Error by design: methods for predicting device usability

Neville A. Stanton[a] and Christopher Baber[b]

[a] Department of Design, Brunel University, Runnymede Campus, Egham, Surrey TW20 0JZ, UK
[b] Kodak/Royal Academy Educational Technology Group, School of Electronic and Electrical Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

Abstract

This paper introduces the idea of predicting 'designer error' by evaluating devices using Human Error Identification (HEI) techniques. This is demonstrated using Systematic Human Error Reduction and Prediction Approach (SHERPA) and Task Analysis for Error Identification (TAFEI) to evaluate a vending machine. Appraisal criteria which rely upon user opinion, face validity and utilisation are questioned. Instead a quantitative approach, based upon signal detection theory, is recommended. The performance of people using SHERPA and TAFEI are compared with heuristic judgement and each other. The results of these studies show that both SHERPA and TAFEI are better at predicting errors than the heuristic technique. The performance of SHERPA and TAFEI are comparable, giving some confidence in the use of these approaches. It is suggested that using HEI techniques as part of the design and evaluation process could help to make devices easier to use.

Author Keywords: design methods; errors; psychology of design; user behaviour

We are all familiar with the annoyance of errors we make with everyday devices, such as switching on an empty kettle, or making mistakes in the programming sequence with video cassette recorders. People have a tendency to blame themselves for ‘human error’. However, the use and abuse of the term has led some to question the very notion of ‘human error’ [1]. ‘Human error’ is often invoked in the absence of technological explanations. Chapanis [2] wrote that back in the 1940's he noted that ‘pilot error’ was really ‘designer error’. Chapanis was interested in why pilots often retracted the landing gear instead of the landing flaps after landing. On further investigation, he found that the designer had put two identical toggle switches side-by-side, one for the landing gear the other for the flaps. Chapanis proposed that either the controls should be separated and coded or the landing gear switch should be rendered inoperable after landing. This was a challenge to contemporary thinking at that time, and shows that design is all important in human error reduction. In other words, the design of a product can yield the potential for error. Half a century after Chapanis's observations, the idea that one can design error-tolerant devices is beginning to gain credence [3 and 4]. One can argue that human error is not a simple matter of one individual making one mistake, so much as the product of a design which has permitted the existence and continuation of specific activities which could lead to errors [5]. Concerns arise over the cost of design errors, both in terms of the frustration of device users and the potential financial burden that may be incurred [4]. The problem is, what tools do designers' possess that would enable them to anticipate the error potential of a device? Fortunately, a parallel stream of activity has been developing human error identification techniques that might just be up to the task [6].

Predicting human error may strike the reader, at first sight, as implausible. After all, one school of thought assumes that if anything can go wrong, it will. This implies that the prediction of error would require an infinite collection of possible things that could go wrong. However, if we know an activity that is to be performed, and the characteristics of the product being used, then it should be possible to indicate the principle types of errors which may arise. This is the general approach taken by all methods aimed at predicting human error: first define what actions need to be performed and then indicate how these actions might fail. Notice that the aim is not necessarily to predict all errors, rather to predict the most likely or the most annoying. Techniques have been developed for the detailed and systematic assessment of a person's activities. A structured approach enables an analyst to identify potential points in tasks where errors could have significant negative consequences. From this assessment, preventive strategies can be sought to minimise the consequences or reduce the likelihood of error.

Psychologists have been investigating the origins and causes of human error since the dawn of the discipline [5]. Traditional approaches suggested that error was an individual phenomenon, the individual who appears responsible for the error. Indeed, so-called ‘Freudian slips’ were treated as the unwitting revelation of intention: errors revealed what a person was really thinking but did not wish to disclose [5]. More recently, error research in the cognitive tradition has concentrated upon classifying errors within taxonomies and determining underlying psychological mechanisms [7]. The taxonomic approaches [5 and 8] have led to the classification of errors into different forms, for example: capture errors, description errors, data driven errors, association activation errors and loss of activation errors. Reason [5] and Wickens [9] identify psychological mechanisms implied in error causation, for example: the failure of memory retrieval mechanisms in lapses, poor perception and decision making in mistakes and motor execution problems in slips. Taxonomies offer a means of classifying what has happened, whereas consideration of psychological mechanisms offers an explanation of why it has happened. Reason [5]; in particular, has argued that we need to consider the activities of the individual and the devices they are using if we are able to consider what may go wrong. This approach does not conceive of errors as unpredictable events, rather as wholly predictable consequences of an individual's activities and device design. We feel that this is a much healthier approach to device design than the ‘autopsy’ approach of only considering errors after things have gone wrong. We have been told of manufacturers whose error analysis strategy is to keep a list of customer complaints. We will argue the case for predictive approaches to supplement this procedure.

1. Research issues for human error identification

An abundance of methods for identifying human error exist, some of which may be appropriate for the analysis of consumer products. In general, most of the existing techniques have two key problems. The first of these problems relates to the lack of representation of the external environment or objects. Typically, human error analysis techniques do not represent the activity of the device and material that the human interacts with, in more than a passing manner. Hollnagel [10] emphasises that Human Reliability Analysis (HRA) often fails to take adequate account of the context in which performance occurs. Second, there tends to be a good deal of dependence made upon the judgement of the analyst [11]. Different analysts, with different experience may make different predictions regarding the same problem (called intra-analyst reliability). Similarly, the same analyst may make different judgements on different occasions (inter-analyst reliability). This subjectivity of analysis weakens the confidence that can be placed in any predictions made.

The development of HEI techniques could benefit from the approaches used in establishing psychometric techniques as two recent reviews demonstrate [12 and 13]. In two comprehensive reviews, Bartram and colleagues compare the performance of psychometric techniques across a range of criteria such as: reliability, validity, application and analysis time, costs, skills required, comprehensiveness of documentation, and so on. The methodological approach adopted by Bartram may also be applied to the entire field of ergonomics methods [14 and 15]. There are a number of issues that need to be addressed in the analysis of human error identification techniques. Some of the judgements for the criteria developed by Kirwan [16] could be deceptive justifications of a technique’s effectiveness, as they could be based upon:

- user opinion,

- face validity,

- utilisation of the technique.

User opinion is suspect because of three main reasons. First it assumes that the user is a good judge of what makes an effective technique. Second, user opinion is based on previous experience, and unless there is a high degree of homogeneity of experience, opinions may vary widely. Third, judgements may be obtained from an unrepresentative sample. Both Kirwan [16] and Baber and Stanton's [17] studies used very small samples. Face validity is suspect because a HEI technique might not be able to predict errors just because it looks as though it might, which is certainly true in the domain of psychometrics [18]. Finally, utilisation of one particular technique over another might be more to do with familiarity of the analyst than representing greater confidence in the predictive validity of the technique. Therefore more rigorous criteria need to be developed.

