Predicting design induced pilot error using HET (human error template) – A new formal human error identification method for flight decks

N. A. Stanton, D. Harris[†], P. M. Salmon, J. M. Demagalski[†], A. Marshall*, M. S. Young, S. W. A. Dekker[‡] and T. Waldmann[§]

Department of Design, Brunel University
Middlesex, UK

[†]Cranfield University
Bedford, UK

*Marshall Ergonomics Ltd
Hampshire, UK

[‡]School of Aviation, Lund University
Ljungbyhed
Sweden

[§]College of Engineering
University of Limerick,
Limerick, Ireland

ABSTRACT

Human factors certification criteria are being developed for large civil aircraft with the objective of reducing the incidence of design-induced error on the flight deck. Many formal error identification techniques currently exist which have been developed in non-aviation contexts but none have been validated for use to this end. This paper describes a new human error identification technique (HET – human error template) designed specifically as a diagnostic tool for the identification of design-induced error on the flight deck. HET is benchmarked against three existing techniques (SHERPA – systematic human error reduction and prediction approach; human error HAZOP – hazard and operability study; and HEIST – human error In systems tool). HET outperforms all three existing techniques in a validation study comparing predicted errors to actual errors reported during an approach and landing task in a modern, highly automated commercial aircraft. It is concluded that HET should provide a useful tool as a adjunct to the proposed human factors certification process.

NOMENCLATURE

χ^2 Value in the Chi-square distribution used to test for significant differences between three (or more) independent groups when using the Kruskall-Wallis non-parametric analysis of variance where the dependent variable is measured on an ordinal scale.

p Probability of making a type I decision error.

U Critical value in the Mann-Whitney U test to test for significant differences between two independent groups where the dependent variable is measured on an ordinal scale. The parameter U is the number of times a value in one group precedes a value in another group when values are sorted in ascending order.

z A value in the standard normal distribution that may be related directly to a probability value to determine statistical significance when using the Wilcoxon Matched Pairs Signed Ranks test to establish if there is a statistically significant difference between two related samples where the dependent variable is measured on an ordinal scale.

Paper No. 3026. Manuscript received 6 October 2005, revised 18 November 2005 accepted 2 December 2005.

1.0 INTRODUCTION

For 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[1]. If this rate remains unchanged, with the current projected increase in the demand for air travel this will mean that there will be one major hull loss almost every week by the year 2015. As the reliability and structural integrity of aircraft has improved the number of accidents directly resulting from such failures has reduced dramatically, hence so has the overall number of accidents. However, human reliability has not improved to the same extent. Figures vary but it is estimated that up to 75% of all aircraft accidents now have a major human factors component. Human error is now the primary risk to flight safety[2].

The roots of human error are manifold and have complex interrelationships with all aspects of the operation of a modern airliner. However, during the last decade 'design induced' error has become of particular concern to the airworthiness authorities, particularly in the highly automated third (and fourth) generations of automated airlines. However, Chapanis[3] noted that that back in the 1940s many aspects of 'pilot error' were really 'designer error'. This was a challenge to contemporary thinking at the time and shows that good design is all-important in human error reduction. He was particularly interested in why pilots often retracted the landing gear instead of the landing flaps after landing the aircraft. He identified the problem as 'designer error' rather than 'pilot error', as the designer had put two identical toggle switches side-by-side, one for the gear and the other for the flaps. It was proposed that the controls were separated and coded. The separation and coding of controls is now standard human factors practice. Half a century after Chapanis's original observations, the idea that one can design error-tolerant devices is beginning to gain credence[4]. The high levels of automation in the new generation airlines have without a doubt offered considerable advances in safety over their forbearers, however new types of error have begun to emerge on these flight decks[5]. This was exemplified by accidents such as the Nagoya Airbus A300-600 (where the pilots could not disengage the go-around mode after inadvertent activation as a result of a combination of 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 on the flight management computer and a lack of logic checking resulted in a CFIT accident); and the Strasbourg A320 accident (where the crew inadvertently set an excessive rate of descent instead of manipulating the flight path angle as a result of both functions utilising a common control interface and an associated poor display).

As a result of such accidents, the US Federal Aviation Administration (FAA) commissioned an exhaustive study of the pilot-aircraft interface on modern flight decks[6]. The report identified several major flight deck design shortcomings and deficiencies in the design process. There were criticisms of the flight deck interfaces, such as pilots' autoflight mode awareness/indication; energy awareness; confusing and unclear display symbology and nomenclature, and a lack of consistency in FMS interfaces and conventions. The report also heavily criticised the flight deck design process, identifying in particular a lack of human factors expertise on design teams and placing too much emphasis on the physical ergonomics of the flight deck, and insufficient on the cognitive ergonomics. Fifty-one specific recommendations came out of the report, including:

> *"The FAA should require the evaluation of flight deck designs for susceptibility to design-induced flightcrew errors and the consequences of those errors as part of the type certification process."*

In July 1999 the US Department of Transportation assigned a task to the Aviation Rulemaking Advisory Committee to provide advice and recommendations to the FAA administrator to 'review the existing material in FAR/JAR 25 and make recommendations about what regulatory standards and/or advisory material should be updated or developed to consistently address design-related flight crew performance vulnerabilities and prevention (detection, tolerance and recovery) of flight crew error'[7]. The European Joint Aviation Authorities (JAA – now European Aviation Safety Agency EASA), as a part of the airworthiness regulatory harmonisation efforts, also subsequently adopted this task. The rules and advisory material being developed as part of this process will be applied to both the type certification and supplemental type certification processes for large transport aircraft[8,9,10]. In the meantime, in 2001 the JAA issued an interim policy document[11] that will remain in force until the new harmonised human factors regulations encompassed in Part 25 come into force. In Europe a notice of proposed amendment was issued in 2004 as a step in the rulemaking process[12].

