Predicting Design-Induced Error in the Cockpit *

Neville A. Stanton **,1, Don Harris[2], Paul M. Salmon[3], Jason Demagalski[4], Andrew Marshall[5], Thomas Waldmann[6], Sidney Dekker[7], and Mark S. Young[8]

[1]*Transportation Research Group, School of Civil Engineering and the Environment, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.*
[2]*Department of Systems Engineering and Human Factors, School of Engineering, Cranfield University, Cranfield, Bedford, MK43 0AL, UK.*
[3]*Monash University Accident Research Centre, Building 70, Clayton Campus, VIC 3800, Australia*
[4]*National Air Traffic Services, Corporate and Technical Centre, 4000 Parkway, Whiteley, Fareham, Hampshire PO15 7FL, UK.*
[5]*Marshall Ergonomics Ltd, 25 Third Avenue, Havant, Hampshire, PO9 2QR*
[6]*College of Engineering, University of Limerick, Ireland*
[7]*School of Aviation, Lund University, SE 260 70 Ljungbyhed, Sweden*
[8]*School of Engineering and Design, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK.*

ABSTRACT

This paper describes the Human Error Identification (HEI) Technique called the Human Error Template (HET). HET has been developed specifically for the aerospace industry in response to Certification Specification (CS) 25.1302. In particular, it is intended as an aid for the early identification of design-induced errors, and as a formal method to demonstrate the inclusion of human factors issues in the design and certification process of aircraft flight-decks, including supplemental type certification. The template-based approach was chosen because it appeared to be quick to learn and easy to use. HET uses a hierarchical task analysis as its starting point. A checklist of twelve (12) external error modes is used to determine which might lead to credible errors for each task step. For each credible error a description is given and the outcome described. If the likelihood of the error and the consequences are both high then that task step is rated as a 'Fail'. The error mode taxonomy developed comprises: fail to execute a task, task execution incomplete, in the wrong direction, wrong task executed, task repeated, on the wrong interface element, too early, too late, too much, too little, misread information, and other. HET was then compared to SHERPA, HAZOP and HEIST. Thirty seven (37) analysts were employed in this study based on a landing scenario. HET showed significantly better Sensitivity Index scores than any of the other methods, and the greatest number of correct error predictions (hits). The results from the HET validation study demonstrate that HET meets all the criteria set. It is easy to learn, the error taxonomy has been specifically designed for flight-deck tasks, it is auditable, and it has been proved to be both reliable and valid. HET is recommended for use in the design, evaluation and certification of aircraft flight-decks.

Keywords: Human error identification, Cockpit design, Certification, Validation

* Manuscript received, Sep. 23, 2009, final revision, Nov. 19, 2009
** To whom correspondence should be addressed, E-mail: n.stanton@soton.ac.uk

I. INTRODUCTION TO AIRCRAFT SAFETY

Commercial aviation is without doubt one of the safest forms of passenger travel. For the majority of the past half-century there has been a steady decline in the commercial aircraft accident rate. However, over the last two decades it has been noticeable that the serious accident rate has remained relatively constant at approximately one per million departures (Boeing Commercial Airplanes Group, 2000). If this rate remains unchanged, with the current projected increase in the demand for air travel (assuming that the market eventually recovers after recent world events) this will mean that there will be one major hull loss almost every week by the year 2015.

Initial efforts to enhance aircraft safety were aimed at system reliability, structural integrity and aircraft dynamics. The airworthiness regulations governing the design of commercial aircraft, for example Joint Airworthiness Requirements (JAR) part 25: Large Aeroplanes (UK Civil Aviation Authority, 1978) and Federal Aviation Regulation (FAR) part 25: Airworthiness Standards (US Department of Transportation, 1974) still reflect these earlier concerns. As reliability and structural integrity have improved over the last 50 years, the number of accidents resulting from such failures has reduced dramatically and hence, so has the overall number of accidents and accident rate. What this has meant, though, is that up to 75% of all aircraft accidents have a human factors component in them. Human error is now the primary risk to flight safety (Civil Aviation Authority, 1998). It would appear that the human component is now the most ‘unpredictable’ component in the system.

There is nothing particularly new about human error in the cockpit. Chapanis (1999) recalls his work at the Aero Medical Laboratory in the early 1940's where he investigated the problem of pilots and co-pilots retracting the landing gear instead of the landing flaps after landing. His investigations in the B-17 (known as the 'Flying Fortress') revealed that the toggle switches for the landing gear and the landing flaps were both identical and next to each other. Chapanis's insight into human performance enabled him to understand how the pilot might have confused the two toggle switches, particularly after the stresses of a combat mission. He proposed coding solutions to the problem: separate the switches (spatial coding) and/or shape the switches to represent the part they control (shape coding), so the landing flap switch resembles a 'flap' and the landing gear switch resembles a 'wheel'. Thus the pilot can tell by looking at, or touching, the switch what function it controls. In his book, Chapanis (1999) also proposed that the landing gear switch could be deactivated if sensors on the landing struts detect the weight of the aircraft.