2. Are structured techniques better than no technique?

One possible way of benchmarking HEI techniques is through comparison with popular evaluation approaches, such as heuristic evaluation. It has been proposed that heuristic evaluation can be performed by relatively small numbers of assessors, e.g., between 5 and 8 assessors can uncover up to 80% of usability problems[19, 20 and 21]. There have been several studies investigating the benefit of heuristic evaluation relative to other techniques. Unfortunately a review of this field will show some ambiguity and inconsistency. For instance, Nielsen and Philips [22] found that user trials were better than heuristics, whereas Jeffries *et al.* [23] found that heuristics were better than user trials. In a study comparing three techniques, Westwater and Johnson [24] found that checklists were superior to user trials, which in turn were superior to heuristics. Whilst the jury is still hung on the relative merits of heuristics there is little doubt regarding the low cost of the technique. The idea of heuristics represents an interesting benchmark for this work, in that such evaluation is proposed to be ‘quick and dirty’, but may yield useful results. Given the potential high cost of user trials, if it can be shown that HEI techniques offer quick and reliable predictions of device use, then one could seriously consider their application to product evaluation. We propose further that HEI offers benefits over heuristic evaluation in that one can evaluate products when they are in their conceptual design stage (rather than having more detailed prototypes for evaluation). Furthermore, much of the heuristic evaluation literature appears to validate predictions against the predictions themselves, i.e., when writers speak of 80% of usability problems being found, they mean either 80% of the total number of problems identified by the technique or 80% of the problems identified by an expert (also using heuristic techniques). This notion of self-validating a method strikes us as somewhat odd, and in our work we seek to validate the method using an external data source. If a technique can be shown to be a reliable predictor of human error when applied by a single analyst, and in the absence of a prototype, then it could be applied at the concept stages of design [15]. One of the aims of our approach is to develop methods to support analytical prototyping [25]. This approach is based on a model of the user. HEI techniques are not just ways of identifying error, but also methodologies for determining aberrant human performance. Finally, while the use of less than 10 evaluators might be attractive for time and cost, it is not easy to see that this small sample size can produce statistically meaningful data (the power of any test applied to such data will be relatively weak). Consequently, in study one, we set ourselves the following goals: to compare HEI with heuristic evaluation, to use an external data source for validation (in this case, data produced by actual user trials), and to use a sample size in excess of 30.

In this paper we will consider two human error identification techniques, the Systematic Human Error Reduction and Prediction Approach (SHERPA) and the Task Analysis for Error Identification (TAFEI).

3. Systematic human error reduction and prediction approach (SHERPA)

SHERPA [26] uses Hierarchical Task Analysis (HTA [27]) together with error taxonomy to identify credible errors associated with a sequence of human activity. In essence, the SHERPA technique works by indicating which error modes are credible for each task step in turn, based upon an analysis of work activity. This indication is based upon the judgement of the analyst, and requires input from a subject matters expert to be realistic. A summary of the procedure is shown in Figure 1.

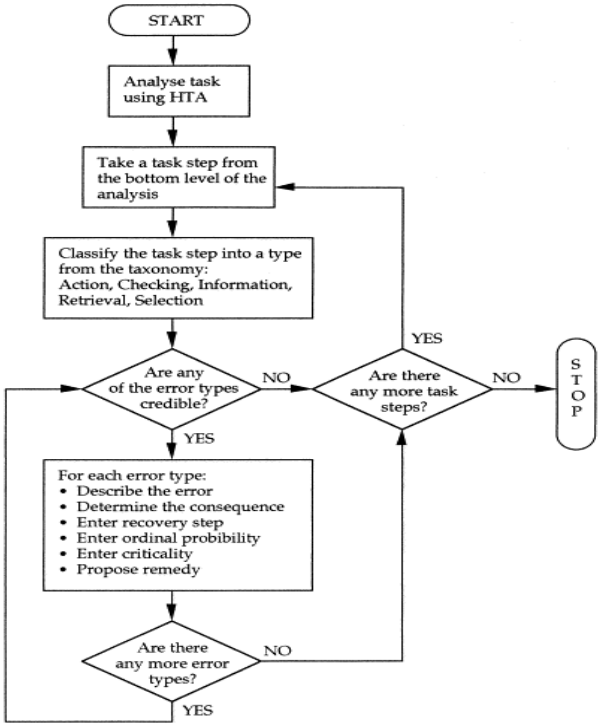

START
Analyse task using HTA
Take a task step from the bottom level of the analysis
Classify the task step into a type from the taxonomy: Action, Checking, Information, Retrieval, Selection
Are any of the error types credible?
NO
YES
Are there any more task steps?
YES
NO
STOP
For each error type:
• Describe the error
• Determine the consequence
• Enter recovery step
• Enter ordinal probibility
• Enter criticality
• Propose remedy
Are there any more error types?
NO
YES

Figure 1 Summary of the SHERPA procedure

The process begins with the analysis of work activities, using Hierarchical Task Analysis. HTA is based upon the notion that task performance can be expressed in terms of a hierarchy of goals (what the person is seeking to achieve), operations (the activities executed to achieve the goals) and plans (the sequence in which the operations are executed). An example of HTA for the purchase of confectionery from a vending machine is shown in Figure 2.