Compliance with any airworthiness requirement must be established through inspection, demonstration, evaluation, analysis and/or test. To demonstrate compliance with the forthcoming human factors airworthiness requirements, formal error analysis will be one of the most rigorous ways of evaluating the pilot interface and demonstrating that the likelihood of 'design-induced error' is as low as is reasonably practicable. Formal error analysis is not new; however it is a novel approach as a means of demonstrating compliance with a certification requirement. Any technique used for a formal approval process must be reliable, valid and, for the purposes of certification, the method should also be capable of being used by non-human factors experts within the certification authorities (e.g. the certification test pilots). As a direct corollary, any such technique should also be capable of being used by the flight deck design teams to verify in the early stages of design that their flight deck interfaces are likely to comply with the certification requirement. In addition to enhancing safety, there is also a strong economic argument for the early identification of inadequacies in the pilot interface. It has been suggested that there is a 1:10:100 ratio in the cost to correct interface adequacies at the design, development and operational stages, respectively[13].

Any error prediction methodology for the flight deck must be designed to encompass the specific demands of the aviation environment. However, there is some caution and scepticism in the aerospace industry with regard to formal methods that produce a probability of error associated with any aspect of crew performance. Advisory Circular AC25.1309-1A14 also suggests that the reliable quantitative estimation of the probability of crew error is not possible. As a result, emphasis is placed upon the identification of potential errors using formal methods, not their quantification.

This paper describes the development, testing and comparative benchmarking of a new human error identification (HEI) technique, the human error template (HET). HET has been developed specifically for the aerospace industry as a diagnostic tool 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 amended and supplemental type certification.

2.0 DESCRIPTION OF THE HET METHODOLOGY

HET is a checklist style approach to error prediction that comes in the form of an error proforma containing twelve error modes. Figure 1 shows a flowchart of how the HET methodology should be conducted. The HET is applied to each bottom level task step in a hierarchical task analysis[15] (HTA) of the task in question. The technique requires the analyst to indicate which of the HET error modes are credible (if any) for each task step, based upon their judgement.

The HET error taxonomy consists of 12 basic error modes that were 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:

1 Failure to execute.

2 Task execution incomplete.

3 Task executed in the wrong direction.

4 Wrong task executed.

5 Task repeated.

6 Task executed on the wrong interface element.

7 Task executed too early.

8 Task executed too late.

9 Task executed too much.

10 Task executed too little.

11 Misread Information.

12 Other.

Second, for each credible error the analyst provides a description of the form that the error would take. Third, the analyst has to determine the outcome or consequence associated with the error. Finally, the analyst estimates 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. 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 comprehensive as it is 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. An example of a HET output is shown in Fig. 2. An extract of the corresponding HTA upon which it is based can be found in Fig. 3. A full description of the methodology can be found elsewhere[16].

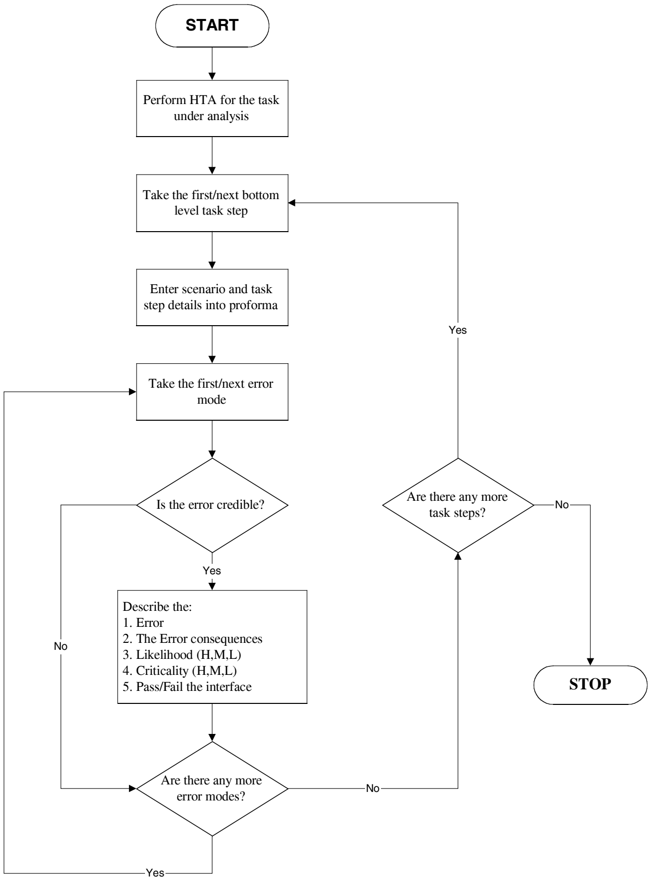

START
Perform HTA for the task under analysis
Take the first/next bottom level task step
Enter scenario and task step details into proforma
Take the first/next error mode
Is the error credible?
Yes
No
Describe the:
1. Error
2. The Error consequences
3. Likelihood (H,M,L)
4. Criticality (H,M,L)
5. Pass/Fail the interface
Are there any more error modes?
Yes
No
Are there any more task steps?
Yes
No
STOP

Figure 1. HET methodology flowchart.

