Predicting Pilot Error on the Flight Deck: Validation of a New Methodology and a Multiple Methods and Analysts Approach to Enhancing Error Prediction Sensitivity

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

[1]Brunel University, BIT-LAB, Uxbridge, Middlesex, UB8 3PH, UK,

[2]Cranfield University, Human Factors Group, School of Engineering, Cranfield, Bedford, MK43 OAL, UK

[3]Marshall Associates, London, UK,

[4]University of Limerick, Ireland,

[5]Lund University, Sweden

*Corresponding author: paul.salmon@brunel.ac.uk + 44 (0) 1895 265543

Abstract

The Human Error Template (HET) is a recently developed methodology for predicting designed induced pilot error. This article describes a validation study undertaken to compare the performance of HET against three contemporary Human Error Identification (HEI) approaches when used to predict pilot errors for an approach and landing task and also to compare individual analyst error predictions to an approach to enhancing error prediction sensitivity: the *multiple analysts and methods* approach, whereby multiple analyst predictions using a range of HEI technique are pooled. The findings indicate that, of the four methodologies used in isolation, analysts using the HET methodology offered the most accurate error predictions, and also that the multiple analysts and methods approach was more successful overall in terms of error prediction sensitivity than the three other methods but not the HET approach. The results suggest that when predicting design induced error, it is appropriate to use domain specific approaches and also a toolkit of different HEI approaches and multiple analysts in order to heighten error prediction sensitivity.

Keywords: Human error, Human Error Identification, Error Prediction, Reliability & Validity.

Introduction to Human Error Identification

Human error remains a problem of great concern to human factor's professionals and within complex sociotechnical systems around 75% of all accidents and safety compromising incidents are attributed to human error. There are many means of reducing or mitigating human error and one approach involves the use of structured methodologies to predict the errors that are likely to be made by operators during task performance. Human Error Identification (HEI) works on the premise that an understanding of an employee's work task and the characteristics of the technology being used allows analysts to predict, a priori, potential errors that may arise from the resulting interaction (Baber & Stanton, 1996, Stanton & Baber, 2002). The use of HEI techniques is now widespread, with applications in a wide range of domains including the nuclear power and petro-chemical processing industries (Kirwan, 1996), air traffic control (Shorrock & Kirwan, 2002), aviation (Harris, Stanton, Marshall, Demagalski, Young & Salmon, 2005), naval space operations (Nelson et al, 1998), medicine and public technology (Baber & Stanton, 1996).

Despite the superfluity of HEI techniques available (a methods review identified over 50 approaches – see Stanton, Salmon, Walker, Baber & Jenkins, 2005) and their increased application, they are relatively rarely used in the domain of the civil flight deck. This is surprising since it has previously been established that the major cause of aviation accidents is human error (McFadden and Towell, 1999). Estimates vary, but data suggest that human error has been identified as a causal or contributory factor in as many as 75% of the accidents that occur in commercial aviation (Civil Aviation Authority, 1998). Further, a number of high profile aviation incidents have been attributed, at least in some part, to design-induced human error, including the Nagoya

Airbus A300-600 accident (where the pilots could not disengage the go-around mode after inadvertent activation due to a lack of understanding of the automation and poor design of the operating logic in the autoland system), the Cali Boeing 757 accident (where the poor interface design of the flight management computer and a lack of logic checking led to a controlled flight into terrain accident) and the Strasbourg A320 accident at Mont St Odile (where the crew inadvertently set an excessive descent rate instead of manipulating the flight path angle as a result of both functions using a common control interface and an associated poorly designed display).

As part of a DTI/EUREKA! funded project entitled, “Prediction of Human Errors on Civil Flightdecks”, the authors developed a new HEI methodology designed to be used specifically for predicting design induced pilot error on civil flight decks. The Human Error Template (HET; Marshall et al, 2003) was developed specifically for use in the certification of civil flight deck technology. The impetus for this came from a US Federal Aviation Administration report (FAA, 1996), which identified many major deficiencies in the design process of modern commercial airliner flight decks. In particular, the recommendations within the report made explicit the requirement for flight deck designs to be evaluated for their susceptibility to design-induced flightcrew errors and also to identify the likely consequences of those errors during the type certification process (Harris et al, 2005).

As safety-critical systems become more and more complex, the use of HEI is an important feature of system design and analysis that can contribute to the enhancement of system safety. Accurate error prediction allows systemic flaws to be removed from system designs and from already existing systems. Valid approaches

for predicting errors in safety-critical systems, as well as procedures that potentially could enhance the accuracy of HEI efforts are therefore an important provision. The aims of this study were two fold. Firstly, we wished to assess the performance of the HET methodology against three contemporary HEI methods, SHERPA (Embrey 1986), Human Error HAZOP and the Human Error Identification in Systems Tool (HEIST; Kirwan 1994). The purpose of this was to validate the HET methodology as a tool for predicting design induced pilot error on civil flightdecks. It was anticipated that the HET methodology would be more accurate at predicting design induced pilot error than the three contemporary methods. This assumption was based upon the fact that the HET methodology was developed specifically use on flightdecks. This meant that its error mode taxonomy was tailored especially for use on civil flightdecks, whereas the other three methods were developed for control room and nuclear power plant tasks, and so their EEM taxonomies were not domain specific. Secondly, we wished to compare the performance of an approach designed to enhance the accuracy of error predictions, namely the multiple methods and multiple analysts approach, in which the error predictions of different analysts using different methods are pooled in order to enhance error prediction sensitivity.

The Human Error Template

The HET methodology uses an aviation specific external error mode (EEM) taxonomy that was developed from a review of existing HEI methods and an evaluation of incidences of design-induced pilot error. The HET EEM taxonomy comprises the following 12 error types:

- *Fail to execute* e.g. pilot fails to perform a particular task or action.
- *Task execution incomplete* e.g. pilot fails to perform a task or action in its entirety.