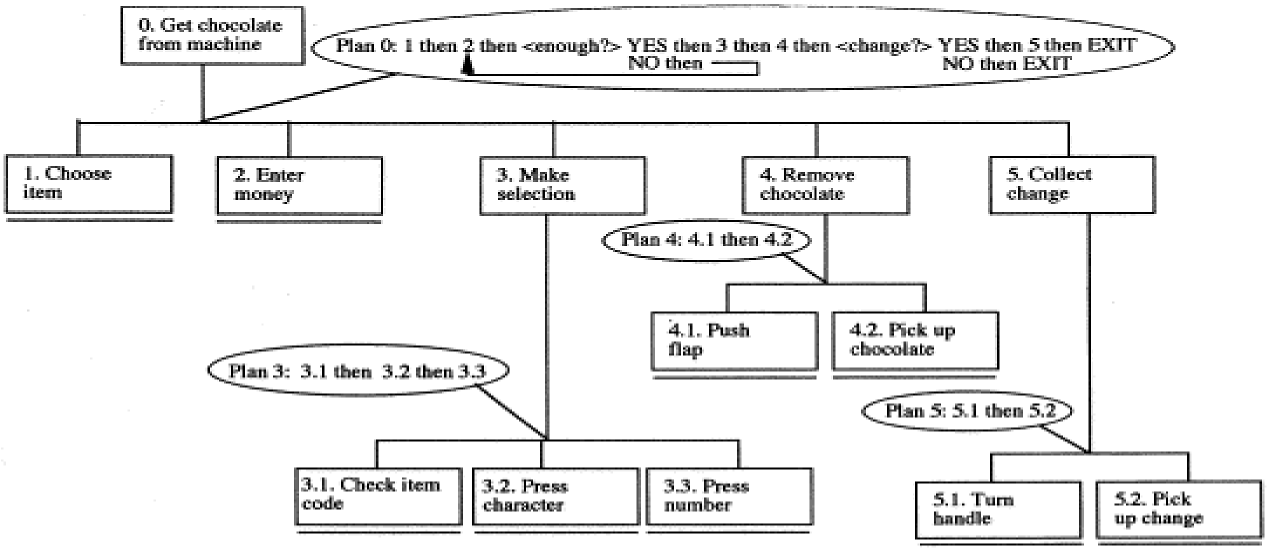
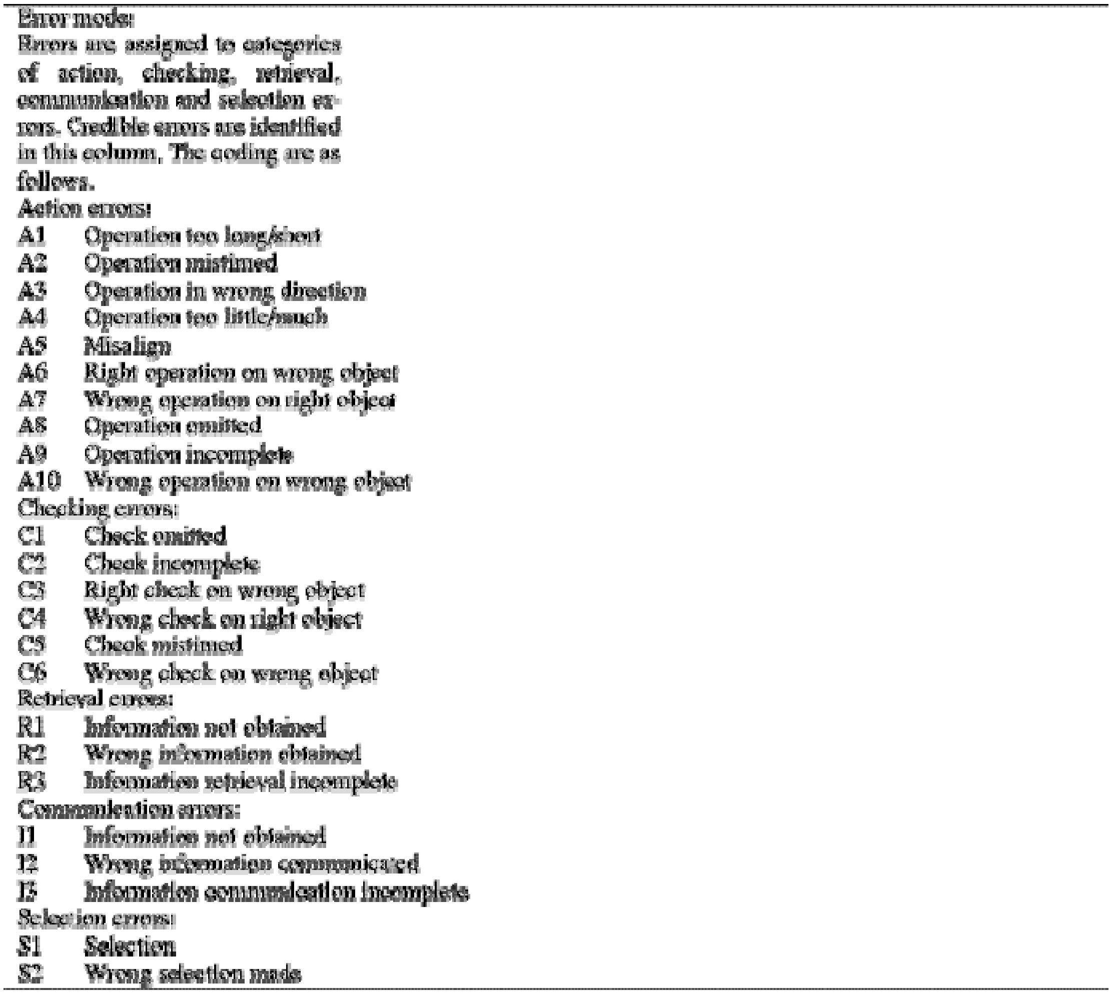


Figure 2 HTA for the purchase of chocolate from a vending machine

For the application of SHERPA, each task step from the bottom level of the analysis is taken in turn. First each task step is classified into a type from the taxonomy, into one of the following types:

- action (e.g. pressing a button, pulling a switch, opening a door),
- retrieval (e.g. getting information from a screen or manual),
- checking (e.g. conducting a procedural check),
- selection (e.g. choosing one alternative over another),
- information communication (e.g. talking to another party).

This classification of the task step then leads the analyst to consider credible error modes associated with that activity, as shown in Table 1.



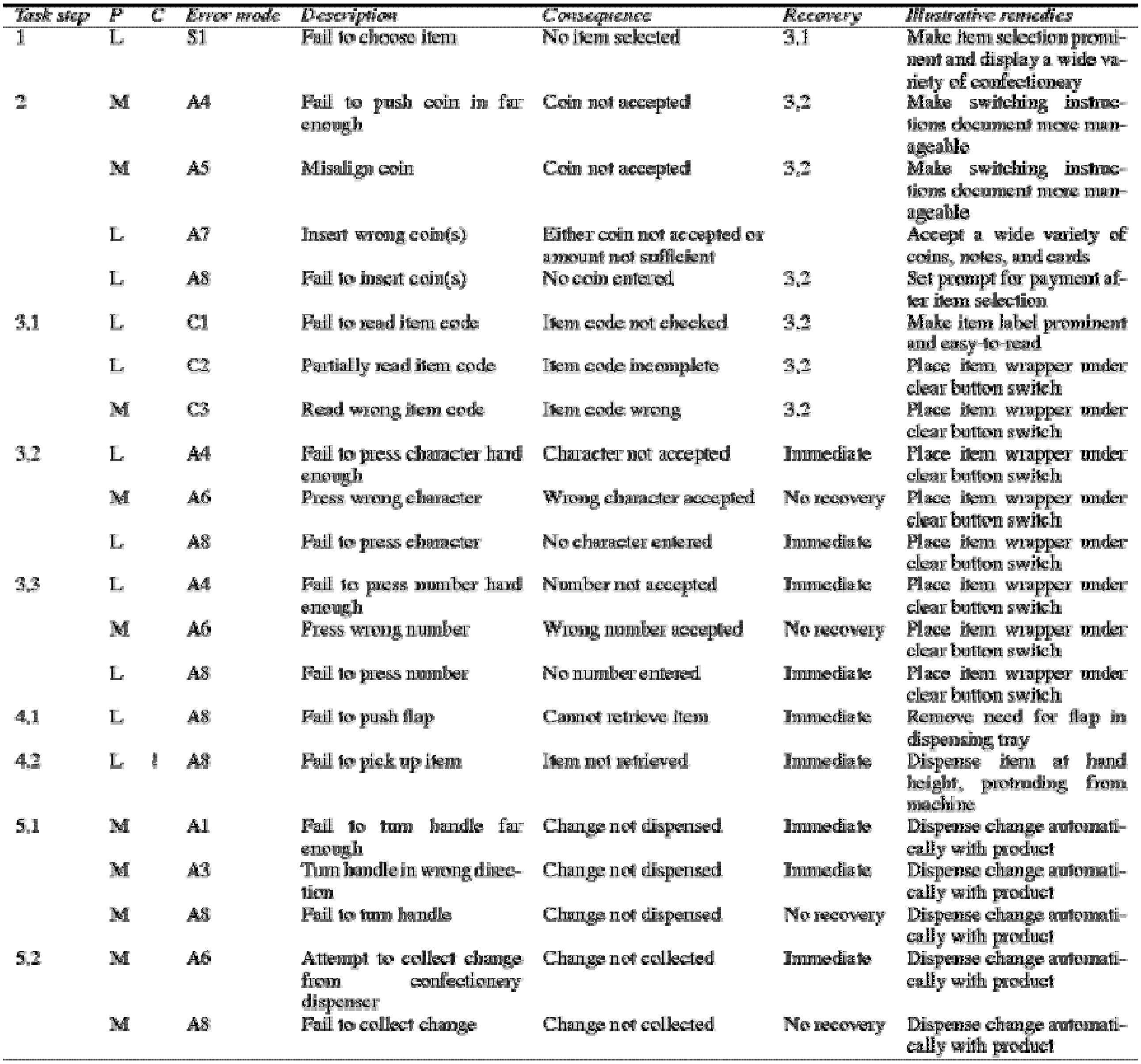
Error mode:

Errors are assigned to categories of action, checking, retrieval, communication and selection errors. Credible errors are identified in this column. The coding are as follows.

