5+ model YIJHC : 1418 Prod.Type:FTP pp.1–16(col.fig.:3–5) ED:BibhuPrasad PAGN:VG SCAN:

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Int. J. Human-Computer Studies ■ (■■■■) ■■■–■■■

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

Cue effectiveness in mitigating postcompletion errors in a routine procedural task

Phillip H. Chung, Michael D. Byrne*

Rice University, 6100 Main Street, MS-25, Houston, TX 77005-1892, USA

Received 8 July 2005; received in revised form 20 August 2007; accepted 6 September 2007

Communicated by S. Wiedenbeck

Abstract

Postcompletion errors, which are omissions of actions required after the completion of a task's main goal, occur in a variety of everyday procedural tasks. Previous research has demonstrated the difficulty of reducing their frequency by means other than redesigning the task structure [Byrne, M.D., Davis, E.M., 2006. Task structure and postcompletion error in the execution of a routine procedure. Human Factors 48, 627–638]. Nevertheless, finding a successful strategy for mitigation of this type of error may uncover important mechanisms underlying interactive behavior. Two experiments were carried out to test visual cues for their ability to reduce the frequency of postcompletion errors in a computer-based routine procedural task. A cue that was visually salient, just-in-time, and meaningful entirely eliminated the error, whereas cues that were not as specific were ineffective. These results are beyond the predictive capability of extant error identification methods and common design guidelines but are consistent with the work of Altmann and Trafton [2002. Memory for goals: an activation-based model. Cognitive Science 26, 39–83] and Hollnagel [1993. Human Reliability Analysis, Context and Control. Academic Press, London]. Finally, a computational model developed in ACT-R is presented as a first step towards validation of the major findings from the two experiments.

Keywords: Postcompletion error; Human error; Interface design; Modeling; ACT-R; Cognitive architecture; Routine procedural task; Error mitigation; Error intervention; Visual cue; Visual attention; Visual salience; Goal memory

1. Introduction

In one of the most influential works on human error, Reason (1990) defines human error as "all those occasions in which a planned sequence of *mental* or *physical* activities fails to achieve its intended outcome, when these failures cannot be attributed to the intervention of some chance agency." This definition does not attempt, however, to explain *why* or *how* such failures occur. Similarly, the intent of most popular error taxonomies (e.g., Norman, 1988) is not the prediction of errors. For the present perhaps the best we can look to in terms of error prediction is some probabilistic measure based on an analysis of the task and interface. Such human error identification approaches include Task Analysis for Error Identification (TAFEI) developed by Baber and Stanton (1994) and more statistical approaches, such as Swain and Guttman's (1983) Technique for Human Error Rate Prediction (THERP), both popular in safety-critical settings. All are somewhat lacking, however, in their account for the specific cognitive mechanisms and processes leading to erroneous behavior, as well as the conditions surrounding them. Moreover, once a potential error is identified using such a technique, the responsibility falls solely on the evaluator or designer to generate an appropriate solution.

The generation of a useful theory to support human error prediction in computer-based routine procedural tasks would require significant data beyond what currently exists. Such a theory, based on our current understanding of human cognition, would certainly lead to safer design solutions. There are major hurdles, however, that must first be overcome. As Rasmussen (1987) once suggested, such an endeavor would necessitate a human error "data bank,"

*Corresponding author. Tel.: +1 713 348 3770; fax: +1 713 348 5221.
E-mail address: byrne@acm.org (M.D. Byrne).

1071-5819/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.ijhcs.2007.09.001

from which predictions could be extrapolated. The creation of such an artifact would be a major undertaking, as capturing data on low-frequency errors is difficult and time consuming to set up in the laboratory (Wood, 2000). Fortunately, the last century's introduction of automation and computers has established an outstanding arena for this problematic element of human behavior to showcase itself. Subsequently, much effort has been given to analyze such errors in retrospect (e.g., root cause analysis, error taxonomies) and generate general design guidelines against them. Still, according to John and Kieras (1996), "No methodology for predicting when and what errors users will make as a function of interface design has yet been developed and recognized as satisfactory."

This paper reports two laboratory experiments on human error in interactive tasks as well as a novel approach to studying human error using the ACT-R cognitive architecture (Anderson et al., 2004). The use of ACT-R to develop a theory of human error goes one step beyond existing error taxonomies or error prediction methods often based on a traditional task models (e.g., Hierarchical Task Analysis (HTA); Annett and Duncan, 1967). Following Rasmussen's concept of a human error "data bank," the theory is dynamic and modeled on human data. An interesting demonstration is provided of the difficult task of human error identification from a cognitive perspective, which existing methods of task representation and error identification seem unable to convincingly manage.