- *Task executed in the wrong direction* e.g. pilot turns a knob or moves a lever in the wrong direction.
- *Wrong task executed* e.g. pilot performs a wrong task or action.
- *Task repeated* e.g. pilot presses the correct button twice.
- *Task executed on the wrong interface element* e.g. pilot presses the wrong button.
- *Task executed too early* e.g. pilot performs a task or action too early in a sequence.
- *Task executed too late* e.g. pilot performs a task or action too late in a sequence
- *Task executed too much* e.g. pilot moves a lever or turns a knob too much.
- *Task executed too little* e.g. pilot does not move a lever or turns a knob sufficiently.
- *Misread Information* e.g. pilot misreads the information presented by a display.
- *Other.*

The HET EEM taxonomy is applied to each bottom level task step in a Hierarchical Task Analysis (HTA; Stanton, 2006) of the flight task under analysis in order to identify any credible errors. The identification of credible errors is based on the analyst's subjective judgement and involves the analyst either observing the task being performed or walking through the task themselves either with the flight deck interface itself or with functional drawings and photographs of the interface. For each credible error (i.e. those judged by the analyst to be possible) the analyst provides a description of the form that the error would take, such as, '*pilot dials in the airspeed value using the heading knob*' or '*pilot fails to lower the landing gear'*. Next, the outcome or consequence associated with the error is described (e.g. the consequence of the pilot dialling in the airspeed using the heading knob would be that the aircraft inappropriately adjusts its heading to that of the erroneously entered speed value).

Finally, judgements on the likelihood of the error occurring (*Low, Medium or High*) and the criticality of the error (*Low, Medium or High*) are made based on domain expertise and experience. If the identified error is given a '*high*' rating for both likelihood and criticality, the interface technology in question is rated as a 'fail', meaning that it is not suitable for certification. An example HET pro-forma for the task step '*Dial the speed/mach knob to enter 150 on the IAS/Mach display*' is presented in Table 1. A flowchart depicting the HET procedure is presented in Figure 1.

Table 1. Example HET output (Source: Marshall et al, 2003)

Scenario: *Land A320 at New Orleans using the autoland system*	Task Step: 3.4.2. *Dial the Speed/MACH knob to enter 150 in the IAS/MACH window*	Interface Elements: *Speed/MACH Knob, IAS/MACH Display, Auto Pilot Panel*

Error Mode	Description	Outcome	Likelihood			Criticality			PASS	FAIL
			L	M	H	L	M	H		
Fail to execute										
Task execution incomplete										
Task execution in wrong direction	Pilot turns the Speed/MACH knob in the wrong direction	Aircraft decreases speed rather than increases speed		✓			✓		✓	
Wrong task executed										
Task repeated										
Task executed on the wrong interface element	Pilot dials in airspeed using the HDG knon rather than the Speed/MACH knob	Aircraft moves to HDG of 150 and stays at current speed			✓			✓		✓
Task executed too early										
Task executed too late										
Task executed too much	Pilot turns the Speed/MACH knob too much	Aircraft takes on incorrect airspeed			✓		✓		✓	
Task executed too little	Pilot does not turn the Speed/MACH knob enough	Aircraft takes on incorrect airspeed			✓		✓		✓	
Misread information										
Other										