Grether (1949) reports on the difficulties of reading the traditional three needle altimeter which displays the height of the aircraft in three ranges: the longest needle indicates 100s of feet, the broad pointer indicates 1000s of feet and the small pointer indicates 10,000s of feet. Previous work had shown that pilots frequently misread the altimeter. This error had been attributed to numerous fatal and non-fatal accidents. Grether devised an experiment to see if different designs of altimeter could have an effect on the interpretation time and the error rate. If misreading altimeters was really was a case of 'designer error' rather than 'pilot error' then different designs should reveal different error rates. Grether tested six different variations of the dial and needle altimeter containing combinations of three, two and one needles with and without inset counter as well as three types of vertically moving scale (similar to a digital display). Pilots were asked to record the altimeter reading. The results of the experiment showed that there were marked differences in the error rates for the different designs of the altimeters. The data also show that those displays that took longer to interpret also produced more errors. The traditional three-needle altimeter took some 7 seconds to interpret and produced over 11 percent errors of 1,000 feet or more. By way of contrast, the vertically moving scale altimeters took less than 2 seconds to interpret and produced less than 1 percent errors of 1,000 feet or more.

Both of these examples, one from control design and one from display design, suggest that it is not 'pilot error' that causes accidents; rather it is 'designer error', i.e., poor representation of system information output to the pilot and confusing system input devices. This notion of putting the blame on the last person in the accident chain (e.g., the pilot), has lost credibility in modern aviation research. Modern day researchers take a systems view of error, by understanding the relationships between all the moving parts in a system, both human and technical, from concept, to design, to manufacture, to operation and maintenance (including system mid-life upgrades) and finally to dismantling and disposal of the system. What is new here is the assertion that design-induced errors may be predicted in advance of the aircraft becoming operational (Stanton and Baber, 1996, 2002).

The traditional approach to certification has been referred to as the ‘system-by-system’ method whereby safety was achieved by ensuring that each system complied with the certification requirements (see Applegate and Graeber, 2001). This approach cannot consider the flight-deck as an integrated whole, and it has to be emphasised that modern commercial airliners now have extremely complex and highly integrated system architectures. The ‘system-by-system’ engineering approach to human factors certification is also inappropriate as human factors engineers design on a ‘task-by-task’ basis, which implicitly crosses the boundaries of many systems, because pilots interact with several systems when performing many flight-related tasks. As a result, inconsistencies in interfaces are much more obvious to them. Many human factors problems lie not within an individual system (hence also within a single regulation) but between systems (regulations). The regulations do not treat the flight-deck as a harmonious, integrated whole. One of the great challenges for the proposed human factors certification of flight-decks is reconciling the conflicts between the ‘traditional’ engineering approach to design

and the human factors engineering approach to design. Should the new human factors regulations reflect the task-based, human-centred, design philosophy, or does this make the whole process too difficult by being at odds with the vast majority of the rest of the certification process?

Since September 2007 the rules and advisory material developed from the output of the HF HWG have been adopted by EASA as Certification Specification (CS) 25.1302 and with supporting advisory material in AMC (Acceptable Means of Compliance) 25.1302 (see also Harris, this volume). This rule applies to the Type Certification and Supplemental Type Certification processes for large transport aircraft certificated in the European Union (see EASA Certification Specification 25, http://www.easa.europa.eu/ws_prod/g/rg_certspecs.php#CS-25, Amendment 5, September 2008). Of particular relevance to this paper is section 'd', which states that:

"To the extent practicable, installed equipment must enable the flight crew to manage errors resulting from the kinds of flight crew interactions with the equipment that can be reasonably expected in service, assuming the flight crew is acting in good faith. This sub-paragraph (d) does not apply to skill-related errors associated with manual control of the aeroplane."

Ways of anticipating which design-induced errors may be manifest in the cockpit are considered in the next section.