Scenario: **Land A320 at New Orleans using the Autoland system**			**Task step:** **3.4.2 Dial the 'Speed/MACH; knob to slow down to 150kt on IAS/MACH display**								
				Likeli-hood			**Critic-ality**				
				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	Aircraft speeds up instead of slowing down		✓			✓		✓	
Wrong task executed											
Task repeated											
Task executed on wrong interface element	✓	Pilot dials using the HDG knob instead	Aircraft 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	Aircraft slows down too much	✓			✓				✓
Task executed too little	✓	Pilot turns the Speed/MACH knob too little	Aircraft does not slow down enough/Too fast for approach	✓				✓		✓	
Misread information											
Other											

Figure 2. Example of HET output.

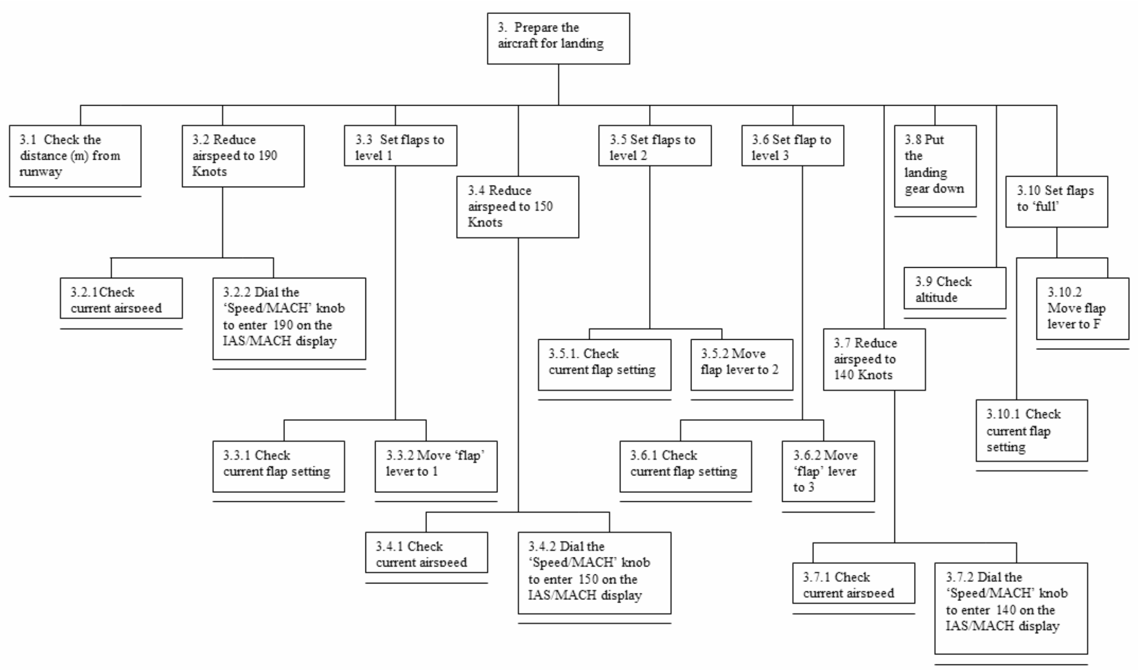

3. Prepare the aircraft for landing
3.1 Check the distance (m) from runway
3.2 Reduce airspeed to 190 Knots
3.3 Set flaps to level 1
3.4 Reduce airspeed to 150 Knots
3.5 Set flaps to level 2
3.6 Set flap to level 3
3.8 Put the landing gear down
3.10 Set flaps to 'full'
3.2.1Check current airspeed
3.2.2 Dial the 'Speed/MACH' knob to enter 190 on the IAS/MACH display
3.9 Check altitude
3.10.2 Move flap lever to F
3.5.1. Check current flap setting
3.5.2 Move flap lever to 2
3.7 Reduce airspeed to 140 Knots
3.10.1 Check current flap setting
3.3.1 Check current flap setting
3.3.2 Move 'flap' lever to 1
3.6.1 Check current flap setting
3.6.2 Move 'flap' lever to 3
3.4.1 Check current airspeed
3.4.2 Dial the 'Speed/MACH' knob to enter 150 on the IAS/MACH display
3.7.1 Check current airspeed
3.7.2 Dial the 'Speed/MACH' knob to enter 140 on the IAS/MACH display

Figure 3. Section of HTA for illustrative purposes – Land Aircraft X at New Orleans using Autoland system.
A full copy of this analysis can be found in Marshall, Stanton, Young, Salmon, Harris, Demagalski, Waldmann and Dekker[16].

3.0 BENCHMARKING HET METHODOLOGY AGAINST EXISTING HEI METHODS

A short list of 32 prospective HEI methods for subsequent evaluation of their suitability for flight deck design and certification was compiled[16]. These methods were then further analysed under ten broad headings to aid down-selection. The ten headings were:

1. what the method measures – error, performance times, mental workload etc.
2. which domain the method was originally developed for – nuclear power, aviation, HCI etc.
3. whether or not the method required domain experts to conduct an analysis.
4. training time required – low, medium or high.
5. resource usage – amount of time and resources spent conducting an analysis with the method – low, medium or high.
6. links to any other methods – whether or not the method requires the input of another method to perform an analysis e.g. SHERPA requires a hierarchical task analysis (HTA) to be conducted first.
7. consistency – would the method produce the same results when used by different analysts – low, medium or high.
8. validation studies – had the method been subjected to any validation studies in the literature.
9. main strengths.
10. main weaknesses.

