CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Brunel University Research Archive

Validating TAFEI:

reliability and validity of a human error prediction technique

Neville A. Stanton* and Christopher Baber

School of Engineering and Design
Brunel University
Uxbridge
Middlesex
UB8 3PH, UK

School of Electronic and Electrical
Engineering
University of Birmingham
Edgbaston, Birmingham
B15 2TT, UK

email: Neville.Stanton@brunel.ac.uk
phone: +44 (0) 1895 265543
fax :

* Correspondence should be directed to Professor Neville Stanton

Validating TAFEI:

reliability and validity of a human error prediction technique

Abstract

This paper reports on the theoretical and empirical developments for an error prediction methodology called Task analysis For Error Identification (TAFEI). Other researchers have noted the need for theoretically-driven approaches that are able to provide practical utility in error prediction. Theoretical developments include the concept of 'rewritable routines', that describes the loop between cognitive processing, action and devices states. This has been proposed as a way of unifying ideas from systems theory and cognitive psychology. The empirical research shows that TAFEI is superior to heuristic methods, which supports the idea that structured methods assist in error prediction. The validation study shows that TAFEI reaches acceptable levels in terms of test-retest reliability and concurrent validity. It is believed that the method has reached a level of maturity after ten years of development work. This is demonstrated by the many uses that the method has been put, including that of a design tool.

KEYWORDS: human error, methodology, reliability, validity, sensitivity.

1. FOUNDATIONS OF TAFEI

In 1994, we introduced a technique for predicting human error when people use devices (Baber and Stanton 1994). The technique was called Task Analysis For Error Identification (TAFEI). The theoretical foundation for this approach was based upon the unification of ideas from general systems theory (von Bertalanffy 1950) and human problem solving (Miller et al. 1960, Hayes-Roth and Hayes-Roth 1979, Suchman 1987, Simon and Young 1988) resulting in a state-space search with boundaries between human and device elements. Initial attempts at validating the approach suggested that the errors predicted using TAFEI were relatively consistent with errors observed when people used the devices under consideration (Baber and Stanton 1996).

During the past decade, our efforts have been directed towards consolidating the theory that underlies TAFEI. At the heart of this theory are the ideas of 'relevance' and 'rewritable routines': we posit that in order to move from current to relevant states, eliminating other possible states, the user needs to retain some (temporary) record of the interaction and to have some means of assigning relevance to states (Baber and Stanton 1998). At each state, this record will be modified. We assume that the record will be held in working memory, presumably in the articulatory loop which has a limited duration (e.g., around two seconds, Baddeley 1986). This means that unless the record is updated, it will decay or be disrupted. This could result from distraction, e.g., interrupting the activity, or competition, e.g., having more than one option that appears relevant. As the record will also guide the next action, we see this as a rewritable routine. The current routine will be performed and the results used to update the record, in preparation for the next routine. To some extent this notion is similar to the ‘partial provisional planning’ hypothesis of Simon and Young (1988) and also resonates with the idea of a

'scratchpad' held in working memory (although the 'scratchpad' has normally been associated with visuo-spatial processing, Baddeley 1986). Figure 1 presents a simple schematic of this process. The possible states (interpreted by the user from the machine) are compared against states which could lead to the goal. The comparator has a two-way connection to the rewritable routines (with the routines influencing the comparator, i.e., by defining relevance, and taking the output to define action).

INSERT FIGURE 1 ABOUT HERE

Figure 1: Simple Schematic of Rewritable Routines

We feel that the role of the comparator could draw upon the psychological theory of the perceptual cycle put forward by Niesser (1976), as it should direct activity, sample the environment and modify stored information. The comparator is the mechanism that engages the environment and is therefore the interactive part of the process. In this model, the 'environment' will largely be confined to the 'System Image' presented by the device. By System Image, we are following the lead of Norman (1988) who proposed that the physical appearance of a product (together with feedback received through other senses, such as hearing or touch) constituted the Image of the product, and that the user will use this Image to guide activity. Thus, the device might contain a small Liquid Crystal Display (LCD) and three buttons; the user might then assume that pressing any of the three buttons might lead to changes on the LCD.