II. HUMAN ERROR IDENTIFICATION

Validation of human error identification (HEI) methods remains a huge problem in the area of human error prediction. However, although there are very few such studies, a number of HEI method validation and comparison studies have been conducted in the past (Williams, 1985; Whalley and Kirwan, 1989; Kirwan, 1992a; 1992b; 1998a; 1998b, Kennedy, 1995; Stanton and Baber, 1996, 2002, Stanton and Stevenage, 1998, 2000). The purpose of these studies is to firstly validate the methods in question by confirming that they actually have a degree of accuracy in predicting potential human error and secondly, to see which of the methods analysed is the most successful in terms of accuracy of error predictions made. Reliability (the degree to which the methods produce the same error predictions over time and with different analysts) of the methods is also a concern and is also frequently tested in such studies. Whalley and Kirwan (1989) evaluated six HEI methods (Heuristics, PHECA, SRK, SHERPA, THERP and HAZOP) for their ability to predict the errors responsible for four actual incidents that had occurred in the nuclear industry. Similarly, Kennedy (1995) examined the ability of a number of HEI methods to retrospectively predict the underlying errors in ten actual disasters, such as the Trident Papa-India air disaster and the Three Mile Island disaster. Kennedy concluded that there was no single method that was universally applicable across different systems and suggested that the best approach was to use a combination of the available methods and analyst interpretation. Kirwan (1992b) developed a list of eight criteria to evaluate twelve HEI techniques (THERP, Human Error HAZOP, SRK approach, SHERPA, GEMS, PHECA, Murphy diagrams, CADA, HRMS, IMAS, Confusion Matrices and CES). In conclusion, Kirwan (1992b) recommended a combination of expert judgement and the SHERPA technique as the most valid approach to error identification. In a more recent comparative study, Kirwan (1998b) used fourteen criteria to compare 38 HEI methods. In conclusion, it was reported that, of the 38 methods, only nine are available in the public domain and are of practical use (Kirwan, 1998). The nine methods were THERP, Human Error HAZOP, SHERPA, CMA/FSMA, PRMA, EOCA, SRS-HRA, SRK and HRMS. Baber and Stanton (1996) tested the predictive validity of SHERPA and TAFEI when used to predict London Underground rail ticket machine errors. It was concluded that both SHERPA and TAFEI provided an acceptable level of sensitivity based upon the data from two expert analysts (Baber and Stanton, 1996). Stanton and Stevenage (1998) also compared a heuristic approach and the SHERPA methodology concerning the prediction of errors made when using a vending machine and reported that SHERPA provided a better means of predicting errors than the heuristic approach did. Moreover, it was reported that SHERPA returned a mean sensitivity index (SI) of 0.76 at time 1, 0.74 at time 2 and 0.73 at time 3, which are very acceptable levels of concurrent validity. Furthermore Stanton and Baber (2002) reported reliability values for the SHERPA methodology of between 0.4 and 0.6 and sensitivity values between 0.7 and 0.8, which is accepted as being high. It is apparent from the literature that SHERPA is the most efficient and reliable HEI method available to human factors professionals. Furthermore a literature review of existing HEI methods conducted by the authors revealed that of 32 available HEI methods, SHERPA, Human Error HAZOP and HEIST were the most suitable for use on the flight-deck. As a result of this review, the three HEI methods were selected for use in this comparative study.

III. HUMAN ERROR TEMPLATE (HET)

HET is a newly developed HEI methodology, developed by the authors, aimed specifically at predicting design-induced pilot error on civil flight-decks (Stanton et al, 2006, 2009). The method is a checklist approach and comes in the form of an error template. HET works as a simple checklist and is applied to each bottom level task step in a hierarchical task analysis (HTA) of the task under analysis, as illustrated in the flowchart shown in figure 1.

The HET technique works by indicating which of the HET error modes are credible for each task step, based upon the judgement of the analyst. The analyst simply applies each of the HET error modes to the task step in question and determines whether any of the modes produce any credible errors or not. The HET error taxonomy consists of twelve error modes that were

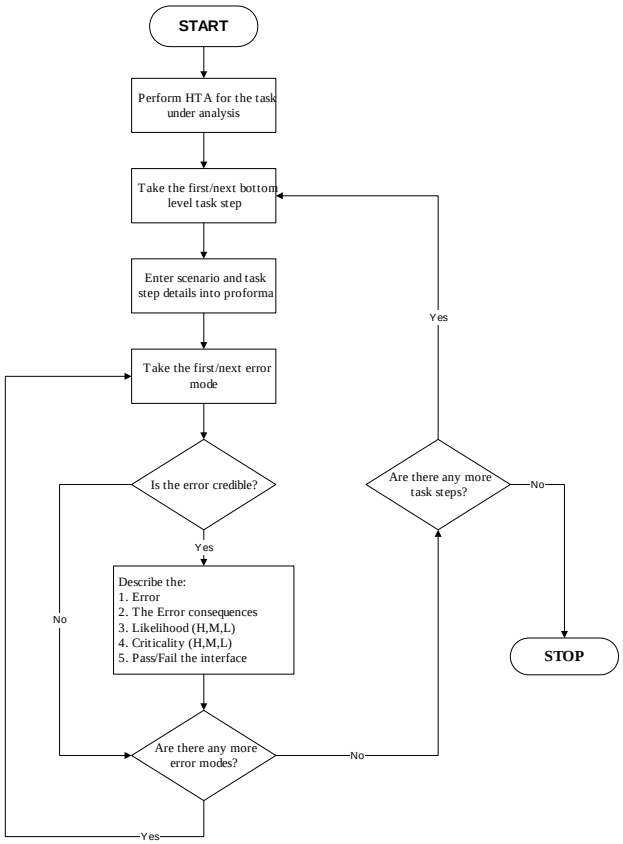

Figure 1 The HET method shown in a flowchart

selected based upon a study of actual pilot error incidence and existing error modes used in contemporary HEI methods. The twelve HET error modes are shown below:

- Fail to execute
- Task execution incomplete
- Task executed in the wrong direction
- Wrong task executed
- Task repeated
- Task executed on the wrong interface element
- Task executed too early
- Task executed too late
- Task executed too much
- Task executed too little
- Misread Information
- Other

For each credible error (i.e. those judged by the analyst to be possible) the analyst should give a description of the form that the error would take, such as, 'pilot dials in the airspeed value using the wrong knob'. Next, the analyst has to determine the outcome or consequence associated with the error e.g. Aircraft stays at current speed and does not slow down for approach. Finally, the analyst then has to determine the likelihood of the error (low, medium or high) and the criticality of the error (low, medium or high). If the error is given a high rating for both likelihood and criticality, the aspect of the interface involved in the task step is then rated as a 'fail', meaning that it is not suitable for certification. An example of a HET output is shown in figure 2 on the next page. The main advantages of the HET method are that it is simple to learn and use, requiring very little training and it is also designed to be a very quick method to use. The error taxonomy used is also comprehensive; it was based on existing error taxonomies from a large number of HEI methods. The HET method is also easily auditable as it comes in the form of an error proforma. The only real disadvantage associated with HET is that for large tasks, it may become laborious to perform. This however is a disadvantage associated with all HEI methods and one that the authors feel cannot be avoided.