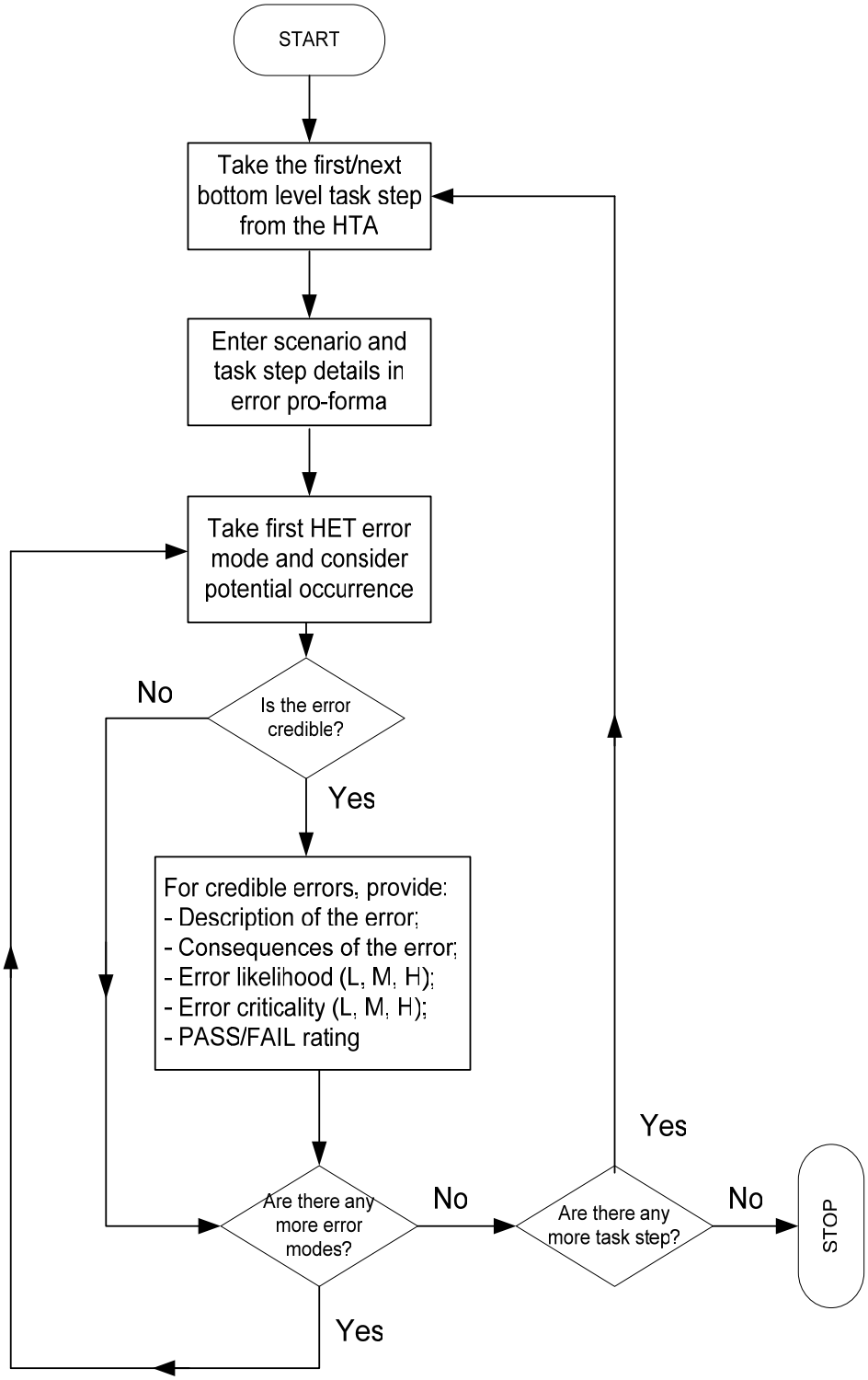

Figure 1. HET Flowchart

Validating the Human Error Template

The validity of HEI techniques requires testing to ensure that they are accurate in the prediction of error, whilst the reliability of HEI techniques requires testing to ensure that the techniques offer the same error predictions when used by different analysts for the same task and when used by the same analyst more than once for the same task. Typically, HEI techniques place a great amount of dependence upon the judgement of the analyst and so different analysts may make different predictions regarding the same problem (*inter-analyst reliability*). Similarly, the same analyst may make different judgments on different occasions (*intra-analyst reliability*).

A number of HEI technique validation studies have been reported in the literature (e.g. Williams, 1989; Whalley and Kirwan, 1989; Kirwan, 1992a; 1992b; 1998a; 1998b, Kennedy, 1995; Baber and Stanton, 1996, Stanton and Stevenage, 1998). For example, Whalley and Kirwan (1989) evaluated six HEI methods for their ability to accurately predict the errors responsible for four incidents that had previously occurred in the nuclear industry. Similarly, Kennedy (1995) examined the ability of a number of HEI methods to predict the errors attributed as causal factors in ten major disasters. In conclusion to an evaluation of 12 HEI approaches, Kirwan (1992b) recommended a combination of expert judgement and the Systematic Human Error Reduction and Prediction Approach (SHERPA; Embrey, 1986) as the most valid approach to HEI. Baber and Stanton (1996) tested the validity of SHERPA and Task Analysis For Error Identification (TAFEI; Baber and Stanton, 1996) when used to predict London Underground rail ticket machine errors. It was concluded that both SHERPA and TAFEI provided an acceptable level of validity based upon the data from two expert analysts. Stanton and Stevenage (1998) also tested the validity of

SHERPA and a heuristic approach when used to predict error on a vending machine task. It was concluded 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 trial one; 0.74 at trial two; and 0.73 at trial three, which represent very acceptable levels of validity.

Multiple Methods and Analysts

It is apparent from the validation studies described above that, although achieving acceptable levels of validity (e.g. SHERPA studies typically return sensitivity index scores of around 0.7) there is room for improvement in terms of the accuracy of HEI error predictions. One such approach could be to use a combination of multiple methods and multiple analysts based on the notion that the accuracy of error predictions may be enhanced by using a range of different but complementary HEI approaches to predict human errors for the same task and also that pooling the error predictions made by a number of different analysts could also enhance the comprehensiveness of the errors predicted. The underlying assumption is that the shortfalls of each HEI technique and each analyst are compensated for by the other techniques and analysts used (i.e. any errors that method A misses, method B will highlight, and any errors that analyst A misses, analyst B may highlight, and so on) which should enhance error prediction sensitivity and accuracy. Kirwan (1998a, b) first proposed the concept of using a range or 'toolkit' of HEI methods to enhance error prediction sensitivity in complex systems. In conclusion to a review of 38 existing HRA/HEI techniques Kirwan (1998a) reported that, since none of the techniques available satisfied all of the 14 criteria against which they were evaluated, a framework or toolkit approach may be the most suitable approach for enhancing the

comprehensiveness of the HEI analysis. Kirwan (1998b) suggested that practitioners to utilise a framework type approach to HEI, whereby a mixture of independent HRA/HEI tools would be used under one framework. One possible framework proposed by Kirwan (1998b) comprised SHERPA, the Human Error Hazard and Operability Study (HAZOP), Errors Of Commission Analysis (EOCA; Kirwan, 1994), Confusion matrix analyses, Fault symptom matrix analysis and the Skill, Rule and Knowledge-based behaviour approach.

Although due to its novelty there appears to be nothing within the academic literature stating the strengths and weaknesses of multiple methods and analysts approaches to HEI, it is apparent that, whilst potentially improving error prediction sensitivity, multiple methods and analysts approaches do have some potential weaknesses. Firstly, the false alarm rate (i.e. errors predicted that do not in fact occur) can potentially be increased due to the pooled error data. However, in safety critical industries it may be acceptable to generate a high rate of false alarms in order to ensure that all potential errors are identified. Secondly, the use of additional methods can significantly increase the level of resources (e.g. time, training etc) required to undertake HEI analyses and also the increased data returned will ultimately increase the time required for data analysis.

The three methods, SHERPA, Human Error HAZOP and HEIST were chosen as a result of a literature review of 32 existing HEI methods from which it was concluded that the SHERPA, Human Error HAZOP and HEIST methodologies were the most suited for use in the prediction of potential design induced error on the flightdeck. A brief description of the three techniques is provided in the following sections. For a

more exhaustive description of the latter three techniques, including example outputs, the reader is referred to Stanton, Salmon, Walker, Baber & Jenkins (2005).