The theory draws upon the ideas of scripts and schema. We can imagine that a person approaching a ticket-vending machine might draw upon a 'vending machine' or 'ticket kiosk'

script when using a ticket machine. From one script, the user might expect the first action to be 'Insert Money', but from the other script, the user might expect the first action to be 'Select Item'. The success, or failure, of the interaction would depend on how closely the user was able to determine a match between the script and the actual operation of the machine. The role of the comparator is vital in this interaction. If it detects differences from the expected states, then it is able to modify the routines. Failure to detect any differences is likely to result in errors. Following Bartlett's (1932) lead, the notion of schema is assumed to reflect a person's "…effort after meaning." [Bartlett 1932: 20], arising from the active processing (by the person) of a given stimulus. This active processing involves combining prior knowledge with information contained in the stimulus. While schema theory is not without its critics (see Brewer 2000) for a review, the notion of an active processing of stimuli clearly has resonance with our proposal for rewritable routines. It might be assumed that there are similarities between the notion of rewritable routines and some of the research on mental models that was popular in the 1980s (see Rogers et al. 1992 for a review). However, we would like to point out the following differences between rewritable routines and mental models:

i. comparison need not involve rule-based or knowledge-based reasoning, i.e., it could consist of skill-based activity. An example of this would direction of motion stereotypes (see Sanders and McCormick 1992 for a review);

ii. comparison requires the interaction between the user's schema and the System Image, i.e., the current state of the device;

iii. Any model developed will only be needed for the current state, which means that

users might recruit a whole host of models during their interaction.

The theory is potentially useful in addressing the interaction between sub-components in systems (i.e., the human and the device). It also assumes a hierarchical order of system components, i.e., all structures and functions are ordered by their relation to other structures and functions, and any particular object or event is comprised of lesser objects and events. General systems theory describes a system in dynamic terms:

- activity results from continual adaptation of the system components
- changes in one component affect other components and the whole system
- systems components become linked by exchanges (inputs and outputs)
- within a component an internal conversion occurs
- each type of input has a corresponding output
- errors become apparent at boundaries between components

The input~conversion~output cycle of a human-machine system is of particular interest here, as are the boundaries between humans and machines, as this is where errors become apparent. In other words, errors arise from transitions between states and can only be identified when a transition has failed to lead to the expected state. We believe that it is essential for a method of error prediction to examine explicitly the nature of the interaction. Many methods appear to do this in an implicit way, but leave consideration of the interaction to the judgment of the analyst. Whilst expert analysts, with a good understanding of the domain, may be able to do this task satisfactorily on most occasions, it does not make the analysis as objective as it could be (Stanton and Baber 1996). TAFEI explicitly analyses the *interaction* between people and

machines and is concerned with task-based scenarios. This is done by mapping human activity onto machine states.

2. HOW TAFEI WORKS

TAFEI attempts to predict errors with device use by modelling the interaction between user and device. It assumes that people use devices in a purposeful manner, such that the interaction may be described as a "cooperative endeavour", and it is by this process that problems arise. Furthermore, the technique makes the assumption that actions are constrained by the state of the product at any particular point in the interaction, and that the device offers information to the user about its functionality. Thus, the interaction between users and devices progresses through a sequence of states. At each state, the user selects the action most relevant to their goal, based on the System Image.

Procedurally, TAFEI is comprised of three main stages – see figure 2. Firstly, an Hierarchical Task Analysis (HTA) is performed to model the human side of the interaction. It is, of course, possible to employ any technique to describe human activity. However, HTA suits our purposes for the following reasons: i. it is related to Goals and Tasks; ii. it is directed at a specific goal; iii. it allows consideration of task sequences (through 'plans'). As will become apparent, TAFEI focuses on a sequence of tasks aimed at reaching a specific goal. Next, State-Space Diagrams (SSDs) are constructed to represent the behaviour of the artifact. Plans from the HTA are mapped onto the SSD to form the TAFEI diagram. Finally, a transition matrix is devised to display state transitions during device use. TAFEI aims to assist the design of artifacts by illustrating when a state transition is possible but undesirable (i.e., illegal). Making all illegal

transitions impossible should facilitate the cooperative endeavour of device use.

INSERT FIGURE 2 ABOUT HERE

Figure : 2 The series of decision stages involved in the TAFEI technique.

The first step in a TAFEI analysis is to obtain an appropriate HTA for the device, as shown in figure 3. As TAFEI is best applied to scenario analyses, it is wise to consider just one specific goal, as described by the HTA (e.g., a specific, closed-loop task of interest) rather than the whole design. Once this goal has been selected, the analysis proceeds to constructing State-Space Diagrams (SSDs) for device operation.

INSERT FIGURE 3 ABOUT HERE

Figure 3: Hierarchical Task Analysis.