IV. METHODOLOGY

4.1 Participants

A total of 37 participants were involved in this study. These participants were divided into four groups based upon the four different HEI methods. The first group consisted of eight undergraduate students aged between 21 and 23 years old. All participants in this group were male. These participants formed the HET group and received training in the HET method. The second group consisted of nine undergraduate students aged between 21 and 23 years old. Of these six were male and three were female. These participants formed the SHERPA group and received training in the SHERPA method. The third group consisted of nine undergraduate students aged between 21 and 23 years old. Of these seven were male and two were female. These participants formed the Human Error HAZOP group and received training in the Human Error HAZOP method. The fourth and final group consisted of eleven undergraduate students aged between 21 and 23 years old. Of these, eight were male and three were female. These participants formed the HEIST group and received training in the HEIST method. All participants had no previous experience of any HEI methods.

4.2 Materials

All participants were supplied with a training package for the methodology in question. These training packages consisted of a description of the method, a copy of the methods associated error taxonomy, a flowchart showing how to conduct an analysis using the method, an example output of the method and also an example of an analysis carried out using the method. Participants were also given a HTA describing the action stages involved when using a simple machine and also a HTA describing the action stages involved when landing an Aircraft A at New Orleans using the Auto-land system. The participants were also provided with photographs of all flight-deck instrumentation used in the flight task i.e. flap lever, throttle lever, auto-pilot panel, captains primary flight display, landing gear lever and the captain's navigation display. All participants were also provided with suitable proformae for recording their error predictions. Microsoft Flight Simulator 2000 Professional Edition was also used to give the participants a demonstration of the flight task under

analysis.

4.3 Design

A between-subjects design was used in this study. The independent variables were the four different participant groups, the HET group, HAZOP group, HEIST group and SHERPA groups. The dependent variables were the errors predicted by each participant and time taken by each participant to conduct the analysis.

4.4 Procedure

Participants were recruited via sending e-mail advertisements to all undergraduate students. Subjects were recruited into four separate groups.

For each group, participants were initially given a short briefing on the purpose of the experiment and also an overview of the project, "Prediction of Human Error on Civil Flight-decks."

Participants were then given an introduction to the areas of Human Error and Human Error Identification.

Participants were then given a short training session on the method that their particular group were being tested on. This training session consisted of two parts. Firstly, the participants were given a short introduction to the method, which involved explaining why the method was developed, what the method does and how the method works. Secondly, participants were taken step by step through an example of an analysis using the method in question.

Once the participants were comfortable with the method and how the method worked, they were given a HTA of a simple task to analyse. After being given a demonstration of the task and a walk through of the HTA, participants were then required to make error predictions for the task with the method that they had been trained in. Participants were also provided with A3 photographs of the machine they were analysing and its user interface. Error proformae were also provided. At this stage, participants were permitted to confer with other participants. Questions regarding the analysis were also permitted. After the participants had finished

Scenario:Land Aircraft A at New Orleans using the Autoland system			Task step:3.4.2 Dial the 'Speed/MACH' knob to enter 150 on IAS/MACH display									
Error Mode	✓	Description	Outcome	Likelihood			Criticality			PASS	FAIL
				H	M	L	H	M	L		
Fail to execute											
Task execution incomplete											
Task executed in wrong direction	✓	Pilot turns the Speed/MACH knob the wrong way	Plane speeds up instead of slowing down		✓		✓			✓	
Wrong task executed											
Task repeated											
Task executed on wrong interface element	✓	Pilot dials using the HDG knob instead	Plane changes course and not speed	✓			✓				✓
Task executed too early											
Task executed too late											
Task executed too much	✓	Pilot turns the Speed/MACH knob too much	Plane slows down too much		✓			✓		✓	
Task executed too little	✓	Pilot turns the Speed/MACH knob too little	Plane does not slow down enough/Too fast for approach		✓			✓		✓	
Misread information											
Other											

Figure 2 Example of HET output

conducting the analysis, they were taken through an 'expert' analysis for the same task, in order to demonstrate the correct results for the human error identification.

After a short break, participants were then given a HTA for the landing task, 'Land Aircraft A at New Orleans using the Auto-land system'. After an initial walk through of the landing task, participants were then given a step-by-step demonstration of the landing task using Microsoft Flight Simulator 2000 Professional Edition. Once all of the participants were familiar with all of the different tasks involved within the landing task, they were given colour photographs of all of the relevant flight-deck instruments e.g. flap lever, throttle lever, auto-pilot panel, captain's primary flight display, landing gear lever and the captain's navigation display. The participants were then asked to predict any potential design-induced pilot errors for the flight task. Suitable error proformae were provided and each participants start and finish time were recorded. For reliability purposes, the participants returned four weeks later to conduct the same procedure again.

