on which this is based are 'close'. As with the other examples, the inter-rater agreement problems with the tool make this claim inappropriate.

Following its publication in Bransby and Jenkinson (1997), the tool was published in all three versions of the EEMUA (1999, 2007, and 2013) guidance. By the release of the third edition of the guidance the AUQ was presented without caveat as a 'benchmark of operator usability' and an 'indicator of ergonomic acceptability' (table 18 and p.98). Without acceptable levels of agreement from raters, the tool should be treated with caution as a benchmark or as an indication of the operator experience. As it stands, the only commonly used measure of operator experience of alarms appears to have significant issues.

Given the above conclusions, the obvious question is 'do we need a measure of the operator experience?' Rothenberg's best practice guide for alarm management (2009) states that 'the alarm rate be matched to the operator's ability to use the alarms' and then goes on to discuss a multiplicity of measures, none of which measure the actual operator experience (ch. 4). ISA 18.02 (2009) fails to recommend any metrics other than the strictly technical. However, if we go back and consider the operator unfortunate to be on duty on board the Deepwater Horizon in the introduction to this paper, we can see a clear need. The human component of the alarm system failed in that case, or more properly it was set up to fail. It should have been foreseeable that the combination of a high demand on the operator ('there were so many alarms, there were hundreds on that page') and a lack of a clearly defined action to take in the event of multiple gas alarms (USNC, 2011 p.339) would result in delayed or no action. Without a reliable and valid measure of the human experience of alarms, we are in danger of having no warning of future similar failures.

Further, with a reliable measure of the likely operator experience, it will be possible to measure the impact of technical measures of alarm system performance on the likely outcome in the real world. At the moment we do not know what contribution the alarm rate, the presentation of alarms, the nuisance alarm burden or the competence of the operator has on the chance of successful and timely response. Is it 20% of each component, is it 50% for one and 5% for another? Outside of the laboratory and simulator there is no way to tell.

The AUQ seems the only proposed measure of the experience of the human component of the alarm system in widespread use. The concept of measuring, reporting and benchmarking the operator experience has persisted throughout the life of the EEMUA guidance document, some 15 years. The tool could also be said to have some 'face validity' that supports this longevity. On the face of things it seems to measure what is important, 'is there a reasonable action that can be taken on receipt of an alarm and is the alarm useful to my job?' If the tool seems to measure the right things and is 'embedded' in the guidance, is it worth persisting with? Can the AUQ be improved to the point of being reliable?

At first glance, the gap between expectation and reality for the reliability of the AUQ seems daunting. However, there are well understood reasons why this may be the case. Neuendorf (2002 p.145) lists a number of threats to reliability of which several could apply to the AUQ as currently implemented and are discussed below. The coding tool (the AUQ) is one factor that can directly affect the reliability coefficient. If the tool is poorly worded, unclear, or ambiguous in nature it will attract more variability in response (Neuendorf p.145, 132). In the case of the AUQ there are a number of areas that could be considered for improvement. Firstly, the naming of the categories. Rather than being individual and unique, the categories of 'action', check' and 'noted' are in fact supersets of each other in the sense that all alarms that attract actions are also checked and noted, and all alarms that are checked are also noted. The second area that may be problematic is the need to fill in the questionnaire in parallel with the demands of the role of process operator. Copying the alarm identifier, descriptor and making the rating judgement is potentially time consuming. In future it may be possible to integrate the process of rating the alarm with the workflow of responding to the alarm. It is important to realise that reliability is not simply a function of the coding tool, but of other factors. These factors include the raters themselves and the units being rated (Neuendorf, 2002 p.145). Raters are normally trained prior to the implementation of questionnaires. Neuendorf strongly advocates coder training (p.133). In the case of the AUQ there is no training or instruction mandated or suggested, other than the 'one page' instruction sheet attached to the questionnaire. Some level of training and guidance as to how to use the questionnaire will be beneficial in increasing reliability. This presents some challenges in the 'real world' of process plant and budgetary constraints. Implementing a training programme (however simple) for the purpose of filling in a questionnaire may be simply impractical. However, there may be solutions, such as internet training, even YouTube videos.

A further source of reliability problems could lie in the alarms themselves. If they are inherently ambiguous, either in design or presentation or in supporting documentation, they could be the origin of lack of agreement between raters. This creates an interesting paradox; the lack of rater agreement (or reliability) becomes a valuable insight into likely operator performance. Essentially the coefficient of agreement itself becomes a useful measure of the alarm system, as well as the measurement tool. However, without the necessary means to compare identical alarms across multiple raters, there would be no way to measure this 'in situ'. Any study into a prospective measurement tool would need to consider ambiguity in the alarms themselves.

If this measure is to be truly useful, it needs to be both reliable and valid. There is no reason that validity beyond the current 'face validity' could not be established to a reasonable level of confidence. The last 20 years has brought a wealth of practical alarm improvement and operational experience to the domain and there are relevant industry bodies that could support development of a valid tool in this field.

The lack of a pilot study or any opportunity to properly evaluate the AUQ tool after its initial implementation is a potential lost opportunity. It seems doubtful that at any point there was a conscious decision not to conduct a pilot study or to establish validity and reliability. The AUQ came into being 'fully formed' and presumably with no specific intent for it to persist for two decades as the sole measure for alarm usefulness. Going back to first principles and conducting a proper pilot study could pay dividends if deficiencies are identified and corrected. In some ways this research could be considered an element of a somewhat belated and wholly retrospective pilot study.