A SSD essentially consists of a series of states that the device passes from a starting state to the goal state. For each series of states, there will be a current state, and a set of possible exits to other states. At a basic level, the current state might be “off”, with the exit condition “switch on” taking the device to the state “on”. Thus, when the device is “off” it is ‘waiting for…’ an action (or set of actions) that will take it to the state “on”. It is very important to have, on completing the SSD, an exhaustive set of states for the device under analysis. Numbered plans from the HTA are then mapped onto the SSD, indicating which human actions take the device from one state to another. Thus the plans are mapped onto the state transitions (if a transition is

activated by the machine, this is also indicated on the SSD, using the letter 'M' on the TAFEI diagram). This results in a TAFEI diagram, as shown in figure 4.

INSERT FIGURE 4 ABOUT HERE

Figure 4: State-space TAFEI diagram

The most important part of the analysis from the point of view of improving usability is the transition matrix. All possible states are entered as headers on a matrix – see figure 5. The cells represent state transitions (e.g., the cell at row 1, column 2 represents the transition between state 1 and state 2), and are then filled in one of three ways. If a transition is deemed impossible (i.e., it is simply not possible to go from one state to another state), a "-" is entered into the cell. If a transition is deemed possible and desirable (i.e., it progresses the user towards the goal state - a correct action), this is a legal transition and "L" is entered into the cell. If, however, a transition is both possible but undesirable (a deviation from the intended path - an error), this is termed illegal and the cell is filled with an "I". The idea behind TAFEI is that usability may be improved by making all illegal transitions (errors) impossible, thereby limiting the user to only performing desirable actions.

INSERT FIGURE 5 ABOUT HERE

Figure 5: Transition matrix

Examples of applications of TAFEI include prediction of errors in boiling kettles (Baber and

Stanton 1994, Stanton and Baber 1998), comparison of word processing packages (Stanton and Baber 1996, Baber and Stanton 1999), withdrawing cash from automatic teller machines (Burford, 1993), medical applications (Baber and Stanton 1999, Yamaoka and Baber 2000), recording on tape-to-tape machines (Baber and Stanton 1994), programming a menu on cookers (Crawford et al 2001), programming video-cassette recorders (Baber and Stanton 1994, Stanton and Baber 1998), operating radio-cassette machines (Stanton and Young 1999b), recalling a phone number on mobile phones (Baber and Stanton 2001), buying a rail ticket on the ticket machines on the London Underground (Baber and Stanton 1996), and operating high-voltage switchgear in substations (Glendon and McKenna 1995).

All of these examples of applying of TAFEI share common features, which define the operational parameters of the technique. First, they are all applied to the analysis of a scenario of device use. The technique assumes purposeful use of the device, drawn from the goals of a tasks analysis. Second, each of the devices offers a clear and logical sequence of activity. The tasks are discrete, step-by-step, rather than continuous and concurrent. Third, there are clear system boundaries between the device and human elements. Given these requirements, it is no wonder that most of the analyses have tended toward single user, single device systems. Theoretically it should be possible to take a nested systems approach, to analyse systems of greater complexity by addressing different levels and different scenarios. Some movement in this direction has begun with the analysis of operating high-voltage switchgear in substations (Glendon and McKenna 1995). Before analyzing complex systems (such as the case of the substation case study) any further, it would be ideal to assess the performance of the TAFEI technique on relatively simple systems first.

3. VALIDATION OF HUMAN ERROR IDENTIFICATION TECHNIQUES

Whilst there are very few reports of validation studies on ergonomics methods in general (Stanton and Young 1999a), the few validation studies that have been conducted on Human Error Identification (HEI) are quite optimistic (e.g. Kirwan 1992a, b, Baber and Stanton 1996). It is encouraging that in recent years the number of validation studies has gradually increased. Empirical evidence of a methods worth should be one of the first requirements for acceptance of the approach by the ergonomics and human factors community. Stanton and Stevenage (1998) 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 (Stanton and Young 1998) or assumes that the methods provide their own validity (Stanton and Young 1999b).

Hollnagel et al. (1999) argue that either we are faced with elegant theory without error prediction (e.g. Reason 1990) or error prediction without any underpinning theory (e.g. Kirwan 1994). Hollnagel et al. (1999) call 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. (1999) claim a 68.6% match between predicted outcomes and actual outcomes.

Stanton and Stevenage (1998) raise several methodological concerns with some of the approaches used in previous validation studies. These concerns comprise: the number of assessors using each technique is typically very small (e.g. between 1 and 3 assessors) and the

use of subjective rating scales rather than some objective measure of performance (i.e. the comparison of predicted errors with actual errors). In an earlier study, Baber and Stanton (1996) aimed to provide a more rigorous test of the predictive validity of TAFEI. Predictive validity was tested by comparing the errors identified by an expert analyst with those observed during 300 transactions with a ticket machine on the London Underground. Baber and Stanton (1996) suggest that TAFEI provides an acceptable level of sensitivity based on the data from two expert analysts ($r = 0.8$). Stanton and Stevenage (1998) developed this approach and proposed a more formal method of benchmarking the performance of HEI techniques. This approach has been followed in the current study.