Current theory in psychology suggests that the same processes producing successful human performance can also be looked to as the source of error (Baars, 1992; Reason, 1990). Reason (1990) describes this in terms of a "cognitive balance sheet," where correct performance and systematic errors are placed on opposing sides. For example, the emergence of automatic behavior through delegation of control to lower-level processes introduces the opportunity for behavioral "slips" to arise. Hence, in order to understand how errors occur, it is necessary to consider the cognitive mechanisms that govern correct human behavior.

Rasmussen's (1987) "Skill-Rule-Knowledge" (SRK) description of task performance provides a general framework of cognitive control mechanisms, which can be used to describe how errors occur. It identifies three separate levels of cognitive control displayed during task performance: skill-based, rule-based, and knowledge-based. Each corresponds to a different degree of familiarity with the task and environment, with knowledge-based behavior representing the least degree of control and familiarity and skill-based the highest. With experience a person proceeds sequentially through the three stages of the model, moving from lowest (knowledge-based) to highest (skill-based). At the rule-based level, rules for behavior are selected using selection criteria based on the mental model the operator has constructed in their mind about a system.

(1) *Match* to salient features of the environment or internally generated messages.
(2) *Strength* or the number of times a rule has performed successfully in the past.
(3) *Specificity* to which a rule describes the current situation.
(4) *Support* or the degree of compatibility a rule has with currently active information.

Failure modes stem from either the application of bad rules or misapplication of good rules due to incorrect rule selection. These rules may be active simultaneously, with several competing for instantiation. These also control the occurrence of errors at the rule-based level, in keeping with Reason's (1990) idea of a "cognitive balance sheet."

1.1. Postcompletion error

Noting the general lack of specificity in the existing theories of human error, Byrne and Bovair (1997) moved to develop a computational theory for one widely cited (e.g., Rasmussen, 1982; Young et al., 1989) omission error, postcompletion error (PCE). PCEs can be roughly defined as errors that occur when the task structure demands "that some action…is required after the main goal of the task…has been satisfied or completed" (Byrne and Bovair, 1997, p. 32). Some commonplace examples include forgetting to remove the original after making a photocopy, leaving a card in the automated teller machine (ATM) after withdrawing cash, and failing to replace the gas cap after filling up a car. With this particular class of error, the actor possesses the correct knowledge necessary to execute the task—which is usually performed correctly—yet still generates systematic errors.

Even for operators highly familiar with the task, the isolation of a postcompletion step within the task structure makes omissions likely. This is especially true when the actor is further affected by external factors such as a working memory load and/or fatigue, as well as internal human tendencies such as hillclimbing (Polson and Lewis, 1990; Gray, 2000; Byrne and Davis, 2006). Byrne and Bovair (1997) hypothesized that these errors were due to excessive working memory load leading to goal loss, or an omission of a step from the task at hand. Since with PCEs the actor omits a specific subgoal rather than forgetting what to do altogether (the main task goal), the source of the error was thought to more likely be working memory than long-term memory. A more recent study by Reason (2002) examined the photocopy example in detail, finding PCEs to be the most common type of omission in that task. Three high-level explanatory observations were provided:

(1) The emergence of the last copy generates a strong but *false completion signal* since the main goal of copying is achieved before all necessary steps (subgoals) are complete.
(2) The proximity of this false signal to the end of the main

task allows for the *attention to be increasingly diverted* to the subsequent task.
(3) The emergence of the last copy indicates that it is no longer necessary to put in another original leaving it *functionally isolated*.

Because commission of this type of error is reliable under high working memory load, Byrne and Bovair (1997) suggest that the only absolute solution is to design it out. A common design solution is to create a forcing function by rearranging the task such that the user is forced to complete the otherwise potentially omitted step in order to achieve the main task goal. ATMs initially faced the same sort of problem as the auditory and visual reminders implemented in early models failed to reliably remind customers to withdraw their card. As a result, many ATMs today now feature a forcing function that prevents the user from proceeding with a transaction before the card is withdrawn.