Action errors:	
A1	Operation too long/short
A2	Operation mistimed
A3	Operation in wrong direction
A4	Operation too little/much
A5	Misalign
A6	Right operation on wrong object
A7	Wrong operation on right object
A8	Operation omitted
A9	Operation incomplete
A10	Wrong operation on wrong object
Checking errors:	
C1	Check omitted
C2	Check incomplete
C3	Right check on wrong object
C4	Wrong check on right object
C5	Check mistimed
C6	Wrong check on wrong object
Retrieval errors:	
R1	Information not obtained
R2	Wrong information obtained
R3	Information retrieval incomplete
Communication errors:	
I1	Information not obtained
I2	Wrong information communicated
I3	Information communication incomplete
Selection errors:	
S1	Selection
S2	Wrong selection made

Table 1 Error modes for SHERPA

For each credible error (i.e. those judged by a subject matter expert to be possible) a description of the form that the error would take is given as illustrated in Table 2. The consequence of the error on the system needs to be determined next, as this has implications for the criticality of the error. The last four steps consider the possibility for error recovery, the ordinal probability of the error, its criticality and potential remedies. Again these are shown in Table 2.



Task step	*P*	*C*	*Error mode*	*Description*	*Consequence*	*Recovery*	*Illustrative remedies*
1	L		S1	Fail to choose item	No item selected	3,1	Make item selection prominent and display a wide variety of confectionery
2	M		A4	Fail to push coin in far enough	Coin not accepted	3,2	Make switching instructions document more manageable
	M		A5	Misalign coin	Coin not accepted	3,2	Make switching instructions document more manageable
	L		A7	Insert wrong coin(s)	Either coin not accepted or amount not sufficient		Accept a wide variety of coins, notes, and cards
	L		A8	Fail to insert coin(s)	No coin entered	3,2	Set prompt for payment after item selection
3.1	L		C1	Fail to read item code	Item code not checked	3.2	Make item label prominent and easy-to-read
	L		C2	Partially read item code	Item code incomplete	3,2	Place item wrapper under clear button switch
	M		C3	Read wrong item code	Item code wrong	3.2	Place item wrapper under clear button switch
3.2	L		A4	Fail to press character hard enough	Character not accepted	Immediate	Place item wrapper under clear button switch
	M		A6	Press wrong character	Wrong character accepted	No recovery	Place item wrapper under clear button switch
	L		A8	Fail to press character	No character entered	Immediate	Place item wrapper under clear button switch
3,3	L		A4	Fail to press number hard enough	Number not accepted	Immediate	Place item wrapper under clear button switch
	M		A6	Press wrong number	Wrong number accepted	No recovery	Place item wrapper under clear button switch
	L		A8	Fail to press number	No number entered	Immediate	Place item wrapper under clear button switch
4,1	L		A8	Fail to push flap	Cannot retrieve item	Immediate	Remove need for flap in dispensing tray
4,2	L	!	A8	Fail to pick up item	Item not retrieved	Immediate	Dispense item at hand height, protruding from machine
5,1	M		A1	Fail to turn handle far enough	Change not dispensed	Immediate	Dispense change automatically with product
	M		A3	Turn handle in wrong direction	Change not dispensed	Immediate	Dispense change automatically with product
	M		A8	Fail to turn handle	Change not dispensed	No recovery	Dispense change automatically with product
5,2	M		A6	Attempt to collect change from confectionery dispenser	Change not collected	Immediate	Dispense change automatically with product
	M		A8	Fail to collect change	Change not collected	No recovery	Dispense change automatically with product

Table 2 the SHERPA table

4. Task analysis for error identification (TAFEI)

TAFEI [3, 6 and 28] explicitly analyses the *interaction* between people and machines. TAFEI analysis is concerned with task-based scenarios. This is done by mapping human activity onto machine states. An overview of the procedure is shown in Figure 3. TAFEI analysis consists of three principal components: Hierarchical Task Analysis (HTA), State-Space Diagrams (SSDs which are loosely based on finite state machines [29]) and Transition Matrices (TM). HTA provides a description of human activity, SSD provides a description of machine activity and TM provides a mechanism for determining potential erroneous activity through the interaction of the human and the device. From this, both legal and illegal operators (called *transitions* in the TAFEI methodology) are identified.

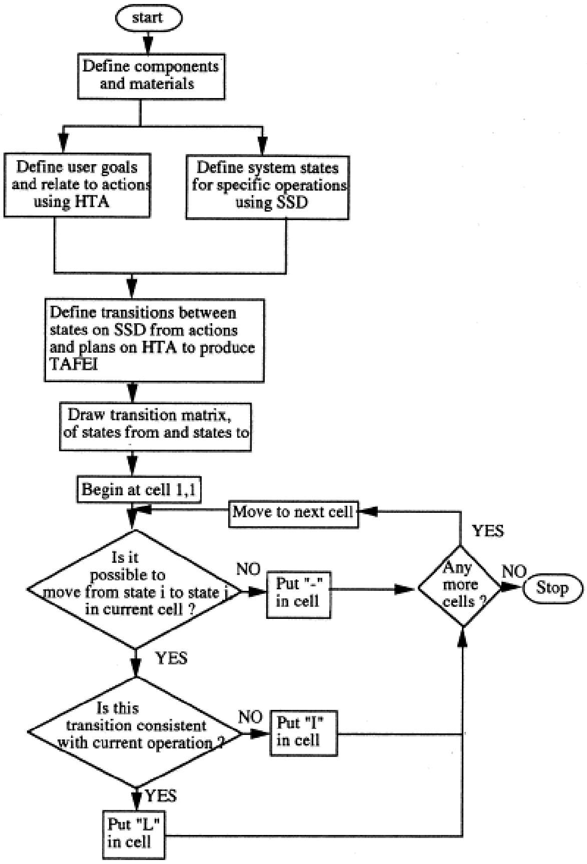

start
Define components and materials
Define user goals and relate to actions using HTA
Define system states for specific operations using SSD
Define transitions between states on SSD from actions and plans on HTA to produce TAFEI
Draw transition matrix, of states from and states to
Begin at cell 1,1
Move to next cell
YES
Is it possible to move from state i to state j in current cell ?
NO
Put "-" in cell
Any more cells ?
NO
Stop
YES
Is this transition consistent with current operation ?
NO
Put "I" in cell
YES
Put "L" in cell

Figure 3. Summary of the TAFEI procedure

In brief, the TAFEI methodology is as follows. First, the system to be addressed needs to be defined. Next, the human activities and machine states are described in separate analyses. The basic building blocks are HTA (describing human activity—see Fig. 2) and state space diagrams (describing machine activity). These two types of analysis are then combined to produce the TAFEI description of human–machine interaction, as shown in Fig. 4.