Systematic Human Error Reduction and Prediction Approach (SHERPA)

SHERPA (Embrey, 1986) was originally developed for use in the nuclear reprocessing industry and is probably the most commonly used HEI approach, with applications in a number of domains, including ticket machines (Baber & Stanton, 1996), vending machines (Stanton and Stevenage, 1998), and in-car radio-cassette machines (Stanton & Young, 1999). SHERPA uses a behavioural taxonomy linked to an error mode taxonomy and is applied to a HTA of the task under analysis. The behavioural and EEM taxonomies are used to identify credible errors that are likely to occur during each step in the HTA. For each credible error identified the analyst provides a description of the form that the error would take, such as, '*pilot dials in wrong airspeed*' and identifies any consequences associated with the error and also any recovery steps that would need to be taken in event of the error being made. Finally, ordinal probability (Low, Medium or High), criticality (Low, Medium or High) and potential design remedies are recorded.

Human Error Hazard and Operability Study (HAZOP)

HAZOP (Kletz, 1974; cited in Swann & Preston, 1995) is a well-established engineering approach that was developed in the late 1960s by ICI (Swann and Preston, 1995) for use in process design audit and engineering risk assessment (Kirwan, 1992a). Typically undertaken as a group approach, HAZOP involves analysts applying guidewords, such as *Not done*, *More than* or *Later than*, to each task step in order to identify potential errors that may occur. Many variations on the

HAZOP approach exist, and the Human Error HAZOP approach was developed for dealing with human error issues (Kirwan and Ainsworth, 1992). In the present study, a set of Human Error HAZOP guidewords (Whalley, 1988; cited in Kirwan & Ainsworth, 1992) were used. Each guideword is applied to each task step to identify any credible errors. Once a description of the error is provided, the consequences, cause and recovery path of the error are described. Finally, redesign suggestions are made to either prevent the error from occurring or mitigate its consequences.

Human Error Identification in Systems Tool (HEIST)

The HEIST technique (Kirwan, 1994) is a component of the HERA methodology (Kirwan, 1998b) and uses *error identifier questions* (e.g. "Could the operator fail to carry out the act in time?") linked to behaviour tables and an external error mode taxonomy that are designed to prompt the analyst to identify potential errors. The task step in question is firstly classified into one of the HEIST behavioural categories and then the associated HEIST behaviour table and error identifier prompts are used to encourage the analyst to identify any errors that could potentially occur during performance of the task in question. For each credible error identified, the system cause or psychological error mechanism and error reduction strategy (both of which are provided in the HEIST behaviour tables) is recorded and the consequences associated with the error are described.

The main differences between the approaches compared relates to the type of approach that they represent, the taxonomies of error modes that they use, how the analyst goes about predicting the errors with the technique and also what additional information is provided once an error has been identified. In terms of the type of HEI

approach, the HET, SHERPA and Human Error HAZOP approaches are examples of taxonomy-based HEI techniques, which are characterised by their use EEM taxonomies to identify potential errors. Typically EEMs are considered for each component step in a particular task or scenario in order to determine credible errors that may arise during the man-machine interaction. Taxonomic approaches to HEI are typically the most successful in terms of sensitivity and are also the cheapest, quickest and easiest to use. However, these techniques depend greatly on the judgement of the analyst and their reliability and validity may at times be questionable. HEIST, on the other hand, is an example of an error identifier prompt-based technique. These approaches use prompts or questions to aid the analyst in identifying potential errors. The prompts are typically linked to a set of error modes and reduction strategies. HEIST is also different in this case since it also considers performance shaping factors. Whilst these techniques attempt to remove the reliability problems associated with taxonomy-based approaches, they add considerable time to the analysis because each prompt must be considered.

Each of the four approaches taxonomies is also distinct in that they were developed specifically for the method and domain in question. The HET EEM taxonomy contains twelve generic error modes that were collated from an analysis of civil aviation accidents and incidents and a review of existing HEI approaches. The SHERPA approach instead uses five separate EEM taxonomies linked to categories of human behaviour and so requires the analyst to firstly classify the task in question into one of these behaviours. The Human Error HAZOP approach uses so-called guidewords which again are different to the other methods taxonomies and are specific to process control. Finally, the HEIST approach again uses a different

taxonomy of EEM that was developed within the nuclear and chemical process control domain.

The purpose of this study was to compare the performance of HET against three contemporary Human Error Identification (HEI) approaches when used to predict pilot errors for an approach and landing task and also to compare, in terms of error prediction sensitivity, the multiple methods and analysts approach with multiple analyst predictions for each method.

Methodology

Participants: A total of 37 Brunel undergraduate students aged between 19 and 21 years old were used as participants in the study. Our justification for using undergraduate participants with no previous experience of human error identification and only limited experience of human factors in general stems from the original requirement for the methodology developed to be usable by non-human factors specialists during the design and certification of flight deck technology. For example, Marshall et al (2003) stated that "the method should also be capable of being used by non-human factors experts within the certification authorities" (p.6). Further, the capture of potential errors in the early design phases of a system require that designers (with limited or no human factors experience) are able to use HEI approaches. The participants were allocated into four groups based upon the HEI methodology that they used during their study (four separate error prediction studies were conducted, one for each HEI methodology). Group one consisted of eight male undergraduate students. These participants formed the *HET* group and received training in the HET methodology. Group two consisted of nine undergraduate students. Of these six were

male and three were female. These participants formed the *SHERPA* group and received training in the SHERPA methodology. Group 3 consisted of a further nine undergraduate students. 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 methodology. The fourth and final group consisted of 11 undergraduate students. Of these, eight were male and three were female. These participants formed the *HEIST* group and received training in the HEIST methodology. All participants had no previous experience of any of the HEI methodologies used nor of flying an aeroplane.

