**Predicting Pilot Error: Assessing the Performance of SHERPA**

Neville A. Stanton, Mark S. Young, Paul Salmon (1), Andrew Marshall (2), Thomas Waldman (3),

Sidney Dekker (4).

(1) Department of Design, Brunel University, Egham, Surrey, TW20 OJZ, UK, +44 (0)1784 431341

(2) Marshall Associates, London, UK

(3) University of Limerick, Ireland

(4) Linkoping University, Sweden

**Abstract**: This paper introduces SHERPA (Systematic Human Error Reduction and Prediction Approach) as a means for predicting pilot error. SHERPA was initially developed for predicting human error in the nuclear industry about 15 years ago. Since that time validation studies to support the continued use of SHERPA have been encouraging. Most research shows that SHERPA is amongst the best human error prediction tools available. Yet there is little research in the open literature of error prediction for cockpit tasks. This study attempts to provide some evidence for the reliability and validity of SHERPA in aviation domain.

**Keywords**: SHERPA, errors, reliability, validity

**Introduction**

Human error is an emotive topic. Psychologists and Ergonomists have been investigating the origins and causes of human error since the dawn of the discipline (Reason, 1990). Traditional approaches suggested that error was an individual phenomenon, the individual who appears responsible for the error. Indeed, so­ called 'Freudian slips' were treated as the unwitting revelation of intention: an error revealed what a person was really thinking but did not wish to disclose. More recently, error research in the cognitive tradition has concentrated upon classifying errors within taxonomies and determining underlying psychological mechanisms (Senders & Moray, 1991). The taxonomic approaches by Norman (1988) and Reason (1990) have led to the classification of errors into different forms, e.g. capture errors, description errors, data driven errors, association activation errors and loss of activation errors. Reason (1990) and Wickens (1992) identify psychological mechanisms implied in error causation, for example the failure of memory retrieval mechanisms in lapses, poor perception and decision-making in mistakes and motor execution problems in slips. Taxonomies offer an explanation of what has happened, whereas consideration of psychological mechanisms offer an explanation of why it has happened. Reason (1990), in particular, has argued that we need to consider the activities of the individual if we are able to consider what may go wrong. This approach does not conceive of errors as unpredictable events, rather as wholly predictable based upon an analysis of an individual's activities. Since the late l 1970's much effort has been put into the development of techniques to predict human error based upon the fortunes, and misfortunes, of the nuclear industry. Despite this development, many techniques are poorly documented and there is little in the way of validation studies in the published literature.

**Validating Human Error Prediction**

Whilst there are very few reports of validation studies on ergonomics methods in general (Stanton and Young, l 999a), the few validation studies that have been conducted on HEI are quite optimistic (e.g. Kirwan, 1992a, b; Stanton and Baber, 1996). It is encouraging that in recent years the number of validation studies has gradually increased. Empirical evidence of a method's worth should be one of the first requirements for acceptance of the approach by the ergonomics and human factors community. Stanton and Stevenage (1998) suggest that ergonomics should adopt similar criteria to the standards set by the psychometric community, i.e. research evidence of reliability and validity before the method is widely used. It may be that the ergonomics community is largely unaware of the lack of data (Stanton and Young, 1998) or assumes that the methods provide their own validity (Stanton and Young, l 999b).

The development of HEI techniques could benefit from the approaches used in establishing psychometric techniques as two recent reviews demonstrate (Bartram et al, 1992, Bartram et al, 1995). The methodological concerns may be applied to the entire field of ergonomics methods. There are a number of issues that need to be addressed in the analysis of human error identification techniques. Some of the judgments for these criteria developed by Kirwan (1992, b) could be deceptive justifications of a technique's effectiveness, as they may be based upon:

• User opinion

• Face validity

• Utilisation of the technique.

User opinion is suspect because of three main reasons. First it assumes that the user is a good judge of what makes an effective technique. Second, user opinion is based on previous experience, and unless there is a high degree of homogeneity of experience, opinions may vary widely. Third, judgments may be obtained from an unrepresentative sample. Both Kirwan (1992, b) and Baber & Stanton's (1996) studies used very small samples. Face validity is suspect because a HEI technique might not be able to predict errors just because it looks as though it might, which is certainly true in the domain of psychometrics (Cook, 1988). Finally, utilisation of one particular technique over another might be more to do with familiarity of the analyst than representing greater confidence in the predictive validity of the technique. Therefore more rigorous criteria need to be developed.