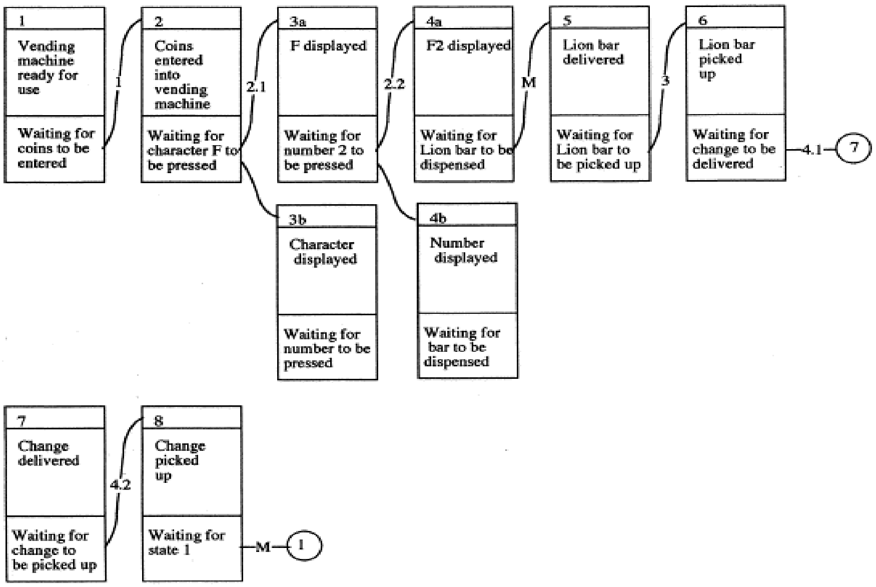
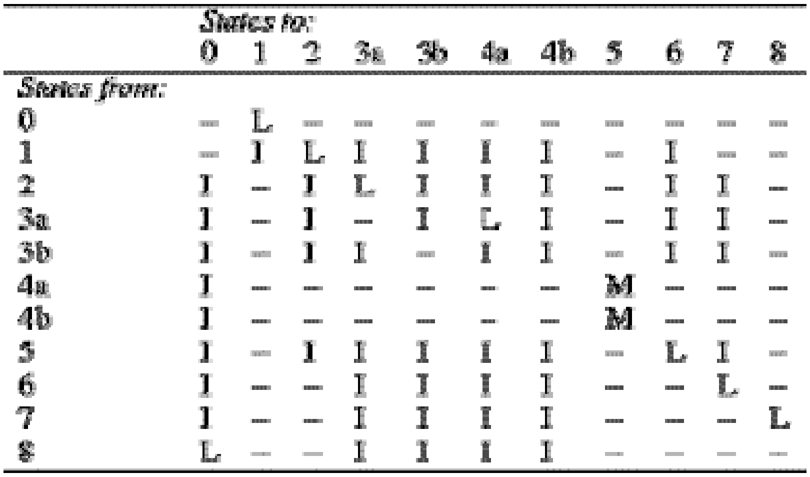

1
Vending machine ready for use
Waiting for coins to be entered
1
2
Coins entered into vending machine
Waiting for character F to be pressed
2.1
3a
F displayed
Waiting for number 2 to be pressed
2.2
4a
F2 displayed
Waiting for Lion bar to be dispensed
M
5
Lion bar delivered
Waiting for Lion bar to be picked up
3
6
Lion bar picked up
Waiting for change to be delivered
4.1
7
3b
Character displayed
Waiting for number to be pressed
4b
Number displayed
Waiting for bar to be dispensed
7
Change delivered
Waiting for change to be picked up
4.2
8
Change picked up
Waiting for state 1
M
1

Figure 4 The TAFEI description

From the TAFEI diagram, a transition matrix is compiled and each transition is scrutinised. In Table 3, each transition has been classified as 'impossible' (i.e. the transition cannot be performed), 'illegal' (the transition can be performed but it does not lead to the desired outcome) or 'legal' (the transition can be performed and is consistent with the description of error-free activity provided by the HTA), until all transitions have been analysed. Finally, 'illegal' transitions are addressed in turn as potential errors, to consider changes that may be introduced.



	States to:										
	0	1	2	3a	3b	4a	4b	5	6	7	8
States from:											
0	–	L	–	–	–	–	–	–	–	–	–
1	–	I	L	I	I	I	I	–	I	–	–
2	I	–	I	L	I	I	I	–	I	I	–
3a	I	–	I	–	I	L	I	–	I	I	–
3b	I	–	I	I	–	I	I	–	I	I	–
4a	I	–	–	–	–	–	–	M	–	–	–
4b	I	–	–	–	–	–	–	M	–	–	–
5	I	–	I	I	I	I	I	–	L	I	–
6	I	–	–	I	I	I	I	–	–	L	–
7	I	–	–	I	I	I	I	–	–	–	L
8	L	–	–	I	I	I	I	–	–	–	–

Table 3 Transition matrix for TAFEIa

5. Experiment one: semi-structured versus unstructured techniques

Experiment one aims to compare the performance of participants using SHERPA and TAFEI with those using heuristic error prediction.

5.1. Method for experiment one

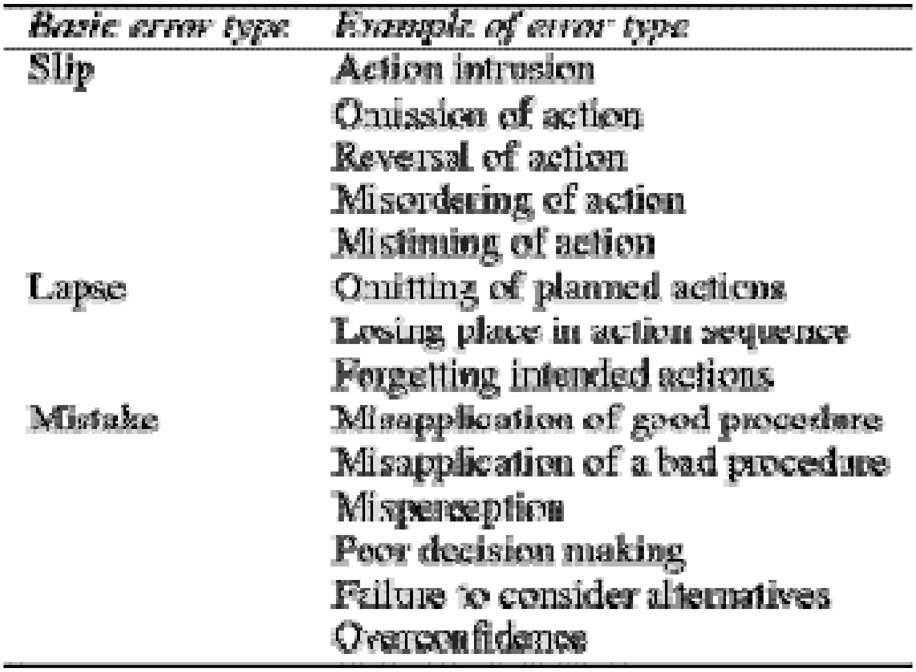
5.1.1. Participants

Three groups of participants were involved in this study. The first group consisted of 36 undergraduate students aged 19–45 years (modal age, 20 years). Of these, 24 were female and 12 were male. These participants formed the control group and received no human error identification (HEI) training.