4.5 Data Reduction

To compute validity statistics, the error predictions made by each subject were compared with error incidence data reported by pilots using the auto-land system for the flight task under analysis. The error predictions from all participants were compared to actual errors reported in a questionnaire based upon the tasks involved in the landing task, 'Land at New Orleans using the Auto-Land system'. Pilots were asked to report any errors that either they had made or they had seen being made by a co-pilot, for each of the task steps in the HTA, 'Land at New Orleans using the Auto-Land system' The sensitivity of or accuracy of each participants error predictions was calculated using the Signal Detection Paradigm. The signal detection paradigm was used as it has been found to provide a useful framework for testing the power of HEI techniques and has been used effectively for this purpose in the past (Stanton and Stevenage 2000). The signal detection paradigm divides the data into the following mutually exclusive categories:

Hit – Predicted error that was reported by the pilots

Miss – Failure to predict an error that was reported by the pilots

False Alarm – Predicted error that was not reported by the pilots

Correct rejections – Correctly rejected error that was not reported by the pilots

These four categories were entered into the signal detection grid for each subject (see figure 3). The signal detection paradigm can be used to calculate the sensitivity index (SI). This provides a value between 0 and 1, the closer that SI is to 1, the more accurate the techniques predictions are. The formula used to calculate SI is shown below in figure 4 (from Stanton and Stevenage, 1998).

		Errors Reported	
		YES	NO
Errors Predicted	YES	HIT	FALSE ALARM
	NO	MISS	CORRECT REJECTION

Figure 3 Signal Detection matrix used to determine the frequency of hits, misses, false alarms and correct rejections

$$Si = \left(\frac{\left(\frac{Hit}{Hit + Miss} \right) + 1 - \left(\frac{False\ Alarm}{FA + Correct\ Rejection} \right)}{2} \right)$$

Figure 4 Sensitivity Index formula

V. RESULTS

To find out whether the observed differences in the sensitivity index (observed in figure 5) were greater than those expected by chance, the Kruskal-Wallis One-Way analysis of variance test was undertaken. The difference in the sensitivity index between the four methods was statistically significant (Chi-Square (3df) = 29.2257, $p<0.0001$). This means that there is a real difference in the sensitivity index for the four HEI methods. To explore differences between pairs of methods the Mann-Whitney U test was used. The sensitivity index for the prototype group was significantly higher than the SI for the SHERPA group ($U=19$, $p<0.0001$). The sensitivity index for the prototype group was significantly higher than the SI for the Human Error HAZOP group ($U=19$, $p<0.0001$). The sensitivity index for the prototype group was significantly higher than the SI for the HEIST group ($U=19$, $p<0.0001$). This means that participants using the HET methodology were significantly more accurate in their predictions than those participants using any of the other methods. Furthermore there were no statistically significant differences between the remaining comparisons of the methods, i.e. no difference between SHERPA and HIEST ($U=155$, p=NS), no difference between SHEPRA and HAZOP (U125, p=NS), and no difference between HAZOP and HEIST ($U=137$, p=NS).

To determine whether or not there was any statistically significant difference between the participant SI scores, hit rate and false alarm rate, at time 1 and time 2, a Wilcoxon Matched Pairs Signed Ranks test was used. It was found that there was no statistically significant difference between the participants SI scores at time 1 and time 2 ($Z=-1.2737$, 2-Tailed $p=.2028$). This found that there was a statistically significant difference between the Hit Rate scores at time 1 and time 2 ($Z=-2.2567$, 2-Tailed $p= .0240$). The participant hit rate scores were statistically significantly higher at time 2 as

shown in figure 6. This found that there was a statistically significant difference between the False Alarm Rate scores at time 1 and time 2 (Z=-2.3166, 2-Tailed p=.0205). The participant false alarm scores were statistically significantly higher at time 2 than at time 1 as shown in figure 7. A summary of the sensitive index, hit rate and false alarm rate for all of the methods over time 1 and time 2 is shown in figure 8. To determine whether or not there was any difference between the time taken for each participant to complete the analysis at time 1 and at time 2, a t-test for Paired Samples was conducted. This demonstrated that there was a statistically significant difference between the time taken for the analysis at time 1 and time 2 ($t33=9.7$, $p<0.001$). This means that the time taken to perform the analysis at time 1 was significantly longer than the time taken to perform the analysis at time 2.

VI. GENERAL DISCUSSION

The aim of this study was to demonstrate that participants using the newly developed HET methodology would be more accurate at predicting potential design-induced pilot error on a landing task than participants using three contemporary HEI methods (SHERPA, Human Error HAZOP and HEIST). The study also aimed to demonstrate that participant SI scores, hit rate scores, false alarm rate scores and also time taken to complete the error analysis would improve significantly at when the analysts perform the same analysis for a second time (Stanton et al, 2006, 2009). In terms of accuracy of error predictions, participants using the HET methodology were the most accurate in their error predictions for the flight task under analysis. This finding supports the original hypothesis that the HET methodology would be the most successful at predicting potential design-induced pilot error for a given flight task. As the HET error mode taxonomy was developed from actual pilot error incidences and from an exhaustive analysis of contemporary approaches to human error identification, it is the most appropriate for use on civil flight-decks. The other methods used (SHERPA, Human Error HAZOP and HEIST) suffer in that they utilise error mode taxonomies that were developed specifically for nuclear power plant control room tasks (Kirwan, 1994). It is apparent that the differences in performance of the four methods are due to the constraints imposed on the possible errors that can be predicted by the error mode taxonomies used by the methods. The possible errors that can be predicted by each method are determined by HET's error mode checklist, SHERPA's behaviour and error mode taxonomy, Human Error HAZOP's guidewords and by HEIST's error identifier questions. For example, the guidewords used in the Human Error HAZOP methodology do not allow the analyst to predict an error such as, "Pilot enters airspeed using the heading knob instead of the speed/Mach knob", i.e. pilot presses wrong button. This was one of the actual errors reported by pilots in the original questionnaire. The HET checklist error taxonomy, however, prompts the analyst for this error, with the error