Shackel (l990) proposed a definition of usability comprising effectiveness (i.e. level of performance: in the case of HEI techniques this could be measured in terms of reliability and validity), learnability (i.e. the amount of training and time taken to achieve the defined level of effectiveness) and attitude (i.e. the associated costs and satisfaction). These criteria together with those from Kirwan (1992, b: i.e., comprehensiveness, accuracy, consistency, theoretical validity, usefulness and acceptability) and the field of psychometrics (Cronbach, 1984; Aiken, 1985) could be used to assess HEI techniques (and other ergonomics methods) in a systematic and quantifiable manner.

**Systematic Human Error Reduction and Prediction Approach (SHERPA)**

SHERPA (Embrey, 1986) uses Hierarchical Task Analysis (HTA: Annett et al. 1971) together with an error taxonomy to identify credible errors associated with a sequence of human activity. In essence the SHERPA technique works by indicating which error modes are credible for each task step in tum, based upon an analysis of work activity. This indication is based upon the judgement of the analyst, and requires input from a subject matters expert to be realistic.

The process begins with the analysis of work activities, using Hierarchical Task Analysis. HTA is based upon the notion that task performance can be expressed in terms of a hierarchy of goals (what the person is seeking to achieve), operations (the activities executed to achieve the goals) and plans (the sequence in which the operations are executed). Then each task step from the bottom level of the analysis is taken in tum. First each task step is classified into a type from the taxonomy, into one of the following types:

• Action (e.g. pressing a button, pulling a switch, opening a door)

• Retrieval (e.g. getting information from a screen or manual)

• Checking (e.g. conducting a procedural check)

• Selection (e.g. choosing one alternative over another)

• Information communication (e.g. talking to another party)

This classification of the task step then leads the analyst to consider credible error modes associated with that activity. From this classification the associated error modes are considered. For each credible error (i.e. those judged by a subject matter expert to be possible) a description of the form that the error would take is given. The consequence of the error on the system needs to be determined next, as this has implications for the criticality of the error. The last four steps consider the possibility for error recovery, the ordinal probability of the error, its criticality and potential remedies.

**Studies of SHERPA**

Kirwan (1992b) conducted a comparative study of six potential HEI techniques. For this study he developed eight criteria on which to compare the approaches. In his study, Kirwan recruited 15 HEI analysts (three per technique, excluding group discussion). Four genuine incidents from the nuclear industry were used as a problem to focus the analysts' effort. This is the main strength of the study, providing a high level of ecological or face validity. The aim of the was to see if the analysts could have predicted the incidents if the techniques had been used. All the analysts took less than two hours to complete the study. Kirwan presented the results for the performance of the techniques as both subjective judgments (i.e.: low, medium and high) and rankings (i.e. worst and best). No statistical analysis was reported in the study, this is likely to be due to methodological limitations of the study (i.e. the small number of participants employed in the study). From the available techniques, SHERPA achieved the highest overall rankings and Kirwan recommends a combination of expert judgement together with the SHERPA technique as the best approach.

A study by Baber & Stanton (1996) aimed to test the hypothesis that the SHERPA technique made valid predictions of human errors in a more rigorous manner. In order to do this, Baber & Stanton compared predictions made by an expert user of SHERPA with errors reported by an observer. The strength of this latter study over Kirwan's is that it reports the use of the method in detail as well as the error predictions made using SHERPA. Baber & Stanton's study focuses upon errors made during ticket purchasing on the London Underground, for which they sampled over 300 transactions during a non-continuous 24-hour period. Baber and Stanton argue that the sample was large enough as 90% of the error types were observed within 20 transactions and after 75 transactions no new error types were observed. From the study, SHERPA produced 12 error types associated with ticket purchase, nine of which were observed to occur. Baber & Stanton used a formula based upon Signal Detection Theory (Macmillan & Creelman, 1991) to determine the sensitivity of SHERPA in predicting errors. Their analysis indicated that SHERPA produces an acceptable level of validity when used by an expert analyst. There are, however, two main criticisms that could be aimed at this study. First, the number of participants in the study was very low; in fact only two SHERPA analysts were used. Second, the analysts were experts in the use of the technique; no attempt was made to study performance whilst acquiring expertise in the use of the technique.