The second and third groups consisted of 36 participants each drawn to match the pool as above. These participants acted as novice analysts using SHERPA and TAFEI. All participants were equally familiar with the machine upon which the human error analysis was conducted.

5.1.2. Materials

All of the participants were provided with Reasons [5] error classification system as part of their training (see Table 4). All TAFEI analysts were provided with a hierarchical task analysis (HTA) chart describing the action stages involved when using a vending machine to obtain a bar of chocolate and a state-space diagram of machine states. All SHERPA analysts were provided with a hierarchical task analysis (HTA) chart describing the action stages involved when using a vending machine to obtain a bar of chocolate and a taxonomy of human errors. Finally, both TAFEI and SHERPA participants were provided with a proforma for recording their error predictions.



Basic error type	*Example of error type*
Slip	Action intrusion
	Omission of action
	Reversal of action
	Misordering of action
	Mistiming of action
Lapse	Omitting of planned actions
	Losing place in action sequence
	Forgetting intended actions
Mistake	Misapplication of good procedure
	Misapplication of a bad procedure
	Misperception
	Poor decision making
	Failure to consider alternatives
	Overconfidence

Table 4 Basic error types with examples

5.1.3. Selection of device

When considering heuristic evaluation, it is often important to distinguish different types of expertise. Thus, one could be an expert in the task, the technology or the methodology [20]. Consequently, it was decided that we required task and technology that could be assumed to be familiar to participants, so that we could assume some level of expertise on these dimensions. A confectionery vending machine was chosen. This was familiar to all participants (and had been used by participants).

5.1.4. Procedure

For both groups, participants were given the scenario of buying one item (a Lion Bar, costing 24 p) from the vending machine using a 50 p coin and thus requiring change. They were required to try to predict the errors that would occur during this operation. To this end, all participants received training by means of a two hour lecture and video on human error. The training began with a general introduction to human error research based upon the work of Reason [5]. A classification system for analysis of human error was presented to distinguish between slips, lapses and mistakes. These error types were defined in terms of an Information Processing Model [9] and examples of each error type were discussed in various contexts. In particular, the link was made between product design and human error [8 and 30]. This was followed by a forty-five minute video on human error which related everyday errors to those errors found in a more unforgiving environment (i.e., the errors contributing to the Tenerife runway disaster of March, 1977). It is proposed that this training (and the classification of human error shown in Table 4) constituted a set of heuristics that participants in the Heuristic condition could apply in their evaluation. Finally, participants using the SHERPA and TAFEI techniques received specific instructions in the use of the technique via a one-hour training session. This comprised an introduction to hierarchical task analysis and an explanation of the staged approach taken by each technique, as outlined earlier. A worked example was provided and participants then proceeded to generate their own analysis of errors using a familiar everyday device (i.e., a kettle).

5.1.5. Error prediction

Participants in the heuristic group were required to indicate the errors which they thought would occur during this scenario. Participants using the SHERPA and TAFEI methods of error prediction received verbal and written training in the use of the method.

5.1.6. Error classification

The error predictions from all participants were compared to the errors actually observed in 75 independent transactions with the machine. These errors were obtained by observing people using the vending machine to purchase goods from it. These observations were recorded for analysis. Observation of the 75 transactions revealed nine discrete types of error and it was possible for more than one error type to occur within a single transaction. The transactions were observed without the prior knowledge of the user and these 75 transactions provided a sample of errors that contained all the error types that were likely from a larger set of observations. In an independent study by Baber and Stanton [17] it was shown that a data set of over 300 person–machine interactions revealed 90% of the error types within the first 20 interactions. Moreover, no novel error types were evident after 75 interactions. The comparison of predicted and observed errors yielded three dependent variables:

1. hits (predicted errors that were seen to occur),

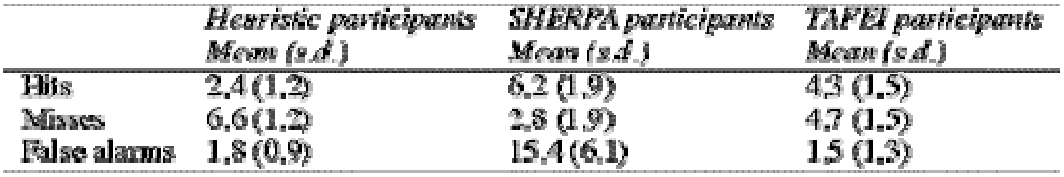
2. false alarms (predicted errors that did not occur), and

3. misses (errors that occurred but were not predicted).

The frequency of misses was obtained by subtracting the number of hits from the total number of errors observed (n=9). These three dependent variables formed the basis for subsequent analyses.

5.2. Results for experiment one

For each participant, the frequency of hits, misses and false alarms when predicting errors with a vending machine were calculated. Table 5 below summarises these data across the control group and the group using the TAFEI method for human error identification.



	Heuristic participants *Mean (s.d.)*	*SHERPA participants* *Mean (s.d.)*	*TAFEI participants* *Mean (s.d.)*
Hits	2.4 (1.2)	6.2 (1.9)	4.3 (1.5)
Misses	6.6 (1.2)	2.8 (1.9)	4.7 (1.5)
False alarms	1.8 (0.9)	15.4 (6.1)	1.5 (1.3)

Table 5 Mean hits, misses and false alarms for participants in the heuristic, SHERPA and TAFEI groups

From Table 5, the participants using the SHERPA and TAFEI technique correctly predicted more errors and missed fewer errors than those in the heuristic group. Participants in the SHERPA group predicted more hits than those in the TAFEI group, but also produced more false alarms.

These results suggest that when participants first use the SHERPA and TAFEI methods of human error identification, they are able to correctly predict more errors, and hence, miss fewer errors, than participants who use a heuristic technique. In this respect, using SHERPA and TAFEI seems to be better than an heuristic approach to error prediction. Thus we are able to confirm that using structured methods to predict human error, results in greater accuracy than using a heuristic approach, despite some claims to the contrary regarding the benefits of heuristics [31].

6. Experiment two: SHERPA versus TAFEI

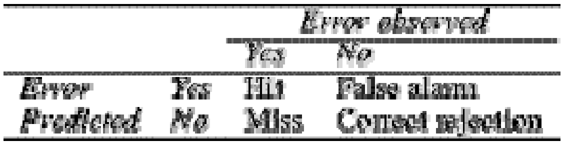
While experiment one has suggested that SHERPA and TAFEI out-perform the heuristic technique, in the next experiment we wanted to see how practice might improve performance.

6.1. Method for experiment two

The SHERPA study employed 25 undergraduates to undertake a SHERPA analysis of the task steps involved in the purchase of an item of confectionery from a vending machine. Following a period of instruction, participants undertook the SHERPA analysis on three separate occasions. This was done to test the reliability of the approach (i.e. the consistency of the analysis over time). Validity was examined by comparing predicted errors with observed errors. The analysis was based upon a Sensitivity Index (SI) from the signal detection paradigm reported by Baber and Stanton [17].

Similarly, the TAFEI study employed 36 undergraduates to undertake a TAFEI analysis of the confectionery vending machine interaction. Following a period of instruction, participants undertook the TAFEI analysis on three separate occasions. Again, analysis of the reliability and validity of the approach was undertaken.