1.2. Hierarchical control structures and goal management

Many of the assumptions behind the theory of PCE reside on the concept of hierarchical control structures and their retention by skilled operators. In previous studies by Byrne and Bovair (1997) and Serig (2001), participants reliably generated errors at the postcompletion steps both within individual subtasks as well as within the larger task, in keeping with the idea of a hierarchical task structure. Cognitive modeling work by Kieras et al. (1997) has also provided strong evidence to suggest that even well-practiced experts, such as telephone assistance operators, do not abandon such task hierarchies. Altmann and Trafton (1999) propose that this ability to break down complex tasks and problems into hierarchies and subgoals, "may be to complex cognition what the opposable thumb is to complex action."

Traditionally, these types of goal-based processing strategies have relied solely on a "task-goal" stack that essentially predicts perfect memory for old goals. However, Altmann and Trafton's (2002) goal-activation model offers an alternative account to this approach that provides a more straightforward account for the types of errors found in human behavior. In essence memory and the environment (i.e., dual-space, Rieman and Young, 1996; internal Q3 and external representations, Zhang and Norman, 1994) are substituted for a goal stack, and task goals are considered as ordinary memory elements with encoding and retrieval processes that must overcome noise and decay. Retrieval cues from the environment dictate the reactivation of suspended goals with perceptual heuristics acting as a substitute for the stack-native last-in, first-out rule. This model makes several predictions about PCEs and the characteristics of a successful cue:

(1) Any *salient* cue (e.g., a loud beep) should be sufficient to prime a postcompletion action.
(2) It should *not* be necessary to put the postcompletion action on the critical path.
(3) Reminders at the start will *not* help a PCE at the end because they are masked by other goals.
(4) Just-in-time priming from environmental cues are the *only* reliable reminder for postcompletion actions.

1.3. Task modeling and human error identification

Hierarchical control structures and goal management are major components of the popular cognitive task modeling approaches. HTA (Annett and Duncan, 1967), developed to aid the investigation of complex non-repetitive tasks, forms the basis of human error identification methods such as Baber and Stanton's TAFEI (1994). By breaking down the task goal structure hierarchically, the human side of the interaction is modeled in conjunction with state-space diagrams detailing the behavior of the artifact. Similarly, Goals Operators Methods and Selection Rules (GOMS; John and Kieras, 1996) offers a means to represent knowledge required by humans in computer-based tasks for correct performance. Such methods make it possible to analyze the dynamics of interaction, even to the point of identifying potential errors (Wood, 2000).

With these and other extant methods of human error identification, such as Cognitive Reliability and Analysis Method (CREAM; Hollnagel, 1998) and THERP (Swain and Guttman, 1983), correct performance must first be modeled in some form before the analyst may proceed to "predict" or identify potential errors in the task procedure. This falls in line with the notion that correct performance and human error are closely tied. Nonetheless, two major weaknesses afflict these techniques (Stanton, 2004), the first of which is their weak account for the external environment (e.g., stress and noise). THERP, most notably, has been criticized for its reliance on fixed error probabilities and lack of account for varying levels of stress. Even representation of the system is limited to simple space-state diagrams at best (e.g., Baber, 1996). As a result, existing methods are poorly suited to predict error in highly perceptual and dynamic tasks, such as flying a plane or driving a car. The second major drawback to these methods is that they rely heavily on the skill of the analyst. This can lead to both inter- and intra-analyst reliability problems for obvious reasons. Despite their weaknesses, however, there are good reasons for why techniques such as THERP continue to thrive in certain application areas such as power plants (Kirwan, 1992). Our intention is not to belittle their significance but rather seek areas for improvement in order to deal with more highly perceptual and dynamic computer-based tasks. We have proposed to do this by using visual cues to examine the visual aspects of human error in computer-based routine procedural tasks.

1.4. Two issues: "Salience" and least-effort

In addressing human error in these instances, it is necessary to consider two critical issues: "salience" and least-effort. Previous research (Chung, 2001) has demonstrated the importance of both principles in error remediation. The latter, as previously discussed, is a result of the human tendency to act as "cognitive misers," who are inclined to take road of least cognitive effort and use decision rules of thumb or heuristics rather than systematically analyzing each decision (e.g., Fiske and Taylor, 1994). Studies of problem solving have demonstrated the pervasive tendency of humans to hillclimb (difference-reduction) or take the shortest perceived route to complete a task or solve a problem (e.g., Newell and Simon, 1972). With hillclimbing, past states do not have to be retained and planning more than the next step is not required, thus reducing the required level of cognitive effort (Anderson, 1995). Polson and Lewis (1990) uncovered a similar propensity in computer-based tasks where perceptual similarity alone is often used to select actions appearing to offer the greatest progress towards the goal. This idea has taken on several iterations such as the label-following heuristic proposed by Englebeck (as cited in Polson and Lewis, 1990) which describes the tendency of novice users to select actions in computer-based tasks by comparing the descriptions of available actions with a description of the goal.