From this process the three most suitable HEI methods against which to benchmark HET were identified as SHERPA[17] (systematic human error reduction and prediction approach); human error HAZOP[18] (hazard and operability study); and HEIST[19] (human error In systems tool). A full description of the application of these techniques to the benchmarking scenario described in Section 4.0 can be found in Marshall, Stanton, Young, Salmon, Harris, Demagalski, Waldmann and Dekker[16].

3.1 SHERPA – Systematic human error reduction and prediction approach

SHERPA uses hierarchical task analysis (HTA) in conjunction with an error taxonomy to identify credible errors associated with a sequence of human activity. This is based on the judgement of the analyst. SHERPA is conducted on each bottom level task step taken from the HTA (q.v. HET). Using judgement, the analyst uses the SHERPA error taxonomy to classify each task step into one of the five following behaviour types: action; retrieval; checking; selection; and Information communication. The analyst then uses the taxonomy and domain expertise to determine any credible error modes for the task in question. For each credible error the analyst provides a description of the form that the error would take. Next, the analyst has to determine any consequences associated with the error and any future task steps that might lead to recovery from the error. An ordinal probability of the error occurring is assigned (low, medium or high), together with criticality of the error (low, medium or high) and any potential design remedies (i.e. how the interface design could be modified to eradicate the error) are recorded.

The main strengths of the SHERPA method are that it provides a structured and comprehensive approach to error prediction, gives an exhaustive and detailed analysis of potential errors and also the SHERPA error taxonomy prompts the analyst for any potential errors. Furthermore, a number of studies have shown encouraging validity and reliability data for the SHERPA technique.

In a comparative study of six human error identification techniques[20] SHERPA achieved the highest overall rankings on a number of assessment criteria for its performance (comprehensiveness, accuracy, consistency, theoretical validity, usefulness and acceptability). In a further study[21] the method also performed well in predicting subsequent actual errors. Empirical studies have shown that SHERPA has acceptable test/re-test reliability[22,23] and has performed reasonably well in an aviation environment[23]. However, SHERPA's main weaknesses are that it is both tedious and time consuming to perform and it does not consider the cognitive components of the error mechanisms. The method's consistency when used by different analysts can also be questioned.

3.2 Human error hazard and operability study (HAZOP)

HAZOP is a well-established engineering approach that developed in the late 1960s[24] for use in process design audit and engineering risk assessment[25]. Originally applied to engineering diagrams the HAZOP technique involves the analyst applying guidewords (e.g. 'not done'; 'more than' or 'later than') to each step in a process to identify potential problems. A more human factors orientated version emerged in the form of the human error HAZOP, aimed at dealing with human error issues[24]. Whalley[18] created a new set of guidewords, more applicable to human error. These human error guidewords (e.g. 'not done', 'repeated', 'less than', 'more than', etc) are applied to each step in an HTA to determine any credible errors (i.e. those judged by the subject matter expert to be possible). Once the analyst has recorded a description of the error, the consequences, cause and recovery path of it are also recorded. Finally, the analyst then records any design improvements to remedy the error.

HAZOP has been used emphatically in many domains. HAZOP style techniques have received wide acceptance by both the process industries and the regulatory authorities[27]. human error HAZOP is relatively quick, easy to use and an exhaustive technique. However, similar to the SHERPA, its main weaknesses are that it is time consuming and also that some of the errors predicted using the tool are questionable.

3.3 HEIST – Human error identification in systems tool

HEIST[17] is a technique that has similarities to a number of traditional HEI techniques (e.g. SHERPA). HEIST can be used by the analyst to identify external error modes by using tables that contain various error prompt questions. There are eight tables in total, under the headings of Activation/Detection; Observation/Data collection; Identification of system state; Interpretation; Evaluation; Goal selection/Task definition; Procedure selection and Procedure execution. The analyst applies each table to each task step from an HTA and determines whether any errors are credible. For each credible error, the analyst then records the system cause or psychological error mechanism and error reduction guidelines (which are all provided in the HEIST tables) and also the error consequence.

The method's main advantage is the use of error identifier questions which prompt the analyst for potential errors. However, the method suffers from a number of domain transfer problems due to the fact that it was developed for the nuclear power industry. HEIST error identifier prompts and error reduction guidelines are quite process control specific, pertaining mostly the nuclear power industry. Herein lies the catch in developing a reliable and valid predictive error technique. If the error identifiers do not key onto the tasks in any meaningful way it is difficult to generate credible errors for any given situation. The more domain appropriate the error identifier prompts and error reduction guidelines, the less generalisable the technique becomes. Generic error identifiers can be applied across domains but are unlikely to perform as well those using a domain as a specific taxonomy. Conversely, domain specific prompts and guidelines are likely to be successful in that domain but are unlikely to perform well across domains. Indeed, as the HET error taxonomy has been developed exclusively for flight decks, similar to HEIST, it also is unlikely to work well in other domains. Furthermore, HEIST can also be time consuming to perform.

