Consolidation of the HEART Human Reliability Assessment Principles

J. Bell, Health and Safety Laboratory, Harpur Hill, Buxton, SK17 9JN and J.C. Williams, Jerry Williams Human Factors Ltd. Hale Barns, WA15 0BW.

The Human Error Assessment and Reduction Technique (HEART) has been used by hazardous industries for 30 years as a risk assessment, accident investigation and design tool. It has given meaningful insight into human error in the nuclear, healthcare, rail, aviation, process industries, and offshore engineering. Yet the method was created at a time when there was no publicly-accessible Internet, no easily-searched databases and, in some cases only limited or, largely, non-existent Human Factors information. Since then some Error Producing Conditions (EPCs) have become much easier to quantify and a great deal more information has become available to check the underpinning methodology. Thousands of papers have been reviewed to identify new EPCs and to confirm, revise and refine some of the existing EPC multipliers to reflect new findings. Checks have also been made on the method's Generic Task Types (GTTs) and the stability or otherwise of the associated variability of these assessments. In many cases, the data underpinning the EPCs were found to be consistent with the original concepts and multipliers, but there were some cases where revision might be helpful. Crucially, while our work has highlighted that there are some important gaps in knowledge, enough is known to articulate clearly how to estimate the risk of human error.

Disclaimer: This publication and the work it describes were co-funded by the Health and Safety Executive (HSE) and Jerry Williams Human Factors Ltd. Its contents, including any opinions and/or conclusions expressed, are those of the authors alone and do not necessarily reflect HSE policy.

Introduction

The Human Error Assessment and Reduction Technique (HEART) was first proposed in 1985 (Williams, 1985). It is a human reliability assessment technique developed to help risk analysts identify the major influences on human performance and the likelihood of error, in a systematic and repeatable way. It is based on the general principle that for each task in life there is a basic probability of failure. Affecting each of these tasks are varying levels of Error Producing Conditions (EPCs) that can influence human reliability in systems operations. The method provides users with human reliability data that can be modified to be specific to their risks. It is a relatively quick, additive factors method that is straightforward to use and is applicable to any industry where human reliability is important. Despite some occasional suggestion to contrary, it is not a nuclear specific method.

The technique identifies nine generic task types (GTTs) and proposes nominal human unreliability values with suggested bounding values, together with 38 error producing conditions. Each EPC has an associated multiplier, which is the maximum amount by which the nominal unreliability might be expected to change depending on the proportion of the EPC applicable at the time of task execution. The EPCs and multipliers range from those that have a significant impact on performance such as being unfamiliar with a task (~17 times more likely to make an error) to those that have a very small impact such as evidence of ill health in the person performing the task (~1.2 time more likely to make an error). A method is provided for computing the assessed likelihood of error and its bounding values.

HEART has been in use for 30 years providing meaningful insight into human error in the healthcare, rail, aviation, nuclear, process and offshore engineering industries. Examples from the published literature include, the analysis of data entry errors in radiotherapy treatment (Chadwick and Fallon, 2012), the tailoring of the method for application in the rail industry (Gibson, et al, 2012), the quantification of risks associated with the transport of radioactive material (IAEA, 2003), and quantification of the likelihood of failure to perform a corrective task in a control room in order to prevent a hazardous event (Nespoli and Ditali, 2010).

In addition, several validation studies have been undertaken (e.g. Bye et al, 2010). These have concluded that HEART enables rapid assessments that broadly correlate with known data and produce insights that are equivalent to those derived from using more complicated methods.

Over the years, some researchers have raised questions about how to interpret some of the subtle terminology that is specific to the method and, as time has elapsed, concern has been raised about the age of data underpinning established HRA methods such as HEART. This paper relates to the work that has been undertaken to consolidate HEART, to add to the data that underpin the methodology by drawing on the last 30 years' Human Factors Literature and to offer greater insights into the application of the method. The bulk of the paper is about the results of work to consider the concept of EPCs, the appropriateness of each and their associated multiplier, but reference is also made to the GTTs and the underpinning method.

Methodology

The aim of the research was to draw together studies that used conceptually similar EPCs in order to generate an estimate of their impact on human performance.