Gray (2000) applied the hillclimbing principle to the interactive behavior involved in programming a VCR. He states that people adhere to a "least-effort principle" in operating the device, mapping prior knowledge to the device and relying on place-keeping. For the task of programming a VCR, people progress using both global and local place-keeping, which "entails knowing what parts of the task have been completed and what parts remain to be accomplished" (p. 221). Task-based and device-specific goal completion is tracked at the global level while progress on the current goal is followed at the local level. This leads to what Gray (2000) calls "display-based difference reduction," where differences between the current state of the world and the final goal state are progressively reduced using perceptual information rather than knowledge in memory. In this manner, local place-keeping becomes a primarily perceptual rather than cognitive task. This is similar to the idea of distributed representations (Zhang and Norman, 1994) or the dual-space nature of human–device interaction (Rieman et al., 1996).

Both Reason's (2002) and Altmann and Trafton's (1999) recommendations for handling PCEs assume that the mitigating cue or reminder is salient or conspicuous. The goal-activation model (Altmann and Trafton, 2002) stipulates that some earlier cue (internal or external) is associatively linked to a subsequent target, such as the postcompletion action. In fact, Altmann and Trafton (2002, p. 64) take the tendency of skilled users to generate correct behavior most of the time as "evidence of deliberate cognitive operations undertaken to meet the priming constraint—to ensure the existence of an associative link to the postcompletion action, and to ensure attention to the right cue at the right time." This is a good working hypothesis as to what occurs when *correct* behavior is displayed.

Nevertheless, in instances where a PCE does arise, it is left to establish if the *incorrect* behavior can be traced to the stage of perception (i.e., failure to see a cue), attention (i.e., failure to attend to a cue), or goal association and retrieval (i.e., retrieving an incorrect goal or simply failing to retrieve a goal). This requires one to determine if a perceptual cue is conspicuous or salient, particularly under the external conditions of the task. In road accidents, for example, failures at the stage of attention are most common, as drivers fail to attend to a plainly visible object such as a road sign (Green and Senders, 2004). In this paper, we are using the term "cue" to refer to something specific on the interface, but it could be more generally applied to the overall perceptual environment. Results from the following experiments demonstrate the difficulty of tracing errors in these instances, an undertaking that extant error identification methods leave to the designer or analyst to overcome using their "expert judgment."

2. Experiment 1

While designers may opt to place the hanging postcompletion action "on the critical path" to reduce or eliminate their omission, such as with many ATMs, this is often impossible or too expensive given an existing system. An alternate solution is to add a cue to remind the user of some action (e.g., a bright warning label or a flashing green light on the card slot of the ATM). It is well known that the selection or noting of visual cues or features is automatic (e.g., Treisman, 1986), and processing of them occurs regardless of whether or not they are informative (e.g., Jonides, 1981; Remington et al., 1992). Evidence suggests that visual cues are even more powerful when people are directed to look for a specific feature (e.g., Most et al., 2000). Even the addition of a simple visual cue, such as an orange dot, can bring about changes in how people interact with physical objects such as doors (Wallace and Huffman, 1990). In light of these theories and experimental evidence, it seems reasonable to suspect that simply cueing or priming a suspended goal would reduce PCEs or omissions in an interactive task.

If a necessary condition for slips is attentional capture of some form, external reminders or environmental cues similarly exploiting attentional capture should reactivate the suspended subgoal for the postcompletion action at the appropriate time. Cognitive aids functioning in this manner are frequently used in industrial settings and aviation to reduce human error. By prompting the actor to make sure that all steps have been completed in the task, they fundamentally augment the limited capacity of working memory that is the root cause of many errors. Further-

Please cite this article as: Chung, P.H., Byrne, M.D., Cue effectiveness in mitigating postcompletion errors in a routine procedural.... International Journal of Human Computer Studies (2007), doi:10.1016/j.ijhcs.2007.09.001

more, many existing everyday devices and computer applications have indicator lights, beeps, alarms, etc. that act as reminders, although their efficacy may sometimes be questionable. While there are undoubtedly many design considerations that must be made before implementing a visual cue in the real world (e.g., clutter on the interface), the aim of the current work was simply to determine the critical characteristics of a successful cue.

