Validating a Process for Understanding Human Error Probabilities in Complex Human Computer Interfaces

Mr Richard Maguire
B.Eng M.Sc C.Eng MIMechE
SE Validation Ltd
10 College Street
Salisbury
Wiltshire
rlm@sevalidation.com

Abstract:

It does seem as though each new generation of equipment becomes increasing complex to operate, understand and interface with. A point agreed to by many of my colleagues who happen to occasionally hire the latest car and have to spend the first hour sat in it trying to tune in radio 2. In defence equipment the drive for better technical and operational capability places a new burden on the operators – as the complexity of the machine increases, there is greater potential for the human to make a significant contribution to that required capability, but also to unwanted irregularities, incidents and even accidents.

Two years ago I led a small team working on understanding human error rates for military aircrew in a new glass cockpit environment. During the research, a method for quantifying the human un-reliability was proposed and demonstrated. This paper now presents the results of a validation exercise undertaken on those derived figures.

Keywords:

HEART, Human Error Assessment, Hierarchical Task Analysis, Fault-tree analysis, Human Computer Interfaces

Introduction:

Since error is endemic in our species [as described by Kirwan (1994)], there are really only two alternatives for modern society; either remove error prone systems completely, or try to understand them better and so minimise the error problems as far as reasonable practicable. Providing there remains a need for complex activities such as air travel and air defence, the first option will not be acceptable, so this limits and forces us to the latter alternative – understanding and mitigation.

In the field of military (all services) rotary wing aircraft, the UK accident rate for 1991 to 2000 averages at around 28 accidents per 100,000 flying hours (UK Inspectorate of flight safety). By comparison, the UK civilian rate for the year 2000 for all aircraft types was around just 6.3 accidents per 100,000 flying hours (Lawrence 2001). In the US Army (not all services) over the period 1987 to 1995 there were nearly 1000 rotary wing accidents costing some $96M and 200 fatalities (Braithwait 1998). These numbers indicate that understanding aircrew error, particularly in rotary wing aircraft, is of significant importance.

The rest of this paper is organised as follows. The concept of identification and quantification of human reliability is summarised; the original research is scoped and presented with the derived results; the validation task is then discussed and finally the validation results are recorded and commented upon.

Quantification of human reliability

The quantification of human reliability is based on having statistically relevant data of human tasks and the associated error rates. Any similar study could refer to the databases and call off the required values and have data that was reasonably fit for purpose. The basic problem with quantitative methods is a lack of data to form the foundation for the assignment of human probabilities to individual task elements. Given that underlying databases are incomplete, experts are asked to provide data that the databases cannot provide (Nagy 2002). This then, leads to a combination of subject matter expert opinion and quantitative analysis supplementing each other, which is open to criticism, argument and may not even be repeatable without careful recording of the expert's demographics. Conventional human reliability analyses are useful in the case of routine highly skilled activities, in the sense that humans may be said to behave very much like machines (Nagy 2002). There is not the need for deep thought, consideration and interpretation of the operating environment. Simple human error analysis methods can certainly be adequate. Increasing complexity of the environment and the human task however, does need a more demanding assessment technique with subsequent validation. Kirwan (1994) suggests a three step method for understanding and decreasing human error, his steps are;

1. Identifying what errors might occur
2. Quantifying the likelihood of occurrence
3. Reducing the error likelihood

Classical risk analysis, as Kirwan records elsewhere, would also establish the severity of the occurrence, and also seek to reduce the impact. But as the method is specific to the subject of human error *occurrence*, it is perfectly acceptable.

Human error identification techniques are numerous, and there are many papers on each technique. As recorded by Wiegmann, Rich and Shappell (2000), Kirwan (1998) describes thirty-eight approaches for error identification. They are categorised by type of approach and are critiqued using a range of assessment criteria. Five broad classifications are developed; taxonomies, psychologically based, cognitive modelling, cognitive simulations and reliability oriented. Several analytical-method classifications are also derived; check-lists, flowcharts, group-based, psychological, representation, cognitive, task analysis, affordance-based, commission identification and crew interactions. The paper does not recommend any single technique, but rather suggests that it is a combinations of techniques and analytical methods that is required.