An extensive literature review was undertaken to first identify studies that had an EPC as an identifiable factor in the experimental design. These papers were scrutinised to establish if the experimental sample size was at least 10 per study condition, that the participants were representative of the working population and, importantly, contained useable error data. Studies using reaction time and other measures of performance cannot be included in the quantification because they require assumptions to be made about influences on causality and therefore, introduce more uncertainty into the analysis. The strict research criterion significantly reduces the number of papers that can be included in the research, but this scientific rigour provides greater confidence in the results.

The methodology involved an assessment of some 25,000 papers, comprising, 2,500 papers published in Human Factors, 8,000 papers in the Proceedings of the Annual Meetings of the Human Factors and Ergonomics Society, 4,000 papers in Ergonomics, 1,000 abstracts in Applied Ergonomics, 2,000 papers in the Journal of Experimental Psychology: Learning, Memory & Cognition, 2,500 abstracts in Behaviour and Information Technology, and some 5,000 Internet sources.

In total, 180 papers met the strict evidentiary selection criterion and were analysed to establish the data points that could be extracted from each paper. The data points were collated into a spreadsheet and for each EPC, the proposed maximum multiplier was derived by considering how the data clustered and making an estimate of the worst case. The process involved using expert judgement to take proper account of apparently spurious data points and potential outliers, to form a justifiable position. This included, for example, making contact with authors to ask for clarification.

As part of the review, evidence was also collected that related to the GTTs (current and potential new ones), the verification, or otherwise, of the assumed log-normal distribution of the human error rates and the additive nature of how elements combine. An additional 160 papers were identified as having potential to add to the GTT knowledge base and although the GTTs look to be well-supported, the human error potential (HEP) distributions would benefit from further scrutiny and research application. These findings are the subject of other papers and will not be discussed in detail here.

Results

The analysis confirmed that 32 (out of 38) HEART original EPC concepts and multipliers appear fit for purpose (Williams and Bell, in press). In each case there is either new supporting evidence for, or there is no new information to challenge, the original conclusions. There was sufficient evidence to recommend that six EPCs be revised slightly and for two new ones to be incorporated into HEART.

Revised EPCs

Six EPCs were modified based on the results of the review. The following sections provide a summary of the basis for the authors' conclusions.

EPC 29. High level emotional stress (change from 1.3 to 2)

Although this factor is frequently referred to in the human reliability assessment literature as being a major factor in some assessments, only a few quantitative studies were identified from the last 30 years. The studies applied a range of 'stressors' and the results showed a difference in performance of approximately 2, between stressed and non-stressed individuals. For example, a paper by Leon and Revelle (1985) found that more anxious subjects had a higher mean error rate that was 2.2 times greater than less anxious subjects. Other studies reported findings in the range of 1.02 (Shackman et al, 2006) through to 1.7 (Tohill and Holyoak , 2000). Overall, the figures suggest an increase in the multiplier from 1.3 to a factor of 2 would be prudent, until further evidence becomes available.

EPC 32. Inconsistency of meaning of displays and procedures (change from 1.2 to 3)

Considering the importance of consistency in product and system design, and although much has been written, very few relevant human reliability data have been published over the last 30 years. In the literature, the emphasis is on automated techniques that have been developed to detect and minimise inconsistency of human action. The few studies that report user error, result in factor increases that are somewhat higher than was originally proposed. For example, Mendel (2009) reported errors that were increased by a factor of 2.2 and concluded that consistency may be especially important for difficult tasks with high cognitive load. Overall, the studies suggest that a multiplier of 3 is more appropriate than the original 1.2 but this EPC would benefit from further investigation.

EPC 35. Disruption of normal work-sleep cycles (change from 1.1 to 1.2)

With few exceptions, most studies report that fatigue induced by sleep loss has been found to reduce human reliability by an amount proportionate to the amount of sleep lost. On average, although there is up to 5% human reliability improvement every afternoon, overall human reliability has been found to reduce by approximately 20% for every 24 hours of sleep deprivation (Williams et al, 2007). Based on a considerable amount of research over the last 30 years (e.g. Wickens et al 2015), it is proposed that the multiplier be increased to x 1.2 (compound) for every 24 hours' sleep lost.