Flight task: The study focussed on the aircraft-landing task using 'Land aircraft X at New Orleans Airport using the autoland system' This task was part of the approach phase of a flight in Aircraft X (a modern, highly automated, 'glass cockpit', medium capacity airliner). This task was chosen as was deemed to be representative of a typical civil aviation landing task in an automated glass cockpit airliner. A HTA was constructed for the flight task based on an observation of a video recording of a similar landing task and consultation with subject matter experts. An extract of the HTA is presented in Figure 2.

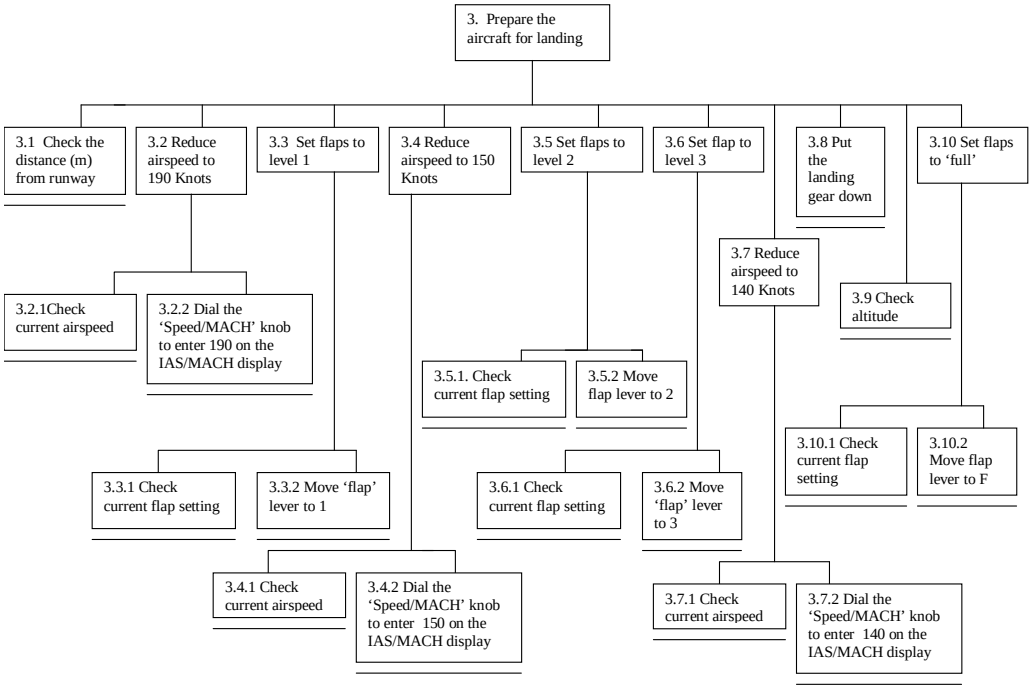

Figure 2. Extract of Landing Task HTA (Source: Marshall et al, 2003).

Materials

All participants were supplied with a training package for the methodology in question. The training packages consisted of a description of the method in question, a copy of the taxonomy associated with the error prediction method; 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 in question. Participants were also given a HTA describing the action stages involved when using a vending machine as part of the training and also a HTA describing the action stages involved when landing aircraft X at New Orleans using the Auto-land system for the main study. 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 (in the appropriate mode), landing gear lever and the Captain's navigation display. All participants were also provided with suitable pro-

formae for recording their error predictions. Microsoft Flight Simulator 2000™ Professional Edition was also used to give the participants a demonstration and walkthrough of the flight task under analysis.

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 HEI exercise.

Procedure

Participants were recruited via e-mail advertisement and the respondents were divided into four separate groups, based upon the four HEI techniques used. For each group, participants were initially given a short briefing on the purpose of the experiment. Following this a lecture-based introduction to the areas of Human Error and HEI was given. Next, participants were given a short training session on the method that their particular group were being tested on. This included a short introduction to the method and a step by step walkthrough of a worked example of a HEI analysis using the method in question. The analysis used for each of the methods was a HEI analysis of a Ford in-car radio cassette system (Stanton and Young, 1999).

Once familiar with their HEI method, participants were given a HTA of a vending machine task (Stanton and Stevenage, 1998) along with A3 photographs of the vending machine and its user interface on which to undertake practice HEI analysis. After a demonstration of the task and a walk through of the HTA, participants used

their allocated method to make error predictions for the vending machine task. At this stage, participants were permitted to confer with other participants and also to ask the experimenter questions regarding the analysis. Once the error predictions were complete, participants were provided with an 'expert' analysis (undertaken by a human factors researcher with considerable experience in HEI) for the vending machine task so that they could compare their error predictions with an experts error predictions for the same task. The experimenter then discussed each of the errors predicted and answered any questions regarding the vending machine error prediction task.

After a short break, participants were then given the HTA for the task, 'Land aircraft X at New Orleans using the Auto-land system', as the experimental condition, along with colour photographs of all of the relevant flight deck equipment. After an initial walk through of the task, participants were given a step-by-step demonstration of the landing task using Microsoft Flight Simulator 2000 Professional Edition. Participants were then asked to predict any potential design induced pilot errors for the flight task. For reliability purposes, participants returned four weeks later to carry out a repetition of the analysis (hereafter referred to as Trial 2) employing the same HEI technique that they had used during the first error prediction exercise (hereafter referred to as Trial 1).

Data Analysis

To compute validity statistics, the error predictions made by each participant were compared with actual error incidence data reported by pilots using the auto-land system for the flight task under analysis (which was obtained via questionnaire

survey). In this survey pilots type-rated on the same aircraft 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 aircraft X at New Orleans airport using the Auto-Land system'. A total of 46 pilots (45% Captains, 37% First Officers, 13.3% Trainee Captains, 4.7% who declined to state their position) with experience ranging from less than 2,000 hours to over 16,000 hours (Mean = 6, 832 hours, SD = 4, 524 hours) responded to the survey. 57 different error types were reported in the survey. A detailed description of these errors can be found in Marshall et al (2003).