Similarly, there are multiple quantification techniques. Quantification has always been a thorny issue, and will likely remain so for many years to come. Some behavioural scientists have argued - at times very forcefully - that quantification in principle is impossible (Hollnagel, 2005). This may be true for a specific forecasted tasks with the obvious lack of a statistical-based referent. However, systematic tasks that are required to be regularly done by multiple persons, and which may be reasonably compared to statistically relevant historical data, will give usefully reasonable results, where none perhaps existed before. Consistently exact values of human error rates to three or four significant figures (as may be available for material failures), is currently just not possible, human behaviour is not that regular. Often however, that is not the principle requirement. This may be more on the lines of getting data that is useful for the purpose i.e. for comparative purposes, or simply to determine values to better than a one significant figure 'guestimate'.

There are occassions where quantification simply has to be done, i.e. when it has been deemed a formal requirement, and you've actually got to do it, for whatever reason. Several notable quantitative techniques are well documented in literature SHERPA, HEART and THERP, for reasons of brevity, this paper will only provide a summary of these.

SHERPA (Systematic Human Error Reduction & Prediction Approach) (Stanton & Wilson, 2000). SHERPA works rather like a human based HAZOP. Each task is classified into one of five basic types (checking, selection, action, communication and information retrieval) and then a taxonomy of error types are applied. For each error type an assessment of likelihood and criticality is made. The analysis can be summarised into a classic risk prioritised format, with a quantitative value being assigned to each task with human error. So whilst there are values with external utility, some quantification is done internally and it may be extended via expert opinion to an external temporal reference.

HEART (Human Error Assessment & Reduction Technique) (Maguire & Simpson, 2003) The HEART method involves a classification of identified tasks into proscribed groups from a look-up table, which leads to a nominal human error probability (HEP). Obvious error-producing conditions are applied to the task scenario under investigation in the form of multiplying value, and these values may be themselves factored according to the scenario. The combination of nominal HEP, error producing conditions and factoring ultimately lead to a final HEP value. Expert opinion is used to validate the selection of the task grouping and the error producing conditions.

THERP (Technique for Human Error Rate Prediction) (Nagy 2002): The THERP approach consists largely of a database of probabilities of different kinds of human error, together with performance shaping factors. The analysis starts with a task analysis, graphically represented as event trees. Event trees are structures with logical operators that are used to consider the different potential outcomes of some initiating fault or failure. Human activities are broken down into task elements, which when considered to fail, become the initiating faults. Performance shaping factors such as stress or time are used to modify the probabilities according to expert judgement. The modified result is an estimate of the likelihood of a particular task being carried out in error.

The initial research:

Many aircraft have safety models with calls for human reliability to show an overall probability of a catastrophic event i.e. fatality or aircraft loss. The initial research was designed to assist in populating safety models with appropriate values. It was published two years ago (Maguire & Simpson, 2003) – a brief resume of the scope, methodology and results is probably required for this paper. A specific aircraft type was not specifically defined other than being a rotary wing machine – although, this in no way limits the use of the methodology. The humans under analysis were aircrew undergoing conversion to type training on the aircraft i.e. they are already pilots and are re-focussing on a new aircraft type. An arbitrary but typical mission profile was specified using subject matter experts from the Empire Test Pilot School at MoD Boscombe Down and the Army Training School at MoD Middle Wallop. The developed scenario was of a training pilot carrying out night flying, over undulating terrain, with tree hazards present and flying as one of a pair. As the flight was for a training purpose, the mission had a duration of two hours and was undertaken in good weather before midnight.

A brief overview of historical accident records (Greenhalgh 1999) indicated three flight phases were particularly prone to human error incidents – low-level transit flying, operating at the hover and at landing. These are also considered to be the flight phases where the constructed safety models could get most benefit. This paper will only consider the landing tasks in detail, serving as a demonstration of the original methodology and the validation task.

Following the guidance from Kirwan (1994), the first phase was to identify what errors might occur. This was done using Hierarchical Task Analysis (HTA) and constructing critical task lists. This was done using a combination of techniques – test-pilot interview, procedural analysis and goal analysis. Fortunately, there was a rich source of flight video data to review, and commentary from the corresponding pilot provided excellent information. A typical derived task hierarchy is shown in Figure 1.