EPC 33. A poor or hostile environment (below 75% of health or life-threatening severity) (change from 1.15 to 2)

Since the original publication of HEART, changes in regulation (Council of the European Union, 1989) have made it less likely that individuals in Europe will be exposed to very harsh environments. Despite this, there has been research over the last 30 years to suggest that whilst the general magnitude of the multiplier might be in roughly the right region, an increase from 1.15 to 2 is appropriate. Mäkinen et al (2006), for example, studied performance over a ten day period at a 25 degree ambient and a 10 degree Celsius experimental condition and reported significant changes in accuracy; the multipliers that emerge from this is study were 1.9 and 1.4 respectively.

EPC 37. Additional team members over and above those necessary to perform task normally and satisfactorily (change from 1.03 to 1.2)

Two studies informed the original multiplier of 1.03 per additional person. From the last 30 years' literature, only one study (Rosenbloom et al, 2007) was found that is relevant to this EPC. This study on testing individuals separately and in pairs showed that the likelihood of driving test failure increased by a factor of 1.29 when an extra person was present. Bearing in mind the original multiplier was derived from two studies and that only one from the last 30 years has been found, it is prudent to retain this EPC and revise the multiplier to 1.2. This EPC does probably affect human reliability and it might be better informed by further research in due course.

EPC 38. Age of personnel performing recall, recognition and detection tasks (ages 25 to 85 years and mentally competent) (change from 1.02 to 1.16)

A considerable amount of research exploring the impact of age on performance has been published in the last 30 years. Data from the some 15 papers (e.g. Ardila et al, 2000, Kvavilashvili et al, 2005, Salthouse et al, 1996, and Whiting et al, 1997), indicate a consistent reduction in reliability that is associated with age and this change is relatively uniform across a wide range of ages. The papers suggest that personnel involved in performing detection, recall and recognition tasks could be approximately 15 -20% less reliable for every ten years age and that, within very narrow limits, this impact will be consistent over the range 25 to 85 years. While it might be argued that experience can 'eliminate' the impact of age on performance, Nunes and Kramer (2009) report that, "high levels of experience do not broadly provide immunity against the detrimental effects of advancing age on basic cognitive functioning". Based on this information, this EPC should be adjusted upward from 1.02 to 1.16 for every 10 years for ages 25 to 85 years (assuming the individuals are not suffering from dementia).

New Error Producing Conditions

In addition to those EPCs that have been revised, the research identified two EPCs that can now be quantified.

Distraction /Task Interruption

Distraction can be defined in a variety of ways, but the phenomenon is generally agreed to be about the drawing of attention away from something with the simultaneous possibility of confusion and increased potential for current task failure. Experiments that have been performed to investigate distraction have tended to be of two basic types, those that look at differences between interruptions and those that look at what happens to performance when attention is switched. Williams et al (2007) initially reviewed these two types of experiments as two distinct types of EPC, but it quickly became apparent that the impact on human reliability is largely the same, regardless of experimental type or method. In their review, Williams et al concluded the impact of distraction ranges from about a factor of two up to a factor of four. This multiplier is a little surprising because, intuitively, a major disruption of performance of perhaps several orders of magnitude might be expected. However, as no new information was identified in the current review to challenge the finding, the EPC Distraction will be included in the update of HEART. For the purposes of assessment, it is suggested that the upper limit be regarded as a factor of four.

Time-of-Day (from diurnal high arousal to diurnal low arousal)

Provisional investigations and assessment of the impact of time-of-day on human reliability suggest that the variation between the diurnal high and low of performance is about a factor of 2.4. The highest reliability is identified as being around 16:00 hrs and the lowest reliability, somewhat unsurprisingly, at around 03:00 hrs. The "neutral" point is at around 20:00 hrs. There is some evidence of a "post lunch dip" that contributes about a factor of 1.3 to overall human unreliability specifically at around 14:00 - 1500hrs. The data analysed and consolidated come from a range of sources, including laboratory, field and accident studies for example Mittler et al (1988) and Folkard (1997). Sufficient is now known that a new EPC and associated multiplier of 2.4 will be included in the update of HEART, in order to capture its potential contribution to human failure with respect to the time-of-day.