As in experiment one, the frequency of hits, misses and false alarms were computed and compared with predicted error rates. In addition, the frequency of correct rejections (where errors that did not occur were correctly not predicted) was calculated by subtraction of the number of hits, misses and false alarms from a theoretical maximum. The four measures that resulted were entered into the signal detection grid shown in Table 6.



		Error observed	
		Yes	*No*
Error	*Yes*	Hit	False alarm
Predicted	*No*	Miss	Correct rejection

Table 6 Signal detection grid recording the frequency of hits, misses, false alarms and correct rejections

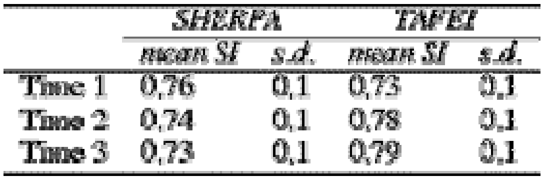
From these four measures, an index of sensitivity (S) was calculated according to the formula below [11]. This gives a value between 0 and 1 with higher values indicating greater sensitivity of error prediction.

$$\frac{\left(\frac{\text{Hit}}{\text{Hit+Miss}} + \left(1 - \frac{\text{False alarm}}{\text{False alarm+Correct rejection}}\right)\right)}{2}$$

The four frequency measures plus this index of sensitivity formed the basis of the subsequent analyses.

6.2. Results for experiment two

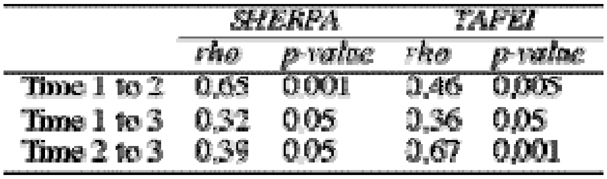
The sensitivity for the two techniques on the three occasions is reported in Table 7.



	SHERPA		*TAFEI*	
	mean SI	*s.d.*	*mean SI*	*s.d.*
Time 1	0,76	0,1	0,73	0,1
Time 2	0,74	0,1	0,78	0,1
Time 3	0,73	0,1	0,79	0,1

Table 7 Means and standard deviations for the Sensitivity Index of SHERPA and TAFEI

As Table 7 shows, there is a good deal of similarity in the sensitivity of the two approaches. This confirms previous studies [11 and 17]. The reliability of the two techniques over the three occasions is reported in Table 8.



	SHERPA		*TAFEI*	
	rho	*p-value*	*rho*	*p-value*
Time 1 to 2	0,65	0.001	0,46	0,005
Time 1 to 3	0,32	0.05	0,36	0,05
Time 2 to 3	0,39	0.05	0,67	0,001

Table 8 Reliability coefficients and probability values for SHERPA and TAFEI

The results suggest that the TAFEI approach appears to improve reliability over time (time 2 to time 3) whereas SHERPA shows quite good levels of reliability initially (time 1 to time 2).

7. Discussion and conclusions

The studies reported in this paper have sought to determine the efficacy of SHERPA and TAFEI as methods for human error identification for use in device evaluation. Both methods out-perform heuristic analysis, suggesting that there is merit to structured approaches. The two methods work in different ways, however. SHERPA is a divergent error prediction method: it works by associating up to 10 error modes with each action. In the hands of a novice, it is typical for there to be an over-inclusive strategy for selecting error modes. The novice user would rather be-safe-than-be-sorry and tend to predict many more errors than actually occur. This might be problematic; ‘crying wolf’ too many times might ruin the credibility of the approach. TAFEI, by contrast, is a convergent error prediction technique: it works by identifying the possible transitions between the different states of a device and uses the normative description of behaviour (provided by the HTA) to identify potentially erroneous actions. Even in the hands of a novice the technique seems to prevent the individual generating too many false alarms, certainly no more than they do using heuristics. In fact, by constraining the user of TAFEI to the problem space surrounding the transitions between device states, it should exclude extraneous error prediction.

A previous study on SHERPA and TAFEI reported by Baber and Stanton [17] compared predictions made by an expert user of SHERPA and TAFEI with errors reported by an observer. Baber and Stanton's study focused upon errors made during ticket purchasing on the London Underground, for which they sampled over 300 transactions during a non-continuous 24 hour period. Baber and Stanton argue that the sample was large enough as 90% of the error types were observed within 20 transactions and after 75 transactions no new error types were observed. From the study, SHERPA produced 12 of the 15 error types associated with ticket purchase, nine of which were observed to occur. TAFEI produced 10 of the 15 error types associated with ticket purchase, all of which were observed to occur. Their analysis indicated that both SHERPA and TAFEI produced acceptable level of validity when used by an expert analyst. There are, however, two main criticisms that could be aimed at this study. First, the number of participants in the study was very low; in fact only two analysts were used. Second, the analysts were experts in the use of the technique; no attempt was made to study performance whilst acquiring expertise in the use of the technique.

Whilst there are very few reports of validation studies on ergonomics methods in general [14 and 15], the few validation studies that have been conducted on HEI are quite optimistic [11, 16, 17 and 32]. It is encouraging that in recent years the number of validation studies has gradually increased. Empirical evidence of a method's worth should be one of the first requirements for acceptance of the approach by the ergonomics and human factors community. Stanton and Stevenage [11] suggest that ergonomics should adopt similar criteria to the standards set by the psychometric community, i.e. research evidence of reliability and validity before the method is widely used. It may be that the ergonomics community is largely unaware of the lack of data [33] or assumes that the methods provide their own validity [15].

Hollnagel, Kaarstad and Lee [34] argue that either we are faced with elegant theory without error prediction [5] or error prediction without any underpinning theory [32]. He calls for a bridge between theory and practice. We certainly sympathise with this call and it is central to the aims of the present paper. In analysing the Cognitive Reliability and Error Analysis Method (CREAM), Hollnagel *et al* [34] claim a 68.6% match between predicted outcomes and actual outcomes.

Despite the fact that SHERPA and TAFEI are structured techniques, there is still a good deal of reliance upon the judgement of the analyst in determining which errors are credible in any given situation. This judgement may be likened to the criterion shift in signal detection theory in determining the difference between signals (errors in this case) and noise. At least two factors play a part in the development of the sensitivity of the response operator curve: domain expertise and expertise in the human error identification method. This is an important point, because one is unlikely to yield good results without the other, as our heuristic evaluation shows. In the heuristic evaluation, the participants had a high level of domain knowledge (they were all regular users of the device under evaluation) but they had no knowledge of the human error identification technique. When domain knowledge was combined with device knowledge, a dramatic improvement in the accuracy of error identification was witnessed. As with all skills, however, improvement comes with practice [35].

The results of experiment one show that SHERPA and TAFEI provide a better means of predicting errors than an heuristic approach, and demonstrates a respectable level of concurrent validity. These findings suggest that SHERPA and TAFEI enable analysts to structure their judgement. However, the results run counter to the literature in some areas (such as usability evaluation) which suggest the superiority of heuristic approaches [20]. The views of Lansdale and Ormerod [36] may help us to reconcile these findings. They suggest that, to be applied successfully, a heuristic approach needs the support of an explicit methodology to 'ensure that the evaluation is structured and thorough' (p 257). The use of the error classification system in the heuristic group provided some structure, but a structured theory is not the same as a structured methodology. Heuristics are typically a set of 10–12 statements against which a device is evaluated. This strikes us as a poorly designed checklist, with little or no structure in the application of the method. SHERPA and TAFEI, on the other hand, provided a semi-structured methodology which formed a framework for the judgement of the analyst without constraining it. It seems to succeed precisely because of its semi-structured nature which alleviates the burden otherwise placed on the analyst's memory while allowing them room to use their own heuristic judgement. Other researchers have found that the use of structured methods can help as part of the design process, such as idea generation activities [37].