4.0 EMPIRICAL BENCHMARKING STUDY

The benchmarking study progressed in three stages. Firstly, the HET HEI technique, together with the SHERPA, human HAZOP and HEIST methodologies were applied to the task of conducting an approach and landing in a modern, highly automated, glass cockpit commercial airliner (Aircraft X). This produced predictions of the errors likely to occur. This was done by the same analysts on two occasions approximately one month apart. Secondly (using an independent research team) using a questionnaire, low-level error data were collected from flight crew currently flying Aircraft X concerning the errors that they had made during the approach and landing flight phase. Kirwan[28] noted that a fundamental problem when validating formal error identification techniques is obtaining ecologically-valid and reliable criterion data. Accidents are very infrequent events and investigation reports do not contain sufficient detail to establish the design-induced errors that contributed to the sequence of events. Incident data are more abundant, however, these reports contain even fewer details about the pilots' actions and any potential shortcomings on the flight deck which potentially provoked design-induced errors. As a result, a self-completion questionnaire had to be used for this task. Finally, once the above two data collection stages were completed, the final stage was to compare the error data with the predictions made by the four different techniques using a signal detection paradigm to assess the predictive validity of the method[29,30].

4.1 Stage One: HEI Predictions using HET, SHERPA, human HAZOP and HEIST

An HTA of a fully-coupled autoland approach to New Orleans airport undertaken in Aircraft X was performed. This consisted of some 22 subtasks under the main headings of setting up for approach, lining up for the runway, and preparing the aircraft for landing. The approach and landing considered was completely normal with no non-routine procedures included. This task analysis formed the basis of the following formal error prediction analysis. An extract of this HTA for illustrative purposes is included in Fig. 3. The full HTA can be found elsewhere[16].

Thirty-seven graduate engineering participants were trained in one of the HEI methods (eight trained in HET; nine in SHERPA and human HAZOP and 11 trained in HEIST). No participants had any prior experience of either civil aviation or HEI techniques. When the instructors were satisfied that the training was completed, the main task was introduced. This required participants to make predictions of the errors that pilots could make in the autoland task.

To make their predictions, participants were given an HTA of the autoland task developed by the authors (described previously); a demonstration of performing an autoland using Microsoft flight simulator; the relevant HEI taxonomies; and colour photographs of Aircraft X's flight control unit, flap levers, landing gear lever, speed brake, primary flight displays, and an overview of the flight deck.

Participants were required to make predictions of the pilot errors on two separate occasions, separated by a period of four weeks. This enabled intra-analyst reliability statistics to be computed. The predictions made were compared with error data reported by pilots using autoland (as described in the following section).

Table 1
Percentage of pilots reporting each of the ten most frequent design induced errors they had made (or knew about) when flying the approach and landing phase in Aircraft X

ITEM	ME	OTHER
Airspeed		
Initially, dialled in an incorrect airspeed on the Flight Control Unit by turning the knob in the wrong direction	39·1%	37·0%
Having entered the desired airspeed, pushed or pulled the switch in the opposite way to the one that you wanted	26·1%	26·1%
Adjusted the heading knob instead of the speed knob	78·3%	65·2%
Altitude		
Entered an incorrect altitude because the 100/1000 feet knob wasn't clicked over	26·1%	28·3%
Heading		
Entered a heading on the Flight Control Unit and failed to activate it at the inappropriate time	34·8%	34·8%
Failed to check HDG (Heading) mode was active	23·9%	19·6%
Approach System		
Tried to engage APPR (Approach) mode too late so that it failed to capture	28·3%	30·4%
Failed to check APPR was active	28·3%	30·4%
Glideslope		
Failed to monitor the glide slope and found that the aircraft had not intercepted it	39·1%	52·2%
Other		
Had an incorrect barometric air pressure set	45·7%	45·7%

4.2 Stage 2: Collection of error data

From the approach and landing HTA for Aircraft X a list was compiled of all the possible errors that could be made during the landing phase of flight using the flight control unit (FCU) as the main controlling interface. Several additional system interfaces were also included such as the speed brake and flaps. A further supplementary list of potential errors was developed from observations made during a series of orientation flights on Aircraft X and comments from interviews with type rated pilots. These data were used to develop a design induced error questionnaire specific to Aircraft X.

The questionnaire was designed to elicit a comprehensive picture of the low-level errors pilots recalled making on the flight deck while flying a fully-coupled autoland approach and landing when flying Aircraft X. To achieve this, respondents were not only asked if they had ever made the error themselves but also if they knew of a fellow pilot who had made the same error. A simple 'yes/no' response format was used. As it was highly probable that the list of questions was not exhaustive, space was provided to report additional errors or for further comments to be given.

Following a pilot administration of the instrument to a sample of senior pilots to check for errors and to refine the wording of the survey items, the final questionnaire was sent to pilots flying the aircraft in three UK Airlines.

The final instrument contained 70 questions concerned with the pilot interfaces on the flight deck. The survey instrument was divided into 13 subsections. It comprised of items regarding speed brake setting (7 questions); flap selection (10 questions); lowering landing gear (1 question); airspeed (11 questions); checking ALT (altitude capture) is engaged (1 question); altitude (8 questions); changing headings (4 questions); checking HDG (heading) mode is engaged (1 question); engaging the approach system (4 questions); checking APPR (approach) mode is engaged (1 question); tracking the localiser (7 questions); tracking the glideslope (2 questions); and other miscellaneous items (13 questions). On return of the questionnaires, several additional interviews were conducted to clarify the additional comments received on many survey instruments.