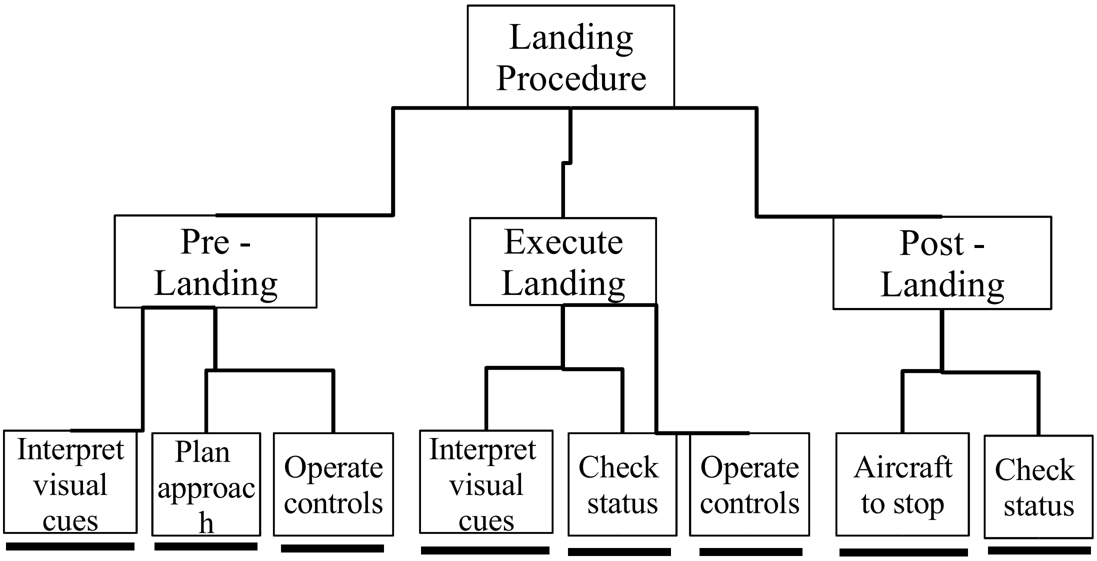


Figure 1Typical task hierarchy segment

The HEART method was utilised for the quantification process. The availability of rich scenario based information, and the need for a faster, cheaper and better response led to this selection. Even in retrospect, this decision has held up well. The HEART method gave a satisfactory set of results, the key stages in their development are shown in Tables 1 to 3.

The next part of the original process was to utilise the task hierarchy to develop a logically operable fault-tree structure for each task segment (Maguire & Simpson 2003). The attempts at these structures led to an increase in required detail being highlighted. For example, the operation of executing the landing had three human tasks initially. The inclusion of some identified crew-resource-management techniques not listed in the flight reference cards or procedures, meant that the extra routine, highly practised, rapid tasks of 'self-check' and 'crew-check' were allowed in the fault-trees. This collaboration between the crew members was shown to reduce the error potential by a full order, encouragement of developing such techniques and collaboration was made in the original research recommendations. The constructed fault-tree was then populated with the values from the HEART analysis. This is shown in this paper in figure 2. A summary of the results from that initial research are presented below, these serves as the object data for the validation task.

Landing phase completed with human errors	5.4e-2 per task
Transit flying phase completed with human errors	2.6e-2 per task
Actions from the hover completed with human errors	3.9e-2 per task

It should be noted that these values do not indicate the frequency of accidents and crashes, but rather the frequency of human errors during these flight phases. Of course the errors may propagate on to accidents, some may be incidents, probably the majority will be just irregularities, which may or may not be officially recorded.

Task	***Description***	***HEART class***	***5th percentile nominal unreliability*** ***(per task call)***
Interpret visual cues	Complex task requiring high level of comprehension and skill	C	0.120
Plan approach	Routine, highly practised, rapid task not involving a high skill level	E	0.007
Operate controls	Completely familiar, well-designed highly practised routine task, performed several times per hour by highly motivated people who are totally aware of the action implications	G	0.00008
Interpret visual cues	Complex task requiring high level of comprehension and skill	C	0.120
Check status	Routine, highly practised, rapid task not involving a high skill level	E	0.007
Operate controls	Completely familiar, well-designed highly practised routine task, performed several times per hour by highly motivated people who are totally aware of the action implications	G	0.00008
Aircraft to stop	Completely familiar, well-designed highly practised routine task, performed several times per hour by highly motivated people who are totally aware of the action implications	G	0.00008
Check status	Routine, highly practised, rapid task not involving a high skill level	E	0.007