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 five and eight assessors can uncover up to 80% of usability problems (Virzi 1992, Nielsen 1993, Landauer 1995). The idea of heuristics represents an interesting benchmark for this work, in that such evaluation is proposed to be 'quick and dirty', but to yield useful results. If it can be shown that HEI techniques are as quick and as reliable, then their use in product evaluation could be considered seriously. 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. 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 experiment 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.

4. EXPERIMENT ONE

4.1. Method for Experiment One

4.1.1. Participants

Two 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 group consisted of 36 participants drawn to match the pool as above. These participants acted as novice analysts using Task Analysis For Error Identification (TAFEI). All participants were equally familiar with the machine upon which the human error analysis was conducted.

4.1.2. Materials

No materials were provided for the control participants, although their training allowed them to

develop heuristics for the task (see below). However, all TAFEI analysts were provided with three items. First, they received a hierarchical task analysis (HTA) chart describing the action stages involved when using a vending machine to obtain a bar of chocolate (see Figure 3). Second, they received a state-space diagram of machine states (see Figure 4). Finally, participants were provided with a proforma for recording their error predictions. The external provision of the HTA and SSD can be justified by keeping the methodology as closely as possible to that developed by Stanton and Stevenage (1998).

4.1.3. 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 (Nielsen 1993). 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 confectionary vending machine was chosen. This was familiar to all participants and had been used by all of them.

4.1.4. Procedure

For both groups, participants were given the scenario of buying one item (a Lion Bar, costing 24p) from the vending machine using a 50p 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 (1990). 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 (Wickens 1992) and examples of each error type were discussed in various contexts. In particular, the link was made between product design and human error (Norman 1988, Thimbleby 1991). 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) constituted a set of heuristics that participants in the Heuristic condition could apply in their evaluation. Finally, participants using the TAFEI technique 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 of the TAFEI 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).

4.1.5. Error prediction

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

4.1.6. Error classification:

The error predictions from all participants were compared to the errors actually observed in 75 independent transactions with the machine. Observation of the 75 transactions revealed 11 discrete types of error and it was possible for more than one error type to occur within a single transaction. These error types are listed in Appendix 1 and errors were identified by a combination of observation and interview with the users of the machine. 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 (1996), 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:

(a) hits (predicted errors that were seen to occur)

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

(c) 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.

4.2. 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 1 summarises these data across the control group and the group using the TAFEI method for human error identification.

INSERT TABLE 1 ABOUT HERE

From Table 1, the participants using the TAFEI technique correctly predicted more errors and

missed fewer errors than the heuristic group. The TAFEI group predict a mean of 4.3 of the 11 errors (39%), while the heuristic group predicted 2.8 of the 11 errors (25%). Three independent samples t-tests examined whether these differences were significant. The results revealed a significant difference in hit rate ($Z_{corrected} = -4.3972$, $df = 70$, $p < 0.001$), and also in miss rate, ($Z_{corrected} = -4.3972$, $df = 70$, $p < 0.001$) in favour of participants using the TAFEI method. The results also showed that there were no statistical differences in false alarms ($Z_{corrected} = 0.1968$, $df = 70$, $p = NS$).

4.3. Discussion for experiment one

These results suggest that when participants first use the TAFEI method 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 TAFEI seems to be better than an heuristic approach to error prediction. These gains associated with using TAFEI do not appear to be at the expense of generating significantly more false alarms. 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 (Nielsen and Mollich, 1990). The study reported by Stanton and Stevenage (1998), using SHERPA, resulted in significantly more false alarms (mean false alarm rate of 15.4 $\pm$ 6.1 vs. 1.5 $\pm$ 1.3 from this study). The relatively low rate of false alarms generated by using TAFEI in this study may be due the inherent differences in the way the two methods work. 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 play-safe-than-be-sorry and they 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. Indeed, this was one of the original

aims for the technique when it was originally developed (Baber and Stanton 1994).

5. EXPERIMENT TWO

While Experiment One has suggested that TAFEI out-performs an heuristic technique, the 'hit' rate of 39% was quite low. Having said this, any technique that identifies errors with sufficient reliability (i.e., few false alarms and no unpredicted errors) could prove beneficial in the early analysis of product designs. In the next experiments, we wanted to see how practice might improve performance.