Stanton & Stevenage (1998) conducted two experimental studies to test the learnability of SHERPA with novice participants. In the first study, the error predictions of 36 participants were compared to those who had no formal error methodology, to see if people using SHERPA performed better than heuristic judgement. Similar to the Baber & Stanton (1996) study, these predictions were made on the task that required people to make a purchase from a vending machine. Participants using the SHERPA technique correctly predicted more errors and missed fewer errors than those using the heuristics. However, they also appeared to incorrectly predict more errors. There appears to be a trade-off in terms of training such that a more sensitive human error identification is achieved, at the cost of a greater number of false positives. This is probably a conservative estimate, as no doubt if the observation period was extended indefinitely, more error types would be observed eventually. In the second study, 25 participants applied SHERPA to the vending task on three separate occasions. The data reported by Stanton & Stevenage show that there is very little change over time in the frequency of hits and misses however, the frequency of false alarms appears to fall over time and consequently, the frequency of correct rejections appears to increase. In terms of the overall sensitivity of error prediction, this shows remarkable consistency over time.

**Predicting Pilot Errors in the Autopilot Task Using SHERPA**

The purpose of this study was to evaluate the SHERPA methodology applied to the analysis of the flight deck for the autoland task. There are many limitations on this study. For starters, there is no attempt to evaluate the dialogue between the pilot, the co-pilot and air traffic control. It is already assumed that autoland will be used. There are also limitations with regard to the collection of error data from pilots, which largely relied upon self-report to a questionnaire survey. Nevertheless, within these limitations, some insight into the success with which SHERPA can be applied to an aviation domain can be gleaned

Eight graduate engineering participants aged between 22 and 55 years took part in this study. All participants were trained in the SHERPA methodology. The training comprised an introduction to the key stages in the method and a demonstration of the approach using a non-aviation example, using an in-car task from Stanton & Young (l999a). Participants were then required to apply the method to another non­ aviation task with guidance from the instructors from a public technology task from Stanton & Stevenage (1998). The purpose of this was to ensure that they had understood the workings of the SHERPA method. A debriefing followed, were participants could share their understanding with each other. When the instructors were satisfied that the training was completed, the main experimental task was introduced. This required participants to make predictions of the errors that pilots could make in the autoland task.

To make their error predictions, participants were given a HTA of the autoland task developed by the authors (comprising some 22 subtasks under the main headings: setting up for approach, lining up for the runway, and preparing the aircraft for landing), a demonstration of autoland via Microsoft flight simulator, the SHERPA error taxonomy, and colour photographs of: the autopilot panel; levers for flaps, landing gear and speed brake; the primary flight displays; and an overview of the cockpit.

Participants were required to make predictions of the pilot errors on two separate occasions, separated by a period of four weeks. This enabled intra-analyst reliability statistics to be computed. The predictions were compared with error data reported by pilots using autoland. This enabled validity statistics to be computed. The signal detection paradigm provides a useful framework for testing the power of HEI techniques. In particular, it identifies type I errors (a miss: when the error analyst predicts the error will not occur and it does) and type II errors (a false alarm: when the error analyst predicts that there will be an error and there is not) in the judgement of the analyst.

Analysis of the data revealed the mean reliability of analysts between time one and time two using SHERPA as approximately 0.7 and mean validity, expressed as a mean of the hit and false alarm rates as approximately 0.6. These values are moderate, but it should be noted that this was the first time the participants had applied the SHERPA method in anger and that they were not aviation experts. The pooled error predictions are compared to the errors reported by pilots in table one. If the error predictions are pooled, the validity statistic rises to approximately 0.9 which is very good indeed.