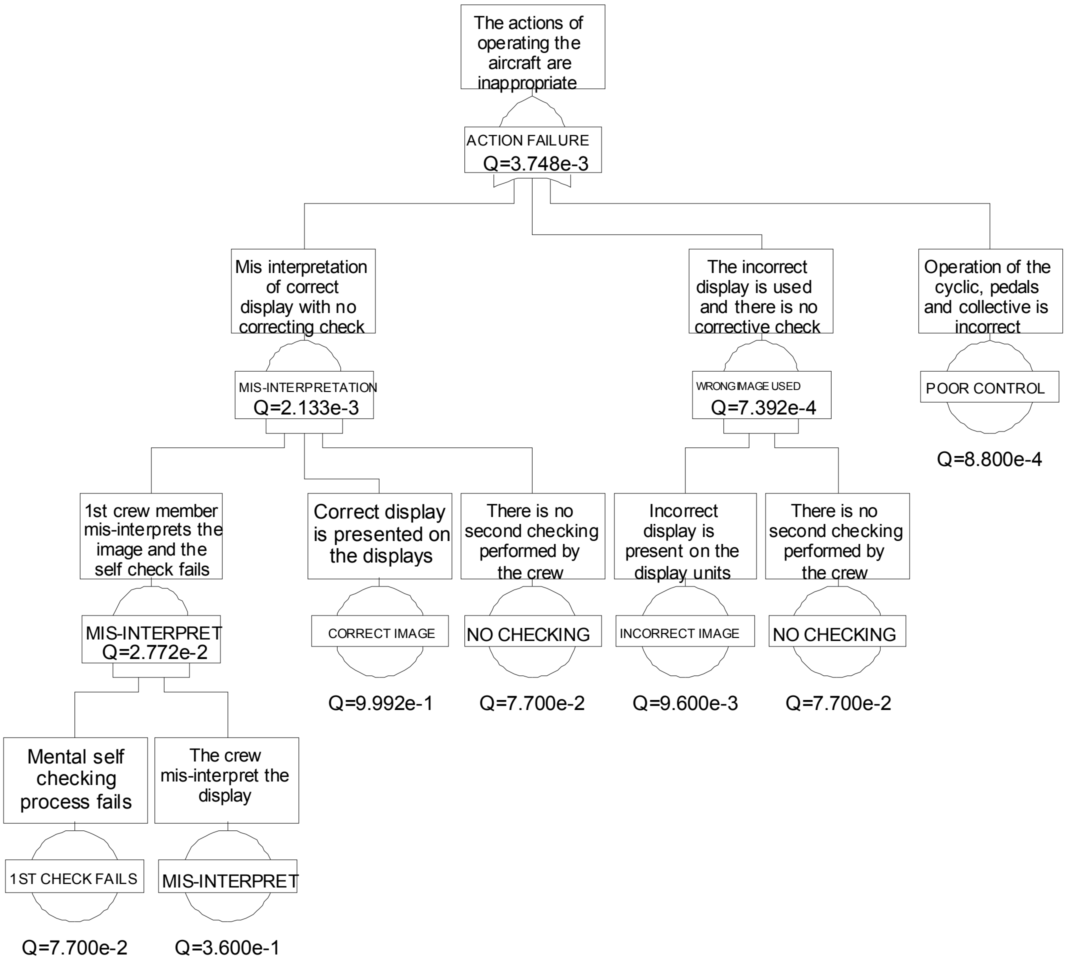
Table 1Landing task classification

Task	***Assigned Error producing conditions***	***Error multiplier effect***
Interpret visual cues	Operator inexperience = 3	3
Plan approach	Shortage of time = 11	11
Operate controls	Operator inexperience = 3	3
Interpret visual cues	Operator inexperience = 3	3
Check status	Shortage of time = 11	11
Operate controls	Shortage of time = 11	11
Aircraft to stop	Shortage of time = 11	11
Check status	None justifiable applicable	1

Table 2Summary of applied EPCs

Task	*Nominal human unreliability*	*Assessed likelihood of error*
Interpret visual cues	0.120	0.360
Plan approach	0.007	0.077
Operate controls	0.00008	0.00024
Interpret visual cues	0.120	0.360
Check status	0.007	0.077
Operate controls	0.00008	0.00088
Aircraft to stop	0.00008	0.00088
Check status	0.007	0.007