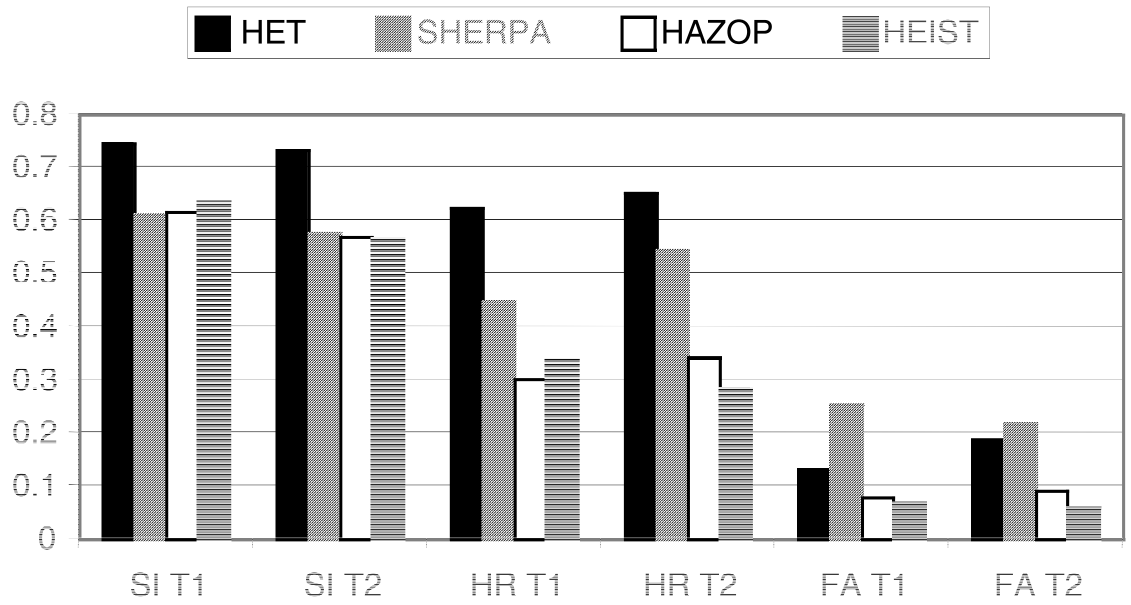
		Errors Reported	
		YES	NO
	YES	HIT	FALSE ALARM
	NO	MISS	CORRECT REJECTION

Figure 4. Signal Detection matrix used to determine the frequency of hits, misses, false alarms and correct rejections.

4.3 Stage 3: Comparison of error data with HEI predictions

4.3.1 Error data sample

Forty-six pilots responded to the survey. Captains comprised 45 7% of the sample; First Officers 37% of the sample and the remainder were either Training Captains (13 3%) or failed to state their position (two respondents). Experience ranged from less than 2,000 hours to over 16,000 with a mean of 6,832 hours (standard deviation of 4,524 hours). Type specific experience ranged from less than 1,000 hours to over 5,000 hours (mean 1,185 hours; standard deviation 1,360 hours).



Caption SI T1 – Sensitivity Index on first application

SI T2 – Sensitivity Index on second application

HR T1 – Hit Rate on first application

HR T2 – Hit Rate on second application

FA T1 – False Alarm rate on first application

FA T2 – False Alarm rate on second application

Figure 5. Bar graph showing mean Sensitivity Index, Hit Rate and False Alarm Rate for each method on first and second application.

4.3.2 Error data

Fifty-seven different types of error were reported, either as responses to the structured survey items or in the additional comments section of the questionnaire. These are summarised in Harris, Stanton, Marshall, Young, Demagalski and Salmon(23) and are described in detail in Marshall, Stanton, Young, Salmon, Harris, Demagalski, Waldmann and Dekker(16). For the purposes of illustration, only the ten most frequently reported errors are summarised in Table 1. In this table the column 'ME' contains the percentage of respondents who had made the error in question themselves; the column labelled 'OTHER' contains the data indicating that they had seen someone else make the error.

4.3.3 Analysis

The predictions made by each of the HEI techniques were compared with error data collected. This enabled validity statistics to be computed using a signal detection paradigm(29,30). This approach has provided a framework for testing the power of formal human error identification methods(22,29). In addition to comparing correct predictions of error with actual errors (hits) it identifies type I analytical errors (a miss: when the error analyst predicts the error will not occur and it does) and type II analytical errors (a false alarm: when the error analyst predicts that there will be an error and there is not). This is described in Fig. 4. 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 technique's predictions are. The formula used to calculate SI is given in equation 1, taken from Stanton and Stevenage(22). The results comparing the HEI predictions with actual errors are given in Fig. 5.

$$\mathrm{SI} = \left(\frac{\left(\frac{\text{Hit}}{\text{Hit + Miss}} \right) + 1 - \left(\frac{\text{False Alarm}}{\text{FA + Correct Rejection}} \right)}{2} \right) \qquad \ldots (1)$$

A Kruskal-Wallis One-Way analysis of variance test was undertaken to establish if the observed differences in the sensitivity index were significantly greater than those expected by chance. The difference in the sensitivity index between the four methods was statistically significant (χ^2, 3 *df* = 29·23, *p*<0·0001) suggesting a genuine difference in the sensitivity index between the four HEI methods. To explore specific differences between pairs of methods a post-hoc Mann-Whitney *U* test was used. The sensitivity index for the HET group was significantly higher than the SI for the SHERPA group (*U* = 19, *p*<0·0001); the human error HAZOP group (*U* = 19, *p*<0·0001) and the HEIST group (*U* = 19, *p*<0·0001). It can be concluded that participants using the HET methodology were significantly more accurate in their predictions than participants using any of the other methods. Furthermore there were no statistically significant differences between the remaining comparisons of the methods.