Errors may also decrease over time if there is a downstream cost for incorrect performance in the task. In this case making an error on a given trial leads to a cost in time or effort further down in the task sequence. For instance, forgetting to save one's work before shutting down a program would require one to go back and redo the work to successfully complete the task. In theory, the cost of having to do the work twice should provide incentive for a change in behavior. If feedback is immediate and obvious, the operator should be motivated to execute the correct behavior in subsequent interactions.

2.1. Predictions

Detailed analyses of simple and practical countermeasures to error and their corresponding effects will increase our understanding of the mechanisms behind human error in human–computer systems. Using computer-based tasks employed previously by Serig (2001) and Byrne and Bovair (1997) to successfully study PCEs, the objective of Experiment 1 was foremost to study the effects of: (1) a simple automated visual *cue* and (2) a downstream cost in the form of a *mode error*. PCEs were chosen for this study, having been shown to be reproducible in previous inquiries and observations from real-world settings.

The computer-based tasks had routine procedures and were administered under the cover story of a Bridge Officer Qualification program (Star Trek). There were three tasks (Navigation, Tactical, Transporter) in total, with the experimental manipulations applied in the Tactical task described in Fig. 1. The goal structure of this task, as explained in training to participants through paper-based manuals, was hierarchical, and steps were grouped accordingly. The experimental cue appears at the step where the participant is required to press the *Tracking* button a second time to disengage the firing system. This was in fact the postcompletion step, with the potential PCE being for participants to move on to the *Main Control* step without first disengaging the system. Finally, a concurrent working memory task was administered to generate a sufficient frequency of PCEs for study as in previous work (Byrne and Bovair, 1997; Serig, 2001). Letters were presented in random order over headphones, and participants were prompted on-screen at varying intervals to recall the last three in order.

The idea of humans as ''cognitive misers,'' their further tendency to hillclimb, evidence for display-based difference reduction (Gray, 2000), and the predictions of goal-activation model (Altmann and Trafton, 1999) all suggest that humans will use means convenient to them and even cut corners to reduce cognitive work. Thus, it was predicted that participants would exploit a just-in-time red visual cue as a reminder to complete the PCE-prone step. Participants were given extensive training and should

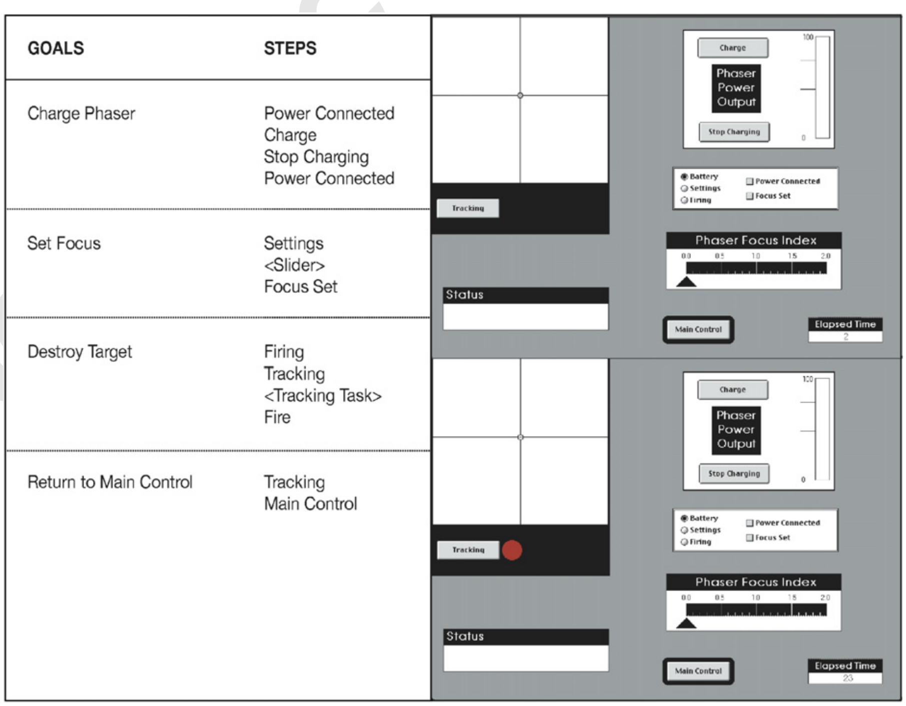

Fig. 1. Task hierarchy and screenshot of the Tactical task. The bottom right figure shows the location of the visual cue to the right of the ''Tracking'' button.