Table 3Summary of assessed likelihood of error for aircrew tasks


The actions of operating the aircraft are inappropriate
ACTION FAILURE
Q=3.748e-3
Mis interpretation of correct display with no correcting check
MIS-INTERPRETATION
Q=2.133e-3
The incorrect display is used and there is no corrective check
WRONG IMAGE USED
Q=7.392e-4
Operation of the cyclic, pedals and collective is incorrect
POOR CONTROL
Q=8.800e-4
1st crew member mis-interprets the image and the self check fails
MIS-INTERPRET
Q=2.772e-2
Correct display is presented on the displays
CORRECT IMAGE
Q=9.992e-1
There is no second checking performed by the crew
NO CHECKING
Q=7.700e-2
Incorrect display is present on the display units
INCORRECT IMAGE
Q=9.600e-3
There is no second checking performed by the crew
NO CHECKING
Q=7.700e-2
Mental self checking process fails
1ST CHECK FAILS
Q=7.700e-2
The crew mis-interpret the display
MIS-INTERPRET
Q=3.600e-1

The Validation Task

The information derived from the HEART analysis was for a customer, and that customer wanted to be sure that the information presented was valid and useful. A secondary research task was given to undertake a validation exercise to prove, or not, the accuracy of the original research. Comprehensive validation efforts have taken place on the HEART method along with a comparison of other human error quantification techniques . This validation exercise involved 30 UK based assessors using the three quantification techniques (10 each) HEART, THERP and JHEDI, to review 30 nuclear power plant tasks to determine the HEP (known to the authors). The results for all three techniques were positive in terms of significant correlations. It was found that 72% of all HEP estimates were within a factor of 10 of the true values, and these results lend support to the empirical validity of the techniques, and to human reliability assessment in general (Kirwan et al 1997).

A similar validation task for aircrew tasks has not been undertaken, so it is worthy from a scientific point of view (as well as a customer's) to carry out a dedicated validation exercise for the HEART results developed in the earlier research (Maguire and Simpson 2003).

Raw data for aircrew un-reliability in a typical glass cockpit-based rotary wing aircraft was available from the US Army Safety Center database as reported by Greenhalgh (1999). This data set gave 1370 recorded night flying events (to match the scenario description of the original research). A cut of the data was taken to give a smaller data set from which to derive a referent for the validation. This gave 235 records for the period October 1989 to October 1991. Analysis of the records gave the breakdown shown in Table 4.

Flight phase	***No. of incidents***	***Human error attributes***
Landing	29	15
Hover	37	9
Transit flying	49	6
Other phases (e.g. roll out)	120	5

Table 4 Breakdown of rotary wing recorded event data

The original study (Maguire & Simpson, 2003) derived values with units of '*per task*' and it is perfectly possible to establish similar units for the actual values based on the data in Table 2 in combination with information and expert opinion on the demographics of the flights that led to the accident data. UK and US experts have given the following estimated information, and it is not anticipated that these values are very wide of the real values.

The total number of night time sorties can be determined from the data in Table 2, by the equation;

$$(a \times b) / (d / 60) \quad = 3000 \text{ sorties per year}$$

$$\text{as the data set is over two years} \quad = 6000 \text{ total sorties}$$

Information Items	***Value***
(a) Annual flight hours for the fleet	15,000 hours
(b) Proportion of flt hours as night flying	40%
(c) Night time sortie duration	120 minutes
(d) Number of task calls for landings per flight	3 landings
(e) Number of task calls for hovering	5 hovers
(f) Number of task calls for transit flying	5 transits

Table 5Summary of flight data demographics

Combining the information in Tables 4 and 5 with the calculated number of sorties, gives a series of values for the nominal human error rates in the three flight phases. This is shown in Table 6.

	Landing	***Hover***	***Transit***
Number of sorties (derived as above)	6000 in total over two years		
Number of task calls per sortie	3	5	5
Number of task calls over two years	18,000	30,000	30,000
Number of recorded human errors	15	9	6
Nominal human error rate	8.3E-004	3.0E-004	2.0E-004

Table 6 Calculated human error rates

Whilst these data items appear to be fully appropriate for the validation exercise, they are limited in their completeness. These values represent the officially recorded data, the original research derived data were for the occurrence of human errors not aircraft accidents. This referent data does need to be supplemented to complete the range of human errors, not just those which are cited in accident reports.

There is an accepted relationship between accident rates, incident rates and irregularity rates. It is known by several terms – The Iceberg of Incidents (Davis 2002) and Accident Ratios (The Engineering Council 1993), and essentially it describes the relationship between major, minor and no-effect events. The ratio between these factors is quoted as 1 : 10 : 600.