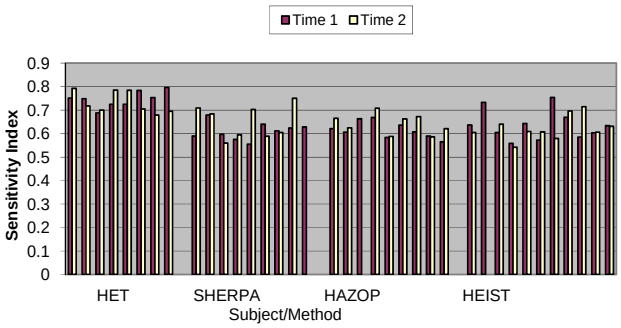

Figure 5 Bar graph showing subjects sensitivity index (SI) scores for time 1 and time 2

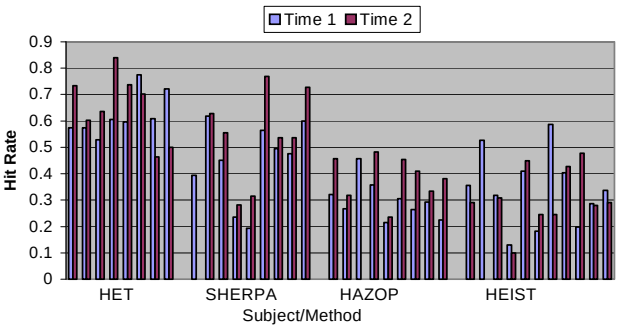

Figure 6 Bar graph showing subjects' Hit Rate scores for time 1 and time 2

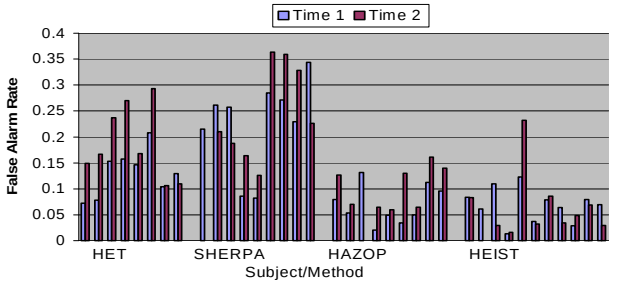

Figure 7 Bar graph showing subjects false alarm rate scores for time 1 and time 2

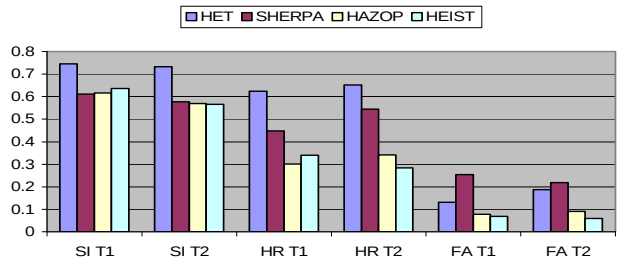

Figure 8 Bar graph showing mean Sensitivity Index, Hit Rate and False Alarm Rate for each method

mode 'Task executed on wrong interface element. Furthermore, it is also suggested that as the HET methodology is a very simple to learn and use checklist type approach, participants using HET were able to pick the method up easier than participants using the other three methods. Of the other three methods, SHERPA, Human Error HAZOP and HEIST, there were no statistically significant differences between the accuracy of error predictions of each of the methods.

It was also expected that the SI scores for each subject would improve from time 1 to time 2, highlighting a learnability effect on participant performance (Stanton and Stevenage, 1998, 2000). However, the results demonstrated that there was no statistically significant difference between the participant SI scores at time 1 and time 2. This is surprising, as it is generally the case that analysts become more efficient at predicting errors the more they use the HEI methods. Further analysis of the results revealed that although hit rate scores were found to statistically significantly increase at time 2 (i.e. participants were predicting more hits and less misses) it was also found that false alarm rate scores also increased statistically significantly at time 2 (i.e. participants were predicting more false alarms and making less correct rejections). This therefore meant that the negative effect of the higher false alarm rates counteracted the positive effect of the higher hit rate scores at time 2. As a result of this, the SI scores did not change significantly at time 2. If the false alarm rate scores had decreased at time 2 along with the increase of the hit rate scores, then the expected improvement in SI scores would have been observed. As this was not the case, and false alarm rate scores actually increased (i.e., it got worse) the expected overall SI improvement was not observed. The participants were improving at predicting more of the actual errors that were reported by the pilots (hits). The problem then, was that participants were predicting significantly more errors that were not reported by the pilots (false alarms) and thus making less correct rejections. This could possibly be the result of the participants becoming more comfortable with the methods and error prediction in general and thus becoming overconfident and predicting more errors than they should be. It may, however, be speculated that a false alarm is an error waiting to happen (Kennedy, 1995). It would be foolish to dismiss the possibility of an error occurring in the real world simply because it had not occurred as yet. Furthermore, as the sample size for the questionnaire study was quite small (i.e., the sample size being only 46 pilots), it could also be surmised that not all of the errors that have occurred, have been reported. This suggests that 'false alarms', as they are currently defined, should be treated with caution. In any case, the SI scores should be treated as conservative estimates, on the basis that the observed error dataset is limited by those sample of pilots questioned.