Conclusion

A valid and reliable measure of the human component of the alarm system is a necessary tool in the alarm management 'toolbox'. At the moment, the Alarm Usefulness Questionnaire in its current format is not that tool, due to problems of reliability. If the necessary improvements to the existing AUQ tool are not sought, or if it proves that it cannot be sufficiently improved, then a measure for the human component of the alarm system is still required if we are to measure and improve the overall performance of alarm systems. How this is achieved will be a challenge for the process industries involving alarm system designers, vendors and the operators of process facilities.

Acknowledgements

The author would like to acknowledge the assistance of Phil Jones at BP Hull and Dr. Miles Richardson at the University of Derby in conducting this research.

References

Bransby, M.L. & Jenkinson, J. (1997). The management of alarm systems. A review of current practice in the procurement, design and management of alarm systems in the chemical and power industries. *HSE Contract research report 166/1998*. Norwich:HMSO

CSB (2007) U.S. CHEMICAL SAFETY AND HAZARD INVESTIGATION BOARD INVESTIGATION REPORT REPORT NO. 2005-04-I-TX REFINERY EXPLOSION AND FIRE. Washington DC:CSB

Department of Transportation (2012a). Codes of Federal Regulations. Title 49: Transportation PART 192—TRANSPORTATION OF NATURAL AND OTHER GAS BY PIPELINE: MINIMUM FEDERAL SAFETY STANDARDS. Washington: US Government Printing Office.

Department of Transportation (2012b). Codes of Federal Regulations. Title 49: Transportation PART 195—TRANSPORTATION OF HAZARDOUS LIQUIDS BY PIPELINE: MINIMUM FEDERAL SAFETY STANDARDS. Washington: US Government Printing Office.

EEMUA (1999). Alarm systems. A Guide to Design, Management and Procurement. EEMUA Publication 191. First Edition. London:EEMUA

EEMUA (2007). Alarm systems. A Guide to Design, Management and Procurement. EEMUA Publication 191. Second Edition. London:EEMUA

EEMUA (2013). Alarm systems. A Guide to Design, Management and Procurement. EEMUA Publication 191. Third Edition. London:EEMUA

EU (2012). Directive 2012/18/EU of the European Parliament and of the Council of 4 July 2012 on the control of major-accident hazards involving dangerous substances, amending and subsequently repealing Council Directive 96/82/EC. Official Journal of the European Union L197 pp 1-37

Freelon, D. G (2010). ReCal: Intercoder Reliability Calculation as a Web Service. *International Journal of Internet Science 5(1)* pp20-23

Gwet, K. L(2011) On the Krippendorff's Alpha Coefficient. Retrieved from www.agreestat.com on 18th January 2014.

Gwet, K. L (2012). The Handbook of Inter-Rater Reliability. 3rd Edition. Maryland:Advanced Analytics

Hayes (2013). KALPHA macro, version 3.2 updated 24th April 2013 retrieved from http://www.afhayes.com on 27th December 2013.

Hayes, A. F. & Krippendorff, K (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures (1)* pp 77-89

Hollender, M (2010). Collaborative Process Automation Systems. North Carolina:International Society of Automation

Hollifield, B. & Habibi, E (2010). The Alarm Management Handbook. A Comprehensive Guide. Second Edition. Houston:PAS

HSE (1997). The explosion and fires at Texaco Refinery, Milford Haven, 24 July 1994. London: Health and Safety Executive (HMSO)

HSE (2011). Buncefield: Why did it happen? The underlying causes of the explosion and fire at the Buncefield oil storage depot, Hemel Hempstead, Hertfordshire on 11 December 2005. UK:Health and Safety Executive

ISA (2009). ANSI/ISA 18.2-2009 Management of Alarm Systems for the Process Industries. International Society of Automation. North Carolina: ISA

Krippendorff, K (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. Retrieved from http://repository.upenn.edu/asc-papers/242 on 19th January 2014

Lombard, M., Snyder-Duch, J. & Bracken, C. C (2002). Content Analysis in Mass Communication. Assessment and Reporting of Intercoder Reliability. *Human Communication Research* (28)4 pp 587- 604

Mitchell, S. K (1979). Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies. *Psychological Bulletin 86 (2)* pp376-390

Neuendorf, K. A (2002). The Content Analysis Guidebook. California:Sage Publications.

Rothenberg, D. H (2009), Alarm Management for Process Control. A Best-Practice Guide for Design, Implementation, and Use of Industrial Alarm Systems. New York:Momentum Press

Taylor, J. & Watkinson, D (2007). Indexing Reliability for Condition Survey Data. *The Conservator 30* pp 49-62

Tinsley, H. E. A. & Weiss, D. J (1975). Interrater Reliability and Agreement of Subjective Judgements. *Journal of Counseling Psychology 22(4)* pp 358-376

USNC (2011). Macondo: The Gulf Oil Disaster, Chief Counsel's Report, 2011. Washington: National Commission on the BP Deep Water Horizon Oil Spill and Offshore Drilling