Discussion

When HEART was first developed, 30 years ago, it was recognised that EPCs can play a significant role in determining human reliability and so an effort was made to articulate them and assess their likely impact on reliability. At the time there were gaps in knowledge and it was anticipated that more research would be published in due course as the need to quantify such effects became apparent. Evidently, the EPC literature is now much larger and more accessible than in the mid '80s, but our research shows there are still substantial gaps in knowledge and areas where research with quantifiable results is lacking.

Despite the lack of quantifiable data, the research does show that significant performance patterns are apparent in the human factors literature. These patterns provide assurance that the EPCs outlined in HEART are valid and remain relevant for managing human performance in modern industry. If there are questions regarding the EPCs they are not about whether they have an impact on performance, or the relative impact of one EPC to another, but rather the size of that impact. To illustrate this point further, the qualitative research shows that somebody under time pressure more likely to make an error than somebody who is bored. Research to quantify of the likelihood of the error indicates that time pressure increases the likelihood of error by approximately 11 times and boredom increases the risk by approximately 1.4.

While 32 of the 38 EPCs are considered fit for purpose and are accepted to have an impact on human reliability, it would be beneficial if more specifically quantifiable research were undertaken. The update of HEART will include all of these EPCs with their original multiplier because they are of relevance and some increasingly so, in contemporary life. For example, EPC 18 "A conflict of short term and long term objectives" has not been researched in a systematic way yet the demands to achieve a 'quick fix' while putting off the main problem until a later date is a recognisable issue in commercial work.

EPC 28 "Little or no intrinsic meaning in a task" needs research because some jobs are becoming increasingly boring due to technological advances that reduce the need for human resourcefulness and action. Currently, the Human Factors community does not have sufficient quantitative information to be informed about the potential impact on human reliability. Stress is a cause of significant health issues, but while researchers know a great deal about the impact of stress on health, the lack of quantitative data means there is more uncertainty in operational risk assessments using human reliability methods. In addition, as the production burden on industry increases employees are required to respond to increasingly pressurised situations. The body of qualitative research could suggest that EPC 29 "High level emotional stress" needs a higher multiplier, but how much higher is not evident. Similarly, research into a lack of exercise (EPC 22) suggests that it may be important, although perhaps at a very low level, but we do not know whether it affects human reliability, and, if so, by how much. Research in this area tends to focus on performance improvement with exercise rather than any detrimental impact due to inactivity.

The limitations of the available data notwithstanding, the basic structure and underpinning data of the HEART method are well-supported by the literature and only minor adjustments are required to refine, support and extend its continued successful application. There is considerable strength in the HEART method because it draws together information from a vast range of studies for each EPC and this means they can be applied confidently to a variety of situations that are not task, context or person specific.

The body of research over the last 30 years has made it possible to consolidate HEART. This paper has summarised the efforts to confirm, revise and refine some of the EPC multipliers and to identify new EPCs; it has also highlighted research needs. Significant performance patterns are apparent, which give confidence that the EPCs are relevant and pertinent to human reliability in modern industry, but there are gaps in our knowledge that need to be addressed. The next steps for the HEART project team are to combine and refine the research data into an updated version of HEART with an accompanying manual. The aim is to provide sufficient background information, practical guidance, an explanation of the terminology and appropriate industry examples that will help assessors make use of HEART for the next 30 years.

References

Ardila, A., Ostrosky-Solis, F., Rosselli, M. & Gómez, C. (2000). Age-related cognitive decline during normal aging: the complex effect of education. *Archives of Clinical Neuropsychology*, 15(6), 495-513.

Bye, A., Lois, E., Dang, V.N., Parry, G., Forester, J., Massaiu, S., Boring, R., Braarud, P.Ø., Broberg, H., Julius, J., Männistö, I. & Nelson, P. (2011). *International HRA Empirical Study – Phase 2 Report: Results from Comparing HRA Method Predictions to Simulator Data from SGTR Scenarios. NUREG/IA-0216, Vol. 2*, Washington D.C.: US Nuclear Regulatory Commission.