5.1. Method for experiment two

5.1.1. Participants

The 36 participants who used the TAFEI method in the previous study, were also employed in this study.

5.1.2. Materials

As for the TAFEI group in Experiment One (see Figures 1, 2 and 3).

5.1.3. Error prediction

The procedure matched that of Experiment One as closely as possible. However, there were several important differences. First, there were no untrained control participants. Instead, participants acted as controls for themselves in that the sensitivity of their error prediction could

be compared within participants over time. A second important methodological difference between the first study and the present ones arises from the fact that across the three occasions, a repeated measures design was used. A one week gap was allowed between the time that data was collected from participants. Consequently, all participants had practice using their error detection method and had immediate feedback in the form of access to the observational data. This was considered to be an important factor given our desire to mirror the procedures used in recent training regimes (see Patrick, 1992). Furthermore, whilst being familiar with the methodological concerns of common method variance, a tradition established in the psychometric arena is to hold as many factors of the testing situation constant as possible. The introduction of different machines would cloud the issue because it would not be possible to determine whether differences in sensitivity of error prediction over time were due to learning, or due to the ease of predicting errors with the various machines. For these reasons, participants made predictions of errors when using one particular machine and this allows us to determine the learnability of the error prediction technique without the confusion of error prediction on different machines. Apart from these conditions of testing, all other methodological details remained unchanged.

5.1.4. Error classification

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 (number of cells in transition matrix as generated by the SSD). The four measures that resulted were entered into the signal detection grid below.

INSERT FIGURE 6 ABOUT HERE

Figure 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 (from Stanton and Stevenage 1998). 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.

5.2. Results for experiment two

For the purposes of the following section the data from all of the experiments are pooled in order to examine the effects of time on the validity and reliability of the TAFEI technique. At each of the three times, the frequency of hits, misses, false alarms and correct rejections were recorded and from these an index of sensitivity was calculated. These are summarised in table 2.

INSERT TABLE 2 ABOUT HERE

As shown in table 2, there is a statistically significant increase in the number of hits (χ^2_2 = 12.2639, $p<0.005$) and consequently a statistically significant decrease in the number of misses (χ^2_2 = 12.2639, $p<0.005$). By conducting comparisons between the hits over time we find that hits increase from time one to time two ($Z=-3.1322$, $p<0.005$) and from time one to time three ($Z=-3.3474$, $p<0.001$), but are stable between time two and time three ($Z=-0.365$, p=NS). This effect is mirrored in the misses, which shows a decrease from time one to time two ($Z=-3.1322$, $p<0.005$) and from time one to time three ($Z=-3.3474$, $p<0.001$), but are stable between time two and time three ($Z=-0.365$, p=NS).

There are no statistical differences in false alarms (χ^2_2 = 3.2917, p=NS) or correct rejections (χ^2_2 = 3.2917, p=NS).

There are statistically significant differences in sensitivity over time ($\chi^2_{2\ 2}$ = 7.722, $p<0.05$). As with the hits, there is an increase in sensitivity from time one to time two ($Z=3.2312$, $p<0.005$) and from time one to time three ($Z=-2.9833$, $p<0.005$), but sensitivity is stable between time two and time three ($Z=-0.624$, p=NS).

In order to compute reliability of the TAFEI method, correlations of sensitivity were undertaken. These show moderate, but statistically significant, correlations of sensitivity between time one and two ($r = 0.46$, $p< 0.01$) and between time one and three ($r = 0.36$, $p< 0.05$). There is a much

higher correlation of sensitivity between time two and three ($r = 0.67$, $p < 0.001$).

5.3. Discussion for experiment two

The data presented in table 2 suggest an early plateau of performance (at time two) for the TAFEI participants, albeit for a relatively simple task. The level of reliability achieved between time two and time three is perfectly acceptable. The study shows that some level of practice needed to improve performance, which is to be expected. By time three, the participants are obtaining the levels of performance on the sensitivity index that Baber and Stanton (1996) report of expert analysts, although the reliability coefficients are somewhat lower.

TAFEI compares favourably with the reliability (ranging between 0.32 and 0.65) and validity (ranging between 0.73 and 0.76) of SHERPA, as reported by Stanton and Stevenage (1998). Both seem to offer acceptable, and to some extent comparable, levels of performance. The studies of SHERPA and TAFEI provide a baseline from which other techniques could be compared using the standardised HTA and error data.

Where TAFEI performs rather better than SHERPA is over the number of false alarms generated. Whilst participants in the study reported by Stanton and Stevenage (1998) were generating a false alarm rate of around 20%, participants in this study were closer to 3%. We believe that is an artifact of the way in which the two methods work in the hands of novices. It is much easier to generate false alarms with SHERPA than with TAFEI (Stanton and Baber 2002).