Time taken to complete the analysis was another dependent variable. It was anticipated that the time taken by the subjects to complete the error analysis would decrease significantly at time 2. The results show that the time taken by the participants did decrease statistically significantly at time 2. This indicates that the subjects were becoming more familiar with the methods in question and also more efficient at using the them (Stanton and Stevenage, 1998; Stanton and Baber, 2002). Another factor involved in this time decrease could be that the participants were becoming more confident at using the methods and also in their error own prediction ability.

VII. SUMMARY AND CONCLUSIONS

In conclusion, the participants using the HET methodology were the most accurate in their error predictions for the landing task, 'Land at New Orleans airport using the autoland system'. The SI scores for the HET methodology were higher than the three other contemporary HEI methods, SHERPA, Human Error HAZOP and HEIST. Of the three contemporary methods, there was no difference in the accuracy of the participants' error predictions. It can therefore be tentatively concluded that of the four HEI methods, the HET methodology was the most successful for use on civil flight-decks. Further conclusions were that the hit rate scores for each of the methods increased significantly at time 2 and that the false alarm rate scores for each methodology also increased significantly at time 2. Time taken to perform the analysis also decreased significantly at time 2. The main objective of the research was to produce a methodology to predict design-induced errors on aircraft flight-decks during design and certification. The criteria identified for the methodology were that it should be:

- easy to learn for non-human factors professionals,
- developed specifically for aviation industry,
- easily auditable,
- suitable for the current FAA/JAR certification procedure, and
- proven to be both reliable and valid.

It has been shown that students can learn how to apply it in 90 minutes, the error taxonomy has been developed specifically for the flight-deck, the completed error proformae are easily interpreted and form a permanent record. The reliability and validity data from the pilot study showed it to be better than current techniques. These data and those from the HET validation show strong support for use of HET. It is the opinion of the authors that HET will be accepted by the regulating authorities as evidence of a formal human factors design error analysis.

ACKNOWLEDGEMENTS

This research was supported in part by a grant from the Department of Trade and Industry as part of the European EUREKA! research programme. The authors are indebted to the help given to the project by staff of the DTI, and in particular to Richard Harrison, Anne Taylor, John Brumwell, Richard Pitman and Gill Richards.

The authors would also like to acknowledge the

assistance of the airline industry and in particular those pilots that filled in questionnaires and provided other help, without whom this project would not have been possible.

Finally, the research team (i.e., the authors of this paper) were awarded the Hodgson Medal and Bronze Award by The Royal Aeronautical Society for their work on flight-deck safety (based on the research published in The Aeronautical Journal, 110 (2), 107-115, 2006).

REFERENCES

[1] Annett, J., Duncan, K. D., Stammers, R. B & Gray, M. J. Task Analysis. Training Information No.6. HMSO: London, 1971.

[2] Applegate, J. D. and Graeber, R. C., "Integrated safety systems design and human factors considerations for jet transport aeroplanes," *Human Factors and Aerospace Safety*, Vol. 1, No. 3, 2001, pp. 201-221.

[3] Baber C., Stanton N.A., "Human error identification techniques applied to public technology: predictions compared with observed use," *Applied Ergonomics*, Vol. 27, No. 2, 2006, pp. 119-131.

[4] Boeing Commercial Airplanes Group, Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations 1959-1999. Seattle WA: BCAG, 2000.

[5] Chapanis, A., The Chapanis Chronicles, Aegean Publishing Company: Santa Barbara, California, 1999.

[6] Embrey, D. E., "SHERPA: A systematic human error reduction and prediction approach," Paper presented at the *International Meeting on Advances in Nuclear Power Systems*, Knoxville, Tennessee, 1986.

[7] Federal Aviation Administration, Report on the Interfaces between Flightcrews and Modern Flight-deck Systems, Washington DC: FAA, 1996.

[8] Grether, W. F., "Instrument reading. 1. The design of long-scale indicators for speed and accuracy of quantitative readings," *Journal of Applied Psychology*, Vol. 33, 1949, pp. 363-372.

[9] Harris, D., (This Volume) "Human Factors for Flight-deck Certification: Issues in Compliance with the new European Aviation Safety Agency Certification Specification 25.1302," *Journal of Aeronautics, Astronautics and Aviation (Series A).*

[10] Joint Airworthiness Authorities, Human Factors Aspects of Flight-deck Design: Interim Policy Paper INT/POL/25/14, Hoofdorp: JAA. 2001.