Table l. Pooled error data

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Errors Observed |  |
|  |  | Yes | No |
| Errors Predicted | Yes | Hits=52 | F.A.=4 |
| No | Misses=5 | C.R.=179 |

**Conclusions**

In conclusion, the results are promising for the use of SHERPA in predicting pilot error. Whilst more studies are needed to investigate different tasks, the current study shows that novices were able to acquire the approaches with relative ease and reach acceptable levels of performance within a reasonable amount of time. This supports the investigation by Stanton & Stevenage and is quite encouraging. The study also shows that HEI techniques can be evaluated quantitatively.

Human error is a complex phenomenon, and is certainly far from being completely understood. Yet in attempting to predict the forms in which these complex behaviours will manifest themselves armed only with a classification systems and a description of the human and machine activities it is amazing what can be achieved. Despite the gaps in our knowledge and the simplicity of the techniques, the performance of the analysts appears surprisingly good. This offers an optimistic view of the future for human error identification techniques. There are a number of other criticisms that need to be addressed, however. Stanton and Stevenage (1998) propose that clearer documentation on the methodologies needs to be provided, and that cross validation studies should be undertaken.

**Acknowledgement**

This research is supported by a grant from the Department of Trade and Industry as part of the European EUREKA! research programme.

**References**

Aitkin, L. R. (1985) Psychological Testing and Assessment. Allyn & Bacon: Boston.

Annett, J.; Duncan, K. D.; Stammers, R. B. & Gray, M. J. (1971) Task Analysis. Training Information

No. 6. HMSO: London.

Baber, C. & Stanton, N. A. (1996) Human error identification techniques applied to public technology: predictions compared with observed use. Applied Ergonomics. 27 (2) 119-131.

Bartram, D.; Lindley, P.; Foster, J. & Marshall, L. (1992) Review of Psychometric Tests

(Level A) for Assessment in Vocational Training. BPS Books: Leicester.

Bartram, D.; Anderson, N.; Kellett, D.; Lindley, P. & Robertson, I. (1995) Review of Personality

Assessment Instruments (Level B) for use in Occupational Settings. BPS Books: Leicester. Cook, M.

(1988) Personnel Selection and Productivity. Wiley: Chichester.

Cronbach, L. J. (1984) Essentials of Psychological Testing. Harper & Row: New York.

Embrey, D. E. (1986) SHERPA: A systematic human error reduction and prediction approach. Paper

presented at the International Meeting on Advances in Nuclear Power Systems, Knoxville, Tennessee.

Kirwan, B. (1992a) Human error identification in human reliability assessment. Part l:

overview of approaches. Applied Ergonomics, 23 pp. 299-318.

Kirwan, B. (1992b) Human error identification in human reliability assessment.

Part 2: detailed comparison of techniques. Applied Ergonomics, 23 pp. 371-381.

Macmillan, N. A. & Creelman, C. D. (1991) Detection Theory: a user's guide. Cambridge University

Press: Cambridge.

Norman, D. A. (1988) The Psychology of Everyday Things. Basic Books: New York.

Reason, J. (1990) Human Error. Cambridge University Press: Cambridge. Senders, J. W. & Moray, N.

P. (1991) Human Error. LEA: Hillsdale, NJ.

Shackel, B. (1990) Human factors and usability. In: Preece, J. & Keller, L. (eds)

Human-Computer Interaction. Prentice-Hall: Hemel Hempstead.

Stanton, N. A. & Baber, C. (1996) A systems approach to human error identification. Safety

Science, 22, pp. 215-228.

Stanton, N. A. & Stevenage (1998) learning to predict human error: issues of reliability,

Validity and acceptability. Ergonomics 41(11), 1737-1756

Stanton, N. A. & Young, M. (1998) Is utility in the mind of the beholder? A review of

ergonomics methods. Applied Ergonomics. 29 (I) 41-54

Stanton, N. A. & Young, M. (l999a) A Guide to Methodology in Ergonomics: Designing for Human Use. Taylor & Francis: London.

Stanton, N. A. & Young, M. (1999b) What price ergonomics? Nature 399, 197-198

Wickens, C. D. (1992) Engineering Psychology and Human Performance. Harper Collins: New York.