GENERAL DISCUSSION

This paper has presented a theory-based human error identification technique. Since its initial development over ten years ago (Baber and Stanton 1994), it has developed some degree of maturity, in terms of the theoretical development and the applications to which it has been put. Both the applied and academic study have been mutually beneficial. Specifically, the theory of rewritable routines has supported methodological development and vice versa (Baber and Stanton, 1997). Parallel developments have been testing the theory in more detail to determine how people employ rewritable routines in their interaction with devices (Baber and Stanton 2001, Stanton and Baber 2002) which relates to some of the contemporary developments in human-computer interaction research (Diaper and Stanton 2004). Developments of rewritable routines theory include attempts to understand the relationship between user experience, system image, human activity and devices states. It is proposed that there is a cyclical relationship between these system interaction elements, whereby the user draws upon prior experience to interpret the system image, which then drivers human action, which determine the device state and the new system image. Then the cycle begins again, where the user draws upon previous experience to interpret the new system image. Some devices (or some states of some devices) may have more powerful system images, which draw on analogies, metaphors, syllogisms, semantics, affordances, stereotypes and cues that are meaningful to the user. If the actions allowed by the device are compatible with these inferred meanings, then interaction is likely to be successful and vice versa. Current research is looking at the role of system image in the recruitment of appropriate rewritable routines, and the relationship between global (an aspect of the routine that is pertinent to the whole episode of interaction) and local routines (an aspect of the routine that is only pertinent to a sub-episode of interaction).

An initial attempt to present validation data early on with expert users (Baber and Stanton 1996) has led to the study of addressing the performance of TAFEI in the hands of novice users. In comparison to error prediction using a heuristic approach, TAFEI is clearly superior. A major claim for heuristics approaches is their speed of use. Experiment One suggests that, given equal, time an heuristic approach is less productive than the structured HEI approach of TAFEI. HEI asks analysts to consider human activity (rather than device characteristics). This means that we are asking our assessors to consider how potential users might experience problems with the device (rather than focusing on device specifics). We feel that this is of benefit for two reasons: i. It forces the analyst to consider human activity, and to consider problems in the light of this activity; ii. it forces the focus of attention away from device features. This means that it is possible to suggest radical revisions to a design (in order to reduce predicted errors), rather than seeking to modify specific features. Finally, HEI techniques do not need real users performing real tasks with real products; thus, the techniques are applicable to very early stages of design (see Baber and Stanton 1999 for a discussion of how this approach can be applied as a design tool and Stanton and Young 1999b for examples how this could form part of an overall approach to analytical prototyping in product development).

Participants also showed performance improvement over the first two experiments, increasing the hits and reducing the misses without compromising the false alarms. Both reliability and validity achieved acceptable levels, and the data compare well with previous studies (Stanton and Stevenage 1998). The signal detection paradigm is a useful way of coding error prediction data, and the present study reinforces this approach. One might argue that a 'hit' rate of 5.4 out of 11 (49%) is relatively poor, but as argued above, any technique that can predict human error

reliably can prove useful in product design and evaluation. Furthermore, Baber and Stanton (1996) have shown that expert performance with TAFEI reaches 90%.

There are a number of other criticisms that need to be addressed, however. Stanton and Stevenage (1998) propose that clearer documentation on the methodologies needs to be provided, and that cross validation studies should be undertaken. Certainly both of these issues have been addressed to some extent with respect to TAFEI. A recent handbook of human factors and ergonomics methods shows how the reporting of methods in the literature could be rather more structured (Stanton et al. 2005). The other validation study presents data from skilled analyst with a more complex device (Baber and Stanton 1996). Further studies with different levels of skill and different complexity of devices surely need to be undertaken in due course. One such study by Stanton and Young (1999b), suggests that novice analysts have some difficult assessing complex devices within a limited amount of time. The factors of analyst's expertise and device complexity are likely to interact (Stanton and Young 2003). With the growing number of studies using TAFEI, as outlined in the introduction, the documentation on the method is growing and freely available, which overcomes another of the criticisms of such methods.

To conclude, the reliability and validity of TAFEI for novices applying the technique to a relatively simple system looks encouraging. The next question is, do the results generalise and scale up? We suspect that, as with most ergonomics methods, there is an element of craft-skill associated with successful practice. Whilst experts can analyse complex systems and novices can analyse simple systems, we doubt that a novice will be able to analyse a complex system successfully. This seems self-evident, as there is both domain knowledge and mastery of the

technique to be possessed. Whilst it has been popular in the usability literature to give the impression that most system problems can be detected quite quickly with unstructured techniques, we feel that this is misguided. The use of structured techniques such as SHERPA and TAFEI have clear benefits in terms of error prediction but need to be used in combination with other methods if broader aspects of performance are to be addressed.