The 30 human error attributed events from the data set can be arranged in the three categories to check the ratio, as recorded. This arrangement is presented in Table 7.

Flight phase	***Major***	***Minor***	***No effect***
Landing	3	2	10
Hover	1	4	4
Transit	3	1	2
Iceberg ratio	1	10	600

Table 7 Comparison of accident severity ratios with Ice-Berg effect

The no-effect category is far too under populated, they appear to have been un-recorded by a factor of around 100 or so in each flight phase. Research cited in Davis (2002) and The Engineering Council (1993) indicate that these no-effect events do take place, but they are left unrecorded due to embarrassment, doctrine or an opinion that these events do not matter.

Supplementing the recorded data with the expected full data set related to the well recorded major events, gives new figures as the referents for the validation exercise.

Flight phase	***Proposed figures from HEART method***	***Figures from referent source***
Landing	5.4 E-02	8.3 E-02
Hover	3.9 E-02	3.0 E-02
Transit	2.6 E-02	2.0 E-02

Table 8 Comparison of HEART derived data and validation referent

By way of comparison with the Kirwan led validation exercises (Kirwan 1996; Kirwan et al 1997; Kirwan 1997), the proposed human error probabilities are likewise with-in a factor of 10 of the referent data. This does lend support to the empirical validity of the original research methodology of overarching task decomposition, fault-tree derivation and HEART quantification. I understand that the customer is satisfied with his human reliability data for his safety models.

Discussion

Although the method appears quite sound, a number of limitations need to be acknowledged before the data may be used. The experts who helped with the original research were UK based and so gave UK opinion on the task hierarchy breakdown. The referent information was from US sources so the differences in approach to flight safety, crew resource management and event recording is likely to be different. It remains unclear as to how much this has affected the results.

The availability of accurate flight demographics is a concern, although even if these values have error bands of +/- 50%, the end comparison is still within the same order of magnitude. A similar case has to be accepted for the quantity of no-effect events that are added back into the referent data set, which due to their size, obviously swamp the more severe putcome events.

However, a validation exercise has been carried out. The referent used for this exercise may be considered reasonable. The comparison between the proposed values and the referent has been shown to be satisfactory, and hence the method and data set derived may be considered fit for the prupose of better understanding human errors.

References

Braithwait Col. M 1998 'Spatial disorientation in US Army rotary wing operations' –. Aviation, space and environmental medicine Vol 69, No 11.

Davis R 2002 'An introduction to system safety management and assurance' - MoD ALTG-LSSO, MoD Abbey Wood, Bristol

Greenhalgh Lt Col J.G. 1999 'BLO aviation report 99/14 BAS(W) tasking matrix 96/1/092 TADS/PNVS FLIR quality – data from US Army Safety Center database Oct 89 to Mar 99' Fort Rucker, USA.

Hollnagel E 2005 'Human reliability analysis' – website based notes

Kirwan B 1994 'A guide to practical human reliability assessment' – Taylor and Francis.

Kirwan B 1996 'The validation of three human reliability quantification techniques – THERP, HEART and JHEDI – part I – Techniques and descriptions' Applied Ergonomics v27/6.

Kirwan B, Kennedy R, Taylor-Adams S & Lambert B 1997 'The validation of three human reliability quantification techniques – THERP, HEART and JHEDI – part II – Results of the validation exercise' v28/1

Kirwan B 1997'The validation of three human reliability quantification techniques – THERP, HEART and JHEDI – part III – Practical aspects of the usage of the techniques' Applied Ergonomics v28/1

Kirwan B 1998 'Human error identification techniques for risk assessment of high risk systems, part 1 : review and evaluation of techniques.' Applied Ergonomics v29/3

UK Inspectorate of Flight Safety figures 1991 to 2000.

Lawrence P 2001'Human factors in aerospace design for operations' –. University of the West of England.

Maguire R. L. & Simpson Dr A. 2003 'Quantification of military helicopter aircrew human error probabilities for safety models' MoD Equipment Safety Assurance Symposium 2003

Nagy G 2002 'Human reliability analysis : From action to context' – Carleton University.

Stanton Prof. N & Wilson J.A. 2000 'Human factors: step change improvement in effectiveness and safety' Drilling Contractor, Jan/Feb 2000.

Wiegmann D A, Rich A. M & Shappell S. A. 2000 'Human error and accident causation theories, frameworks and analytical techniques : an annotated bibliography' – Aviation Research Lab, University of Illinois