A Wilcoxon matched pairs signed ranks test was used to determine if there was a statistically significant difference between the participant SI scores, hit rate and/or false alarm rate, on first and second application of the HEI methods. It was found that there was no statistically significant difference between the participants' SI scores (irrespective of HEI methodology) on first and second application of the methodology ($z = -1{\cdot}27$, $p > 0{\cdot}05$). There was, however, a statistically significant difference between the Hit Rate scores at time 1 and time 2 ($z = -2{\cdot}26$, $p < 0{\cdot}05$). The participant hit rate scores were significantly higher on the second application of the methods. There was also a statistically significant difference between the False Alarm Rate scores on the first and second applications of the methods ($z = -2{\cdot}32$, $p < 0{\cdot}05$). The participant false alarm scores were statistically significantly higher on the second application.

5.0 DISCUSSION

5.1 General discussion

The objective of this study was to demonstrate the utility of the newly developed HET methodology for predicting potential design induced pilot error on a landing task and compare the technique's performance against three contemporary HEI methods (SHERPA, human error HAZOP and HEIST). The study also aimed to demonstrate that participant SI scores, hit rates and false alarm rates would improve significantly when the analysts performed the same analysis for a second time.

In terms of accuracy of error predictions, participants using the HET methodology were the most accurate in their error predictions for the flight task analysed. Of the other three methods, SHERPA, human error HAZOP and HEIST, there were no statistically significant differences between the accuracy of the error predictions made. It should be reiterated that the HET error mode taxonomy was developed from actual pilot error incidences and from an exhaustive analysis of contemporary error prediction, which should make it the most appropriate technique for use on civil flight decks. The other HEI methods (SHERPA, human error HAZOP and HEIST) suffer in that they utilise error mode taxonomies developed specifically for tasks undertaken in nuclear power plant control rooms. The performance of the four HEI methods is largely due to the constraints imposed on the possible errors that can be predicted by the error mode taxonomies they employ. 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' (one of the more frequent errors reported by pilots – see Table 1). The HET checklist error taxonomy, however, prompts the analyst for this error, with the error mode 'Task executed on wrong interface element' (see Fig. 2).

The HET methodology is simple to learn and use. Participants using HET were able to pick the method up easier than participants using the other three methods. It expected that the SI scores would improve between the first and second application of the method, however the results demonstrated that there was no statistically significant difference between the participant scores. Further analysis of the results revealed that although hit rate scores were found to significantly increase on the second application of the methods (i.e. participants were predicting more hits and less misses) it was also found that false alarm rate scores also increased significantly (participants were predicting more false alarms and making less correct rejections). As a result of this, the SI scores did not improve significantly on the second application of the method. Analysts were improving at predicting more of the actual errors reported by the pilots (hits) but also predicting significantly more errors that were not reported by the pilots (false alarms), thus making less correct rejections. There is an alternative view, though, that a false alarm is an error waiting to happen. As a result it would be imprudent to dismiss the possibility of such an error ever occurring simply because it has not yet happened.

As a slight caveat, it can be argued that it is perhaps a little difficult for analysts with little or no experience of the task to make definitive judgements on the probability of an error or its ultimate criticality. Further work is required to establish the reliability and sensitivity of the HET methodology when used by certification test pilots and experienced design engineers. It should also be noted that at the moment the methodology does not encompass the potential error detection/error mitigation processes afforded when flying on a multi-crew flight deck. However, as it is a requirement that for certification purposes all aircraft are capable of operation by a single pilot without imposing undue workload, this was not regarded as a high priority item in the development of HET.

Following comments from subject matter experts on the HET methodology several slight revisions are envisaged for the next version of the technique, including formally including some estimate of the subsequent detection of an error on the recording form (figure 2) and a method of prompting analysts to comment specifically on the ergonomic inadequacies in the pilot interface that promoted any predicted errors.

At the present time it is unlikely that human probabilistic risk assessments can meet the requirements of the aircraft certification process, as noted by the FAA(14). However, the present results indicate that existing human error identification techniques (SHERPA, human error HAZOP and HEIST) developed for use in other domains (e.g. nuclear power, petrochemical, manufacturing and process industries) can be applied with some success in an aviation context. The HET technique, though, shows even higher criterion-referenced validity than these and it can be concluded that with a little further development it should provide a useful diagnostic tool as an adjunct to the proposed human factors certification process(7-12).