REFERENCES

Baber, C. and Stanton, N.A. (1994) Task analysis for error identification, *Ergonomics, 37,* 1923-1942.

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

Baber, C., and Stanton, N.A. (1998) Rewritable routines in human interaction with public technology, *International Journal of Cognitive Ergonomics, 1,* 337-349.

Baber, C. and Stanton, N.A. (1999) Analytical prototyping, In J.M. Noyes and M. Cook (eds) *Interface Technology: The Leading Edge,* (Baldock: Research Studies Press), 175-194.

Baber, C. and Stanton, N. (2001) Analytical prototyping of personal technologies: using predictions of time and error to evaluate user interfaces, In M. Hirose (ed) *Interact'01,* (Amsterdam: IOS Press), 585-592.

Baddeley, A.D. (1986) *Working Memory,* (Oxford: Oxford University Press).

Bartlett, F.C. (1932) *Remembering: A Study in Experimental and Social Psychology*, (Cambridge: Cambridge University Press).

Brewer, W.F. (2000) Bartlett's concept of the schema and its impact on theories of knowledge representation in contemporary cognitive psychology, In A. Saito (ed) *Bartlett, Culture and*

Cognition, (London: Psychology Press), 69-89

Burford, B. (1993) *Designing Adaptive ATMs*, Birmingham: University of Birmingham unpublished MSc Thesis.

Crawford, J. O., Taylor, C. and Li Wan Po, N. (2001) A Case Study of On-Screen Prototypes and Usability Evaluation of Electronic Timers and Food Menu Systems, *International Journal of Human-Computer Interaction*, 13, 187-201.

Diaper, D. & Stanton, N. A. (2004) *Handbook of Task Analysis in Human-Computer Interaction.* (Mahwah, New Jersey: Lawrence Erlbaum Associates).

Glendon, A.I. and McKenna, E.F. (1995) *Human Safety and Risk Management*, (London: Chapman and Hall).

Hayes-Roth, B. and Hayes-Roth, F. (1979) A cognitive model of planning, *Cognitive Science*, *3*, 275-310.

Hollnagel, E., Kaarstad, M. and Lee, H-C. (1999) Error mode prediction. *Ergonomics*, 42, 1457-1471.

Kirwan, B. (1992a) Human error identification in human reliability assessment. Part 1:overview of approaches. *Applied Ergonomics, 23*, 299-318.

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

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

Landauer, T. (1995) *The Trouble with Computers*, Cambridge, (MA: MIT Press).

Miller, G.A., Galanter, E. and Pribram, K.H. (1960) *Plans and the Structure of Behaviour* (New York: Holt, Rinehart and Winston).

Nielsen, J. (1993) *Usability Engineering*, (Boston: Academic Press).

Neilson, J. and Mollich, R. (1990). Heuristic evaluation of user interfaces, *Proceedings of CHI'90,* (New York: ACM), 249-256.

Niesser, U. (1976) *Cognition and Reality: Principles and Implications of Cognitive Psychology*. (San Fancisco: Freeman).

Norman, D. A. (1988). *The Psychology of Everyday Things.* (New York: Basic Books).

Patrick, J. (1992) *Training: Research and Practice*. (London: Academic Press).

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

Rogers, Y, Rutherford, A. and Bibby, P. A. (1992) *Models in the Mind: Theory, Perspective and Application*. (London: Academic Press).

Senders, M. S and McCormick, E. J. (1993) *Human Factors in Engineering and Design*. (New York: McGraw-Hill).

Simon, T. and Young, R.M. (1988) GOMS meets STRIPS: the integration of planning with skilled procedure execution in human-computer interaction In: D.M. Jones and R. Winder (eds.) *People and Computer IV,* (Cambridge: Cambridge University Press), 581-594.

Stanton, N.A. and Baber, C. (1996) A systems approach to human error, *Safety Science, 22,* 215-218.

Stanton and Baber (1998) A systems analysis of consumer products. In: N. A. Stanton (ed.) *Human Factors in Consumer Products*. (London: Taylor & Francis), 75-90.

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

Stanton, N. A., Hedge, A., Salas, E., Hendrick, H. & Brookhaus, K. (2005) *Handbook of Human Factors and Ergonomics Methods*. (London: Taylor & Francis).