The sensitivity of each participant's 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 (e.g. Stanton and Stevenage, 1998, Harris et al, 2005). The signal detection paradigm sorts the data into the following mutually exclusive categories:

1) Hit – An error predicted by the analyst that was also reported by the survey respondents.
2) Miss – The Failure to predict an error that was reported by the survey respondents.
3) False Alarm – An error predicted by the analyst but that was not reported by the survey respondents.
4) Correct rejections – Correctly rejected error that was not reported by the pilots. This represents the number of errors contained in the HEI methods error mode taxonomy that were correctly rejected by the analyst and also not reported by the survey respondents.

These four categories were entered into the signal detection grid for each subject. The signal detection paradigm was then used to calculate the sensitivity index (SI). This returns a value of 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 given in Formula 1 (from Stanton and Stevenage, 1998)

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

Formula 1. Sensitivity Index formula

Results

Treatment of data: The data obtained had to first be grouped so that the *multiple methods and analysts* approach sensitivity could be calculated. For the multiple analyst but single method data each analyst's error predictions using each HEI approach (e.g. SHERPA, HET, HAZOP, and HEIST) were pooled. For the *multiple methods and analysts* approach, six randomly selected participant's error predictions from each method were pooled together. In order to be consistent in the comparison of the individual methods with the pool of 'multiple methods and analysts' some cases had to be discounted, as not all participants turned up to all of the sessions. A core pool of six participants in each of the groups was formed, whose data was then used to form the 'multiple methods and analysts' group. The sensitivity of these error predictions (each method in isolation and the multiple methods and analysts approach data) was then assessed using the sensitivity index formula described above.

The mean Trial 1 and Trial 2 SI scores for multiple analysts using each method and also for the multiple methods and analysts approach are presented in Figure 3.

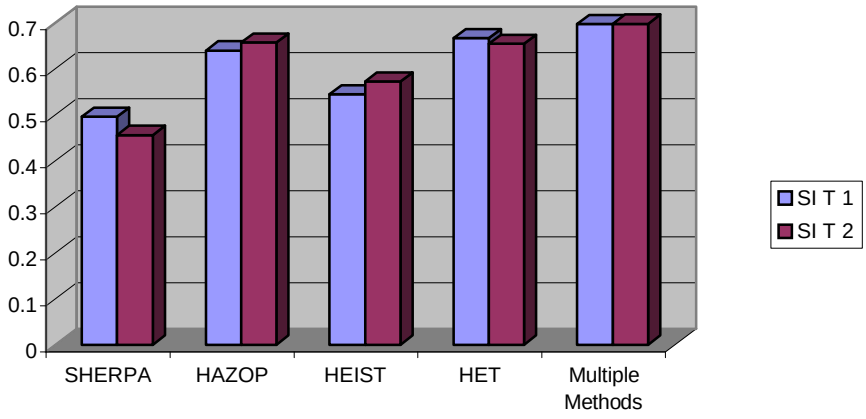

Figure 3. Mean SI Scores (Trial 1 & trial 2)

The mean SI score results show that the multiple methods and analysts approach achieved the greatest mean SI scores (Trial 1 = 0.69, Trial 2 = 0.69), followed by multiple analysts using the HET approach (Trial 1 = 0.66, Trial 2 = 0.65). As sensitivity is made up of hit rate and false alarm rate, each of these was considered separately. The mean trial 1 and trial 2 hit rate scores for multiple analysts using each method and for the multiple methods and analysts approach are presented in Figure 4.

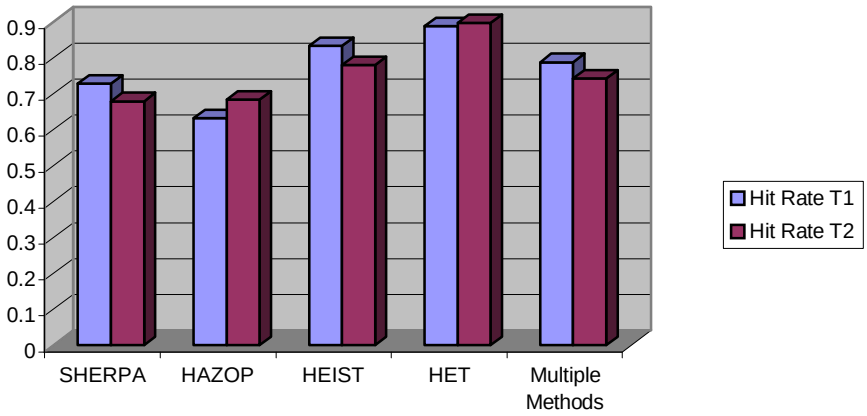

Figure 4. Mean Hit Rate Scores (Trial 1 & trial 2)

The mean hit rate score results show that the multiple analysts using HET approach achieved the greatest mean hit rate scores (Trial 1 = 0.88, Trial 2 = 0.89).

The mean trial 1 and trial 2 false alarm rate scores for multiple analysts using each method and for the multiple methods approach are presented in Figure 5.

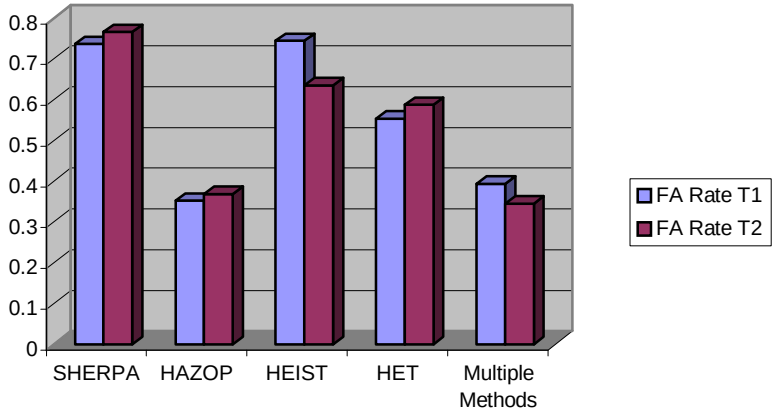

Figure 5. Mean False Alarm Rate Scores (Trial 1 & trial 2).