Chadwick, L. and Fallon, E.F. (2012). Human reliability assessment of a critical nursing task in a radiotherapy treatment process. *Applied Ergonomics*, 43(1), 89-97.

Council of the European Union (1989). *Council Directive of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work (89/391/EEC*), Brussels.

Folkard S. (1997). Black times: Temporal determinants of transport safety. *Accident Analysis and Prevention*, 29(4), 417–430.

Gibson. W, Dennis. C, Thompson, Mills. A and Kirwan. B. (2012). Tailoring the HEART technique for application in the rail industry. *Advances in Safety, Reliability and Risk Management* – Bérenguer, Grall & Guedes Soares (eds) Taylor & Francis Group, London.

IAEA (2003). Input data for quantifying risks associated with the transport of radioactive material. *International Atomic Energy Agency, IAEA-TECDOC-1346*, Vienna

Kvavilashvili, L., Kornbrot, D. E., Mash, V., Cockburn, J. & Milne, A. (2009). Differential effects of age on prospective and retrospective memory tasks in young, young-old, and old-old adults. *Memory*, 17(2), 180–196.

Mäkinen, T.M., Palinkas, L.A., Reeves, D L., Pääkkönen, T., Rintamäki, H., Leppäluoto, J. & Hassi, J. (2006). Effect of repeated exposures to cold on cognitive performance in humans. *Physiology & Behavior*, 87(1), 166-176.

Mendel, J. (2009). *The effect of interface consistency and cognitive load on user performance in an information search task*. Clemson University, South Carolina, MS thesis.

Mittler, M.M., Carskadon, M.A., Czeisler, C.A., Dement, W.C., Dinges, D.F. & Graeber, R.C. (1988). Catastrophes, Sleep, and Public Policy: Consensus Report, *Sleep*, 11(1), 100–109.

Nespoli, C and Ditali, S (2010). Human Error Probability Estimation for process risk assessment with emphasis on control room operations. *Chemical Engineering Transactions* Vol. 19, 219-224.

Nunes, A. & Kramer, A.F. (2009). Experience-based mitigation of age-related performance declines: Evidence from air traffic control. *Journal of Experimental Psychology*, 15(1), 12-24.

Rosenbloom, T., Shahar, A., Perlman, A., Estreich, D. & Kirzner, E. (2007). Success on a practical driver's license test with and without the presence of another testee. *Accident Analysis & Prevention*, *39*(6), 1296-1301

Salthouse, T. A., Hambrick, D. Z., Lukas, K. E., & Dell, T. C. (1996). Determinants of adult age differences on synthetic work performance. *Journal of Experimental Psychology: Applied*, 2(4), 305-329.

Shackman, A. J., Sarinopoulos, I., Maxwell, J.S., Lavric, A., Pizzagalli, D.A. & Davidson, R.J. (2006). Anxiety Selectively Disrupts Visuospatial Working Memory. *Emotion* 6(1), 40–61.

Tohill, J.M. & Holyoak, K.J. (2000). The impact of anxiety on analogical reasoning. *Thinking & Reasoning*, *6*(1), 27-40.

Whiting, W. L. & Smith, A. D. (1997). Differential age-related processing limitations in recall and recognition tasks. *Psychology and Aging*, 12, 216-224.

Wickens, C.D., Hutchins, S.D., Laux, L. and Sebok, A. (2015). The Impact of Sleep Disruption on Complex Cognitive Tasks A Meta-Analysis. *Human Factors*, 57(6), 930–946.

Williams, J.C. (1985). HEART- A proposed method for achieving high reliability in process operation by means of human factors engineering technology. In *Proceedings of a Symposium on the Achievement of Reliability in Operating Plant*, Safety and Reliability Society (Vol. 16), pp.5/1-5/15.

Williams, J.C., Bell, J.L. & Daniels, C.L. (2007). The Quantitative Impact of Fatigue, Distraction and Age on Human Reliability. In Aven & Vinnem (eds) *Risk, Reliability and Societal Safety – Proceedings of ESREL 2007*. pp.2053-2058.

Williams, J.C. and Bell, J.L. (in press). Consolidation of the error producing conditions used in the Human Error Assessment and Reduction Technique (HEART). Safety and Reliability Society Journal.