6.0 CONCLUSIONS

This paper demonstrates that HET can be applied as a flight deck design evaluation tool, although it is acknowledged that the initial HTA may be time consuming. Ideally, the analyst applying HET should have some knowledge of the skills and procedures required to fly an aeroplane (although that was not the case in this study – none of the analysts has any formal aviation knowledge), or at least experience of the systems used on the flight deck. However, it is likely that with only moderate training, certification test pilots could achieve even higher validity coefficients than those reported in this paper. Future research should aim to conduct tests of the HET methodology with subject matter experts in a variety of aviation domains to determine the extent of cross-validation.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Boeing Commercial Airplanes Group. Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations 1959-1999, 2000 Boeing, Seattle WA, USA.
2. Civil Aviation Authority. Global Fatal Accident Review 1980-96 (CAP 681), 1998, Civil Aviation Authority, London.
3. CHAPANIS, A. *The Chapanis Chronicles: 50 years of Human Factors Research, Education, and Design*, Aegean Publishing Company, 1999, Santa Barbara, CA, USA.
4. STANTON, N.A. and BABER, C. Error by design: methods for predicting device usability. *Design Studies*, 2002, **23**, pp 363-384.
5. WOODS, D. and SARTER, N. Learning from automation surprises and going sour accidents. institute for ergonomics, Report ERGO-CSEL-98-02, 1998, NASA Ames CA, USA.
6. Federal Aviation Administration. Report on the Interfaces between Flightcrews and Modern Flight Deck Systems, Federal Aviation Administration, Washington DC, USA, 1996.
7. 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 22 July 1999, **64**, (140).
8. Joint Aviation Authorities. Joint airworthiness requirements (change 15): Part 25 – Large aeroplanes, Hoofdorp: Joint Aviation Authorities, 2000.
9. US Department Of Transportation. Federal Aviation Regulations, (Part 25 – Airworthiness Standards). Revised 1 January 2003. US Department Of Transportation, Washington, DC, USA, 2003.
10. European Aviation Safety Agency. Certification Specification 25 CS 25 – Large Aeroplanes. www.easa.eu.int/doc/Agency_Measures/Certification_Spec/decision_ED_2003_02_RM.pdf (Accessed 20 July 2005). Cologne: European Aviation Safety Agency, 2003.
11 Joint Airworthiness Authorities. Human factors aspects of flight deck design: Interim Policy Paper INT/POL/25/14, Joint Airworthiness Authorities, Hoofdorp, 2001.
12 European Aviation Safety Agency (2004). Notice of Proposed Amendment 15/2004 amending the annex to decision no. 2003/2/RM on certification specifications, including airworthiness codes and acceptable means of compliance for large aeroplanes (CS-25). www.easa.eu.int/doc/Rulemaking/NPA/NPA_15_2005.pdf (Accessed 20 July 2005). Cologne, European Aviation Safety Agency, 2005.
13 Human Factors National Advisory Committee for the DTI Innovation and Growth Team. Gaining competitive advantage through human factors: A guide to the civil aerospace industry, 2003, London, Department of Trade and Industry.
14. Federal Aviation Administration. Advisory Circular: System Design and Analysis (AC 25.1309-1A), Federal Aviation Administration, 1998, Washington DC, USA.
15. ANNETT, J. *Hierarchical Task Analysis*, in, STANTON, N.A. HEDGE, A. SALAS, E. HENDRICK, H. and BROOKHAUS K. (Eds) *Handbook of Human Factors and Ergonomics Methods*, 2005, London, Taylor & Francis: London.
16. MARSHALL, A., STANTON, N., YOUNG, M., SALMON, P., HARRIS, D., DEMAGALSKI, J., WALDMANN, T. and DEKKER, S. 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.
17. EMBREY, D.E. SHERPA: A systematic human error reduction and prediction approach. Paper presented at the International Meeting on Advances in Nuclear Power Systems, 1986, Knoxville, Tennessee, USA.
18. WHALLEY, A. Minimising the cause of human error, in, KIRWAN B. and AINSWORTH, L.K. (Eds) *A Guide to Task Analysis*, 1988, London, Taylor and Francis.
19. KIRWAN, B. *A Guide to Practical Human Reliability Assessment, London*, Taylor and Francis, 1988.
20. KIRWAN, B. *Human Reliability Assessment*, in, J.R. WILSON and E.N. CORLETT (Eds), *Evaluation of Human Work*, 1990, London, Taylor and Francis, pp 706-754.
21. KIRWAN, B. Human error identification in human reliability assessment. Part 2: detailed comparison of techniques, *Applied Ergonomics,* 1992, **23**, pp 371-381.
22. STANTON, N.A. and STEVENAGE, S.V. Learning to predict human error: issues of reliability, validity and acceptability, *Ergonomics* 1998, **41**, pp 1737-1756.
23. HARRIS, D., STANTON, N.A., MARSHALL, A., YOUNG, M.S., DEMAGALSKI, J. and SALMON, P.M. Using SHERPA to predict design-induced error on the flight deck. *Aerospace Science and Technology*, 2005, **9**, pp 525-532.
24. SWANN, C.D. and PRESTON, M.L. Twenty five years of HAZOPs. *J loss prevention in the Process Industries*. 1995, **8**, pp 349-353.
25. KIRWAN, B. Human error identification in human reliability assessment. Part 1: Overview of approaches. *Applied Ergonomics*, 1992, **23**, pp 299-318.
26. KIRWAN, B. and AINSWORTH, L.K. *A Guide to Task Analysis*, 1988, Taylor and Francis, London, 1988.
27. ANDREWS, J.D. and MOSS, T.R. *Reliability and Risk Assessment*, 1993, London, Professional Engineering Publishing.
28. KIRWAN, B. Validation of three Human Reliability Quantification Techniques – THERP, HEART and JHEDI: Part I- Technique Descriptions and Validation Issues, *Applied Ergonomics* 1996, **27**, pp 359-374.
29. BABER, C. and STANTON, N.A. *Human Error Identification Techniques Applied to Public Technology: Predictions Compared with Observed use*, 1996, Applied Ergonomics, **27**, pp 119-131.
30. MACMILLAN, N.A. and CREELMAN, C.D. *Signal Detection Theory: a user's guide,* 1991, Cambridge, Cambridge University Press.

ENDNOTE

Paul Salmon is now at the Monash University Accident Research Centre, Clayton, Victoria 3800, Australia. Jason Demagalski is now employed by National Air Traffic Services, NATS Corporate Technical Centre, Whiteley, Fareham, Hampshire, PO15 7FL.