The results of experiment two show that the test–retest reliability of SHERPA and TAFEI remains fairly consistent over time. However, the correlation coefficient for test–retest reliability was moderate (by analogy to psychometric test development) and it would have been desirable to have participants achieve higher levels of reliability. Reliability and validity are interdependent concepts, but the relationship is in one direction only. Whilst it is perfectly possible to have a highly reliable technique with little validity, it is impossible for a technique to be highly valid with poor reliability. As Aitken [18] notes, 'reliability is a necessary condition but not a sufficient condition for validity' (p 93). Consequently, establishing the validity of a HEI technique is of paramount importance. The current investigation is the first reported study of people learning to use a HEI technique; all previously reported studies have been with expert users. As a result, it is conceivable that the moderate reliability values (r=0.4–0.6) obtained here may simply be an artefact of lack of experience. With this in mind, it is important to note that Baber and Stanton [17] report much higher values when users are experts.

In conclusion, the results are generally supportive of both SHERPA and TAFEI. They indicate that novices are able to acquire the approaches with relative ease and reach acceptable levels of performance within a reasonable amount of time. Comparable levels of sensitivity are achieved and both techniques look relatively stable over time. This is quite encouraging, and it shows that HEI techniques can be evaluated quantitatively. Both SHERPA and TAFEI would be respectable methods for designers to use in device design and evaluation. Any methods that enable the designer to anticipate the use of their device should be a welcome prospect. We would certainly recommend incorporating human error analysis as part of the design process, and this is certainly better practice than testing devices on the purchasers and waiting for complaints. The methodologies require something of a mind-shift in design, such that designers need to accept that 'designer-error' is the underlying cause of 'user-error'.

References

1. W.A. Wagenaar and J. Groeneweg, Accidents at sea: multiple causes, impossible consequences. *International Journal of Man-Machine Studies* 27 (1988), pp. 587–598.

2. A. Chapanis. *The Chapanis chronicles*, Aegean, Santa Barbara (1999).

3. C. Baber and N.A. Stanton, Task analysis for error identification: a methodology for designing error-tolerant consumer products. *Ergonomics* 37 11 (1994), pp. 1923–1941.

4. J.S. Busby, Error and distributed cognition in design. *Design Studies* 22 3 (2001), pp. 233–254.

5. J. Reason. *Human error*, Cambridge University Press, Cambridge (1990).

6. N.A. Stanton and C. Baber, A systems approach to human error identification. *Safety Science* 22 (1996), pp. 215–228.

7. J.W. Senders and N.P. Moray. *Human error*, LEA, Hillsdale, NJ (1991).

8. D.A. Norman. *The psychology of everyday things*, Basic Books, New York (1988).

9. C.D. Wickens. *Engineering psychology and human performance*, Harper Collins, New York (1992).

10. E. Hollnagel. *Human reliability analysis: context and control*, Academic Press, London (1993).

11. N.A. Stanton and S. Stevenage, Learning to predict human error: issues of reliability, validity and acceptability. *Ergonomics* 41 (1998), pp. 1737–1756.

12. D. Bartram, P. Lindley, J. Foster and L. Marshall. *Review of psychometric tests (level A) for assessment in vocational training*, BPS Books, Leicester (1992).

13. D. Bartram, N. Anderson, D. Kellett, P. Lindley and I. Robertson. *Review of personality assessment instruments (level B) for use in occupational settings*, BPS Books, Leicester (1995).

14. N.A. Stanton and M. Young. What price ergonomics? *Nature* 399 (1999), pp. 197–198.

15. N.A. Stanton and M. Young. *A guide to methodology in ergonomics: designing for human use*, Taylor and Francis, London (1999).

16. B. Kirwan, Human error identification in human reliability assessment. Part 2: detailed comparison of techniques. *Applied Ergonomics* 23 (1992), pp. 371–381.

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

18. L.R. Aitkin. *Psychological testing and assessment*, Allyn and Bacon, Boston (1985).

19. R.A. Virzi, Refining the test phase of usability evaluation. *Human Factors* 34 (1993), pp. 457–468.

20. J. Nielsen. *Usability engineering*, Academic Press, Boston (1993).

21. T. Landauer. *The trouble with computers*, MIT Press, Cambridge, MA (1995).

22. Nielson, J and Phillips, V L ‘Estimating the relative usability of two interfaces: heuristics, formal and empirical methods compared’, in *InterCHI'93*, ACM, New York (1993) pp 214–221.

23. Jeffries, R, Miller, J R, Wharton, C and Uyeda, K M ‘User interface evaluation in the real world: a comparison of four techniques’, in *CHI'99*, ACM, New York (1991) pp 119–124.

24. M.G. Westwater and G.I. Johnson, Comparing heuristics, user-centred and checklist-based evaluation approaches. In: S.A. Robertson, Editor, *Contemporary ergonomics 1995*, Taylor and Francis, London (1995), pp. 538–543.

25. Baber C, and Stanton, N A ‘Analytical prototyping’ in J M Noyes and M Cook (eds) *Interface Technology: the leading edge*. Research Studies Press, Baldock (1999) pp 175–194.

26. Embrey, D E ‘SHERPA: a systematic human error reduction and prediction approach’, *Paper presented at the International Meeting on Advances in Nuclear Power Systems*, Knoxville, Tennessee (1986).

27. J. Annett, K.D. Duncan, R.B. Stammers and M.J. Gray. *Task analysis. Training information No. 6*, HMSO, London (1971).

28. N.A. Stanton and C.A. Baber, A systems analysis of consumer products. In: N.A. Stanton, Editor, *Human factors in consumer products*, Taylor and Francis, London (1998), pp. 75–90.

29. Angel, E S and Bekey, G A ‘Adaptive finite state models of manual control systems’, *IEEE Transactions on Man-Machine Systems*, March (1968) pp 15–29.

30. H. Thimbleby, Can humans think? *Ergonomics* 34 (1991), pp. 1269–1287.

31. Nielson, J and Mollich, R ‘Heuristic evaluation of user interfaces’, *CHI'90* ACM, New York (1990) pp 249-256.

32. B. Kirwan. *A practical guide to human reliability assessment*, Taylor and Francis, London (1994).

33. N.A. Stanton and M. Young, Is utility in the mind of the beholder? A study of ergonomics methods. *Applied Ergonomics* 29 (1998), pp. 41–54.

34. E. Hollnagel, M. Kaarstad and H.-C. Lee, Error mode prediction. *Ergonomics* 42 (1998), pp. 1457–1471.

35. A.T. Welford. *Fundamentals of skill*, Methuen, London (1971).

36. M.W. Lansdale and T.C. Ormerod. *Understanding interfaces*, Academic Press, London (1994).

37. Jones, E, Stanton, N A and Harrison, D ‘Applying structured methods to eco-innovation. An evaluation of the product ideas tree diagram’ *Design Studies* Vol 22(6) (2001) 519–542.