[11] Kennedy, R. J., "Can human reliability assessment (HRA) predict real accidents? A case study analysis of HRA," In A.I. Glendon & N.A Stanton (Eds), *Proceedings of the risk assessment and risk reduction conference,* 22n March 1994, Aston University, Birmingham. 1995.

[12] Kirwan, B. & Ainsworth, L. K., A Guide to Task Analysis, Taylor and Francis, London. 1992.

[13] Kirwan, B., (1992a) "Human Error Identification in Human Reliability Assessment. Part 1: Overview of approaches," *Applied Ergonomics*, Vol. 23, 1992, pp. 299-318.

[14] Kirwan, B. (1992b) "Human Error Identification in Human Reliability Assessment. Part 2: Detailed comparison of techniques," *Applied Ergonomics*, Vol. 23, 1992, pp. 371-381.

[15] Kirwan, B., A Guide to Practical Human Reliability Assessment, Taylor and Francis, London. 1994.

[16] Kirwan, B., "Human Error Identification Techniques for Risk Assessment of High Risk Systems – Part 1: Review and evaluation of techniques," *Applied Ergonomics*, Vol. 29, 1998a, pp. 157-177.

[17] Kirwan, B., "Human Error Identification Techniques for Risk Assessment of High Risk Systems– Part 2: Towards a Framework Approach," *Applied Ergonomics*, Vol. 29, No. 5, 1998b, pp. 299-318.

[18] Kletz, T., "HAZOP and HAZAN: Notes on the identification and assessment of hazards," In: C. D. Swann & M. L Preston (eds.) *Twenty five years of HAZOPs. Journal of Loss Prevention in the Process Industries*, Vol. 8, No. 6, 1974, pp. 349-353.

[19] Stanton, N. A. & and Baber, C., "A systems approach to human error identification," *Safety Science*, Vol. 22, 1996, pp. 215-228.

[20] Stanton, N. A. & and Baber, C., "Error by design: methods for predicting device usability," *Design Studies*, Vol. 23, No. 4, July 2002, pp. 363-384.

[21] Stanton, N. A., Harris, D., Salmon, P., Demagalski, J. M., Marshall, A., Young, M. S., Dekker, S. W. A., and Waldmann, T., "Predicting Design-Induced Pilot Error using HET (Human Error Template) – A New Formal Human Error Identification Method for Flight-decks," *The Aeronautical Journal*, Vol. 110, No. 2, 2006, pp. 107-115.

[22] Stanton, N. A., Salmon, P., Harris, D., Marshall, A., Demagalski, J. M., Young, M. S., Waldmann, T., and Dekker, S. W. A., "Predicting Pilot Error On The Flight-deck: A Comparison Of Multiple Method and Multiple Analyst Sensitivity," *Applied Ergonomics*, Vol. 40, No. 3, 2009, pp.464-471.

[23] Stanton N. A., Stevenage S. V., "Learning to predict human error: issues of reliability, validity and acceptability," *Ergonomics*, Vol. 41, No. 11, 1998, pp. 1737-15756.

[24] Stanton, N. A & Stevenage, S. V., Learning to predict human error: issues of acceptability, reliability and validity. In: J. Annett & N. A. Stanton (eds.) *Task Analysis.* Taylor and Francis, London, 2000.

[25] Swann, C. D. & Preston, M. L., "Twenty five years of HAZOPs," *Journal of loss prevention in the Process Industries*, Vol. 8, No. 6, 1995, pp. 349-353.

[26] Whalley, Minimising the cause of human error. In B. Kirwan & L. K. Ainsworth (eds.) *A Guide to Task Analysis.* Taylor and Francis, London, 1988.

[27] Whalley , S. J., & Kirwan, B., "An evaluation of five human error identification techniques," *Paper presented at the 5th International Loss Prevention Symposium*, Oslo, June, 1989.

[28] Williams, J. C., "Validation of human reliability assessment techniques," *Reliability Engineering*, Vol. 11, 1989, pp. 149-162.

[29] Newman T. P. and Courteney H. Y., "Certifying for

Safety," In, *Proceedings of Technology and the Flight-deck Conference*, Vancouver, 1997.

[30] UK Civil Aviation Authority, Joint Airworthiness Requirements, (JAR 25 – Large Aeroplanes). London: CAA, 1978.

[31] US Department Of Transportation, *Federal Aviation Regulations, (Part 25 – Airworthiness Standards)*. Revised January 1, 2003, Washington, DC: DOT, 1974.

[32] US Department Of Transportation, *Federal Aviation Regulations*, (Part 61– Certification: Pilots, Flight Instructors, And Ground Instructors). Revised January 1, 2003. Washington, DC: DOT, 1974.

[33] US Department Of Transportation, Federal Aviation Regulations, (Part 121– Operating Requirements: Domestic, Flag, And Supplemental Operations). Revised January 1, 2003, Washington, DC: DOT, 1974.

[34] US Department of Transportation, Aviation Rulemaking Advisory Committee; Transport Airplane and Engine: Notice of new task assignment for the Aviation Rulemaking Advisory Committee (ARAC), *Federal Register*, Vol. 64, No. 140, July 22, 1999.