Stanton, N. A. and Stevenage, S. (1998) Learning to predict human error: issues of reliability,

validity and acceptability. *Ergonomics, 41*, 1737-1756.

Stanton, N. A. & Young, M. (1998) Is utility in the mind of the beholder? A review of ergonomics methods. *Applied Ergonomics.* 29, 41-54

Stanton, N. A. & Young, M. (1999, a) What price ergonomics? *Nature* 399, 197-198

Stanton, N. A. & Young, M. (1999, b) *A Guide to Methodology in Ergonomics: Designing for Human Use.* (London: Taylor & Francis).

Stanton, N. A. & Young, M. (2003) Giving ergonomics away? The application of ergonomics methods by novices. Applied Ergonomics 34, 479-490.

Suchman, L. (1987) Plans and Situated Action. (Cambridge: Cambridge University Press).

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

von Bertalanffy, L. (1950) The theory of open systems in physics and biology. *Science,* 111, 23-29.

Virzi, R.A. (1992) Refining the test phase of usability evaluation, *Human Factors, 34*, 457-468

Wickens, C. D. (1992) *Engineering Psychology and Human Performance*. (New York: Harper Collins).

Yamaoka, T. and Baber, C. (2000) Three point task analysis and human error estimation, *Proceedings of the Human Interface Symposium 2000*, Tokyo, Japan, 395-398.

Table 1: The mean frequency of hits, misses and false alarms for TAFEI and untrained participants.

	UNTRAINED PARTICIPANTS		TAFEI PARTICIPANTS	
No. Participants	36		36	
	Mean	sd	Mean	sd
hits	2.8	1.2	4.3	1.5
misses 6.2	0.9		4.7	1.5
false alarms	1.8	0.9	1.5	1.3

Table 2: The mean frequency of hits, misses, false alarms and correct rejections and the index of sensitivity of SHERPA over time .

	TIME ONE		TIME TWO		TIME THREE	
participants	36		36		36	
	Mean	sd	Mean	sd	Mean	sd
hits	4.3	1.5	5.4	1.8	5.4	1.5
misses	4.7	1.5	3.6	1.8	3.6	1.5
false alarms	1.5	1.3	1.9	1.3	2.2	1.7
correct rejs	67.5	1.3	67.1	1.3	66.8	1.7
sensitivity	0.73	0.1	0.78	0.1	0.79	0.1

Figure 1: Simple Schematic of Rewritable Routines

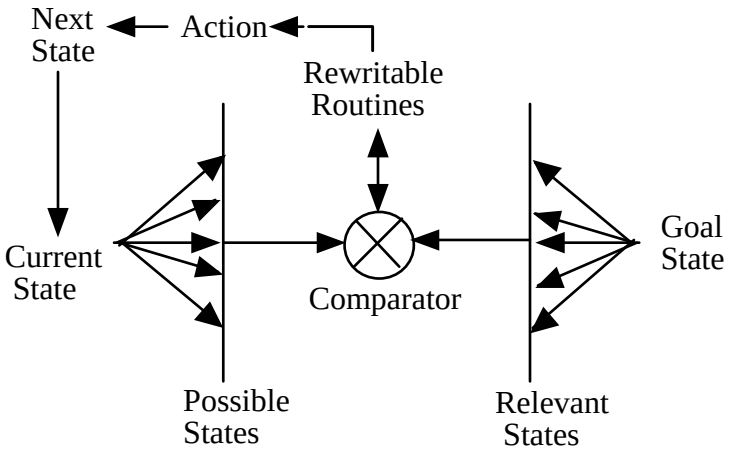

Figure 2: State-space diagram

Figure 3: The series of decision stages involved in the TAFEI technique.

Figure 4: Transition matrix

Figure 6: Signal Detection Grid recording the frequency of hits, misses, false alarms and correct rejections.

		Errors Observed	
		YES	NO
Errors Predicted	YES	hits	false alarms
	NO	misses	correct rejections

Appendix One

Frequency of prediction of each error type using TAFEI

(based on responses of a set of 36 participants taking part at all three testing phases.)

Error Type	Freq at time 1	Freq at time 2	Freq at time 3
Put wrong coins in*	13	11	14
Leave after inserting coins	11	14	19
Fail to put coins in*	16	21	18
Not enough money in*	8	13	12
Pressed wrong character*	27	28	30
Fail to press character*	19	26	18
Pressed wrong number*	25	27	28
Fail to press number	9	10	2
Push flap too early*	1	11	14
Fail to turn handle at all*	19	12	15
Fail to pick up change*	29	30	29
Fail to pick up item	13	16	15

*indicates error observed