The mean false alarm rate scores show that, at trial 1, the multiple analysts using HAZOP approach achieved the lowest false alarm rate score (0.35) whilst the multiple methods approach achieved the lowest false alarm rate score at trial 2 (0.34).

Mann-Whitney ‘U’ statistical tests were performed to establish if the observed differences between the sensitivity index, hit rate and false alarm rate scores for the multiple methods and analysts and multiple analysts approaches were significant. The results are presented in Table 2.

Table 2. Multiple Analysts versus Multiple Methods and Analysts Significance Table

Signal Detection Criteria	Multiple Analysts using			
	HET	HAZOP	SHERPA	HEIST
Hit Rate Trial 1		**.0038 <0.005**	**.0101 <0.05**	
Hit Rate Trial 2	**.0154 <0.05**	**.0161 <0.05**	**.0245 <0.05**	**.0247 <0.05**
False Alarm Rate Trial 1		**.0039 <0.005**	**.0159 <0.05**	**.0062 <0.01**
False Alarm Rate Trial 2	**.0163 <0.05**		**.0064 <0.01**	
Sensitivity Index Trial 1		**.0039 <0.05**	**.0039 <0.005**	**.0039 <0.005**
Sensitivity Index Trial 2			**.0039 <0.005**	**.0065 <0.01**

= Not Significant

The results presented in Table 2 demonstrate that at trial 1, the multiple methods and analysts approach sensitivity index scores were significantly better than the multiple analysts scores using HAZOP, SHERPA and HEIST, but not significantly better than the HET multiple analyst score and at trial 2 that the multiple methods and analysts approach scores were significantly greater than the multiple analyst scores using SHERPA and HEIST, but not the multiple analyst HET and HAZOP scores.

For hit rate at trial 1, the multiple methods and analysts approach score was significantly higher than the multiple analyst HAZOP and SHERPA analyses, but was not significantly greater than the multiple analyst HET and HEIST analyses. For hit rate trial 2, the multiple methods analysis approach score was significantly greater than the multiple analysts' scores for each of the four methods.

For false alarm rate, at trial 1 the statistical analysis indicates that the multiple methods and analysts approach scored lower (and therefore better) than the multiple analysts using HAZOP, SHERPA and HEIST, but that the difference between the multiple methods analysis and the multiple analysts using HET was not significant. For false alarm rate trial 2, the multiple methods and analysts analysis approach scores were significantly lower than multiple analysts using HET and SHERPA, but were not significantly lower than multiple analyst scores for HAZOP and HEIST.

Discussion

Error Prediction Sensitivity

This study had two main objectives. The first objective of the study was to compare the accuracy of the HET approach when used to predict design induced pilot error against three other contemporary HEI approaches developed in other domains. In conclusion, participants using the HET methodology were the most accurate in their predictions for the flight task under analysis, both at Trial 1 and Trial 2. This was most probably attributable to the fact that the HET error mode taxonomy was developed specifically for the use on civil flightdecks, whilst the other techniques were not. SHERPA, Human Error HAZOP and HEIST were developed for the nuclear power and process control domains. This meant that analysts were somewhat constrained in terms of the errors that they could predict, since the taxonomy used in these methods taxonomy may not have contained errors of the type that might occur on civil flight decks. Previous studies have also demonstrated that the HET approach is more accurate than HAZOP, HEIST and SHERPA when used to predict errors for the same flight task (Stanton, Harris, Salmon, Demagalski, Marshall, Young, Dekker & Waldmann, 2006).

The second objective of this study was to test an approach to enhancing the accuracy of error predictions, the *multiple methods and analysts* approach. In terms of the overall accuracy of the error predictions, it was observed that the multiple methods and analysts approach was more significantly more accurate than the multiple analyst approach using HAZOP (at trial 1 only), SHERPA and HEIST at predicting errors for the flight task, but not more accurate than the error predictions offered by the multiple analysts using HET approach (nor HAZOP at trial 2). This finding has implications

for the prediction of human error in complex systems. For example, in some cases, it may be more appropriate to use *multiple methods and multiple analysts* to predict errors, rather than just one analyst and one method in isolation. On the basis of the findings derived from this study, using a group of analysts and HEI methods to predict error can enhance the sensitivity of error predictions, whereas using only one method could potentially lead to critical errors being missed during the error prediction process. This certainly appears to be the case when attempting to predict error in domains for which no HEI approaches have been specifically developed. One way of enhancing the sensitivity of the error predictions made would therefore be to use a combined toolkit of a range of HEI methods from other domains.

This research lends support to Kirwan's (1998a, b) argument for the use of a comprehensive multiple methods (i.e. toolkit) approach. The differences in the taxonomies seem to ensure greater capture of the types of error that occur on the flight deck. Alternatively, an error taxonomy that has been developed specifically for the domain in question appears to perform equally as well when multiple analysts are utilised. It is assumed that the superior accuracy of the multiple methods and analysts approach was due to the error mode taxonomy being more comprehensive as in this case it was effectively four error taxonomies combined. This comprehensiveness of the error taxonomy is likely to lead to an increase in the numbers of errors correctly identified (e.g. hits) and thus reduce the number of errors that are missed (e.g. misses). On the downside, the increased number of error modes could potentially increase the number of wrongly identified errors (e.g. false alarms) and decrease the numbers of errors correctly discarded (e.g. correct rejections). This was not the case in this study, however, with the multiple methods and analysts group performing better

in terms of SI and false alarm rate scores. Further, as pointed out earlier, in some circumstances (i.e. in safety critical systems analysis) it may be acceptable to generate a high false alarm rate if it contributes to the detection of more errors. It is recommended, however, that when using a multiple methods and analysts approach, appropriate subject matter experts with a sufficient level of experience in HEI are used or a combination of subject matter experts and methods experts working together (Stanton & Stevenage, 1998). It should be noted that the analysts in this case were neither experts in the domain of civil aviation nor were they experts in the application of HEI techniques. It would be expected that the signal detection theory statistics should be higher (i.e. error predictions more accurate) if this were the case.

The finding that the multiple method and analysts approach was not significantly more accurate than the multiple analysts using HET approach indicates that a multiple analysts approach (using the same approach) also potentially offers a means of enhancing error prediction sensitivity, albeit if they are using a HEI approach that has been developed specifically for the domain in which the analysis is taking place.

In closing, this study has demonstrated that the HET approach is a viable tool for identifying design induced pilot errors within the civil aviation domain. The level of accuracy attained by inexperienced analysts when using the HET approach to identify such errors is encouraging and suggests that the HET approach, when used by domain experts with significant experience in HEI analysis, can potentially be a very powerful tool for accurately identifying design induced pilot error. Further, this study seems to suggest that error prediction sensitivity can potentially be enhanced through the use of a multiple methods and analysts approach, which indicates that future HEI analyses

efforts should utilise teams of HEI analysts with access to a toolkit of different HEI approaches.

It is recommended that further research into means of enhancing error prediction accuracy be undertaken. Also, further applications of the HET approach within the aviation domain are encouraged. Further, whilst this study has demonstrated that using multiple analysts and methods may enhance error prediction sensitivity, it is clear that further investigation in other domains in which error prediction is dominant is required, such as the process control and air traffic control domains. The usefulness of HEI techniques is already assured. However enhancing the accuracy of the error predictions offered by such techniques can only make them more powerful tools within system design and analysis efforts.

Acknowledgements

We wish to acknowledge that this research was made possible through funding from the Department of Trade and Industry as part of the European EUREKA! Programme. Our thanks go to John Brumwell & Gillian Richards of the CARAD Advanced Systems Programme at the DTI and Richard Harrison (now at QinetiQ) who have provided invaluable support to this research and also to Air2000, British Midland and JMC airlines and their pilots for taking the time to complete the questionnaire.

References

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.

Civil Aviation Authority. 1998. Global Fatal Accident Review 1980-96 (CAP 681), Civil Aviation Authority, London.

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.

Federal Aviation Administration. 1996. Report on the interfaces between flightcrews and modern flight deck systems, Federal Aviation Administration, Washington DC, USA.

Harris, D., Stanton, N. A., Marshall, A., Young, M. S., Demagalski, J. and Salmon, P. M. 2005. Using SHERPA to predict design induced error on the flight deck, *Aerospace Science and Technology*, 9, pp 525-532.

Kennedy, R. J. 1995. 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. Birmingham, UK: Aston University.

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

Kirwan, B. 1992a. Human Error Identification in Human Reliability Assessment. Part 1: Overview of approaches. *Applied Ergonomics*, 23 pp. 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.

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

Kirwan, B. 1996. The validation of three Human Reliability Quantification techniques – THERP, HEART and JHEDI: Part 1 – technique descriptions and validation issues, *Applied Ergonomics*, Vol 27, 6, pp 359 – 373

Kirwan, B. 1998a. Human error identification techniques for risk assessment of high risk systems – Part 1: review and evaluation of techniques, *Applied Ergonomics*, 29 (3), pp 157-177.

Kirwan, B. 1998b Human Error Identification Techniques for Risk Assessment of High Risk Systems– Part 2: Towards a Framework Approach. *Applied Ergonomics*. 29 (5), pp 299-318.

Marshall, A., Stanton, N. A, Young, M., Salmon, P. M., Harris, D., Demagalski, J., Waldmann, T. and Dekker, S. W. 2003. Development of the human error Template – a new methodology for assessing design induced errors on aircraft flight decks. Final Report of the ERRORPRED Project E! 1970 (August 2003), London: Department of Trade and Industry, 2003.

McFadden, K. L., & Towell, E. R. 1999. Aviation human factors: a framework for the new millennium, *Journal of Air Transport Management*, 5, pp 177-184.

Nelson, W.R, Haney, L.N, Ostrom, L.T, Richards, R.E. 1998. Structured methods for identifying and correcting potential human errors in space operations, *Acta Astronautica*, vol 43, pp 211-222.

Shorrock, S.T., Kirwan, B. 2000. Development and application of a human error identification tool for air traffic control, *Applied Ergonomics*, Vol. 33 pp 319-336.

Stanton, N. A. 2006. Hierarchical task analysis: Developments, applications and extensions. *Applied Ergonomics,* 37, pp. *55-79*

Stanton, N.A. & Baber, C. 2002. Error by design: methods for predicting device usability, *Design Studies*, 23, (4), pp 363-384.

Stanton, N.A., & Stevenage, S.V. 1998. Learning to predict human error: issues of reliability, validity and acceptability, *Ergonomics*, 41, pp 1737-1756.

Stanton, N. A, and Young, M. S. 1999. A guide to methodology in ergonomics: Designing for human use, London, Taylor and Francis.

Stanton, N. A., Salmon, P., Walker, G. H, Baber, C. & Jenkins, D. 2005. Human Factors Methods: A Practical Guide for Engineering and Design. Aldershot: Ashgate.

Stanton, N., Harris, D., Salmon, P. M., Demagalski, J. M., Marshall, A., Young, M. S., Dekker, S. W. A. & Waldmann, T. 2006. Predicting design induced pilot error using HET (Human Error Template) – A new formal human error identification method for flight decks. *Journal of Aeronautical Sciences*, February, 107-115.

Swann, C. D. & Preston, M. L. 1995. Twenty five years of HAZOPs, *Journal of loss prevention in the Process Industries*, vol 8 (6), pp 349-353.

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

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