Please cite this article as: Chung, P.H., Byrne, M.D., Cue effectiveness in mitigating postcompletion errors in a routine procedural.... International Journal of Human Computer Studies (2007), doi:10.1016/j.ijhcs.2007.09.001

therefore have acquired adequate knowledge to quickly comprehend the cue. The cue adhered to recommendations by Reason (2002) to be conspicuous (at the critical time) and contiguous (proximal), research by Yantis and Jonides (1988) showing onsets to capture attention better than color or intensity, predictions by Altmann and Trafton (1999) regarding just-in time priming from environmental cues, and real-world prominence on existing devices and applications.

It was also hypothesized that feedback in the downstream cost condition would help participants remember to complete the postcompletion step on subsequent trials. With the target task this condition was instantiated by leaving the console at the postcompletion step if it was omitted on a previous trial. This is essentially a mode error, since when the participant returned to the system, it would fail to respond until the formerly omitted postcompletion step was first executed. Feedback in the form of a time cost and point deduction incurred by having to reorient oneself with the awkward state of the system was expected to motivate participants to pay increased attention to the postcompletion step on subsequent trials. This may potentially be driven by a desire to reduce overall effort, as predicted by the cognitive miser account.

2.2. Method

2.2.1. Participants

A total of 81 (36 male, 45 female) undergraduate and graduate students aged 18–30 from the Rice University participated for credit toward a requirement in a psychology course and/or prizes. Amazon gift certificates ($25, $15, and $10) were awarded to the top three finishers in terms of points, based on correct task performance and speed.

2.2.2. Materials

The materials for this experiment consisted of a paper instruction manual for each of the three tasks (Navigation, Transporter, and Tactical), Apple iMac computers running the Bridge Officer Qualification application written in Macintosh Common Lisp, and Sony MDR-201 headphones. The manuals were thorough in detail and offered illustrations of the interface and diagrams for each step. Organization of the instructions was hierarchical (see Fig. 1), in accordance with the idea of hierarchical control structures. Step-by-step summaries at the end of each manual provided a complete picture of the task for review.

2.2.3. Design

This experiment used a two-factor design with trial at testing as a within-subjects factor and task assignment as a between-subjects factor. Task assignment consisted of four conditions: control (no intervention version of the Tactical task), cued (cued version of the Tactical task), mode error (the downstream cost version of the Tactical task), and a combined cued and mode error condition. Participants were randomly assigned to one of these four conditions. The primary dependent variable was the frequency of PCEs made during the Tactical task (out of the total number of opportunities). Completion time at the postcompletion step was also a measure of interest.

2.2.4. Procedure

Participants were run in two sessions, spaced 1 week apart. The first session served as a training session using written documentation for each of the tasks: Navigation, Tactical, and Transporter. Group assignment was randomized across participants and order of training on the three bridge station tasks was randomized for each. Only the Tactical task contributed to the measure of PCE. Once participants successfully completed the training trial and logged three subsequent error-free trials, they were allowed to move on to the next task. Errors resulted in warning beeps and messages, ejected the operator to the main control, and restarted the task. This prevented participants from completing training without having gone through each of the tasks at least four times with all steps done correctly and completely. This on average did not take longer than 40 min, but participants were given up to an hour. Once training was complete for all three tasks, they were reminded that they would be competing for prizes in 1 week.

The second session consisted of the test trials for the three bridge station tasks. Participants completed 13 trials of each task in random order, for a total of 39 trials for the test day. During the second session, the experiment program emitted warning beeps on error commission to warn individuals but did not eject them to the main control as in training. Moreover, warning messages and reminders were removed and trials were continued until the task goal was met.

The concurrent working memory letter task was also introduced during testing. As in the studies by Serig (2001) and Byrne and Bovair (1997), its function was to increase working memory load during task performance. Participants were presented with auditory stimuli in the form of randomly ordered letters spoken through the headphones at a rate of one letter every 3 s. A tone was presented randomly at intervals ranging from nine to 45 s, upon which the participants were directed to recall the last three letters in order and type them into the text box that appeared on the screen. This was the same for all four conditions.

Participants were encouraged to work both accurately and quickly by means of a scoring system, an onscreen timer, and prizes. The scoring system incremented 25 points for each correctly executed step and decremented 50 points for each incorrect. Up to 100 points were awarded for task completion within a set time. For every incorrect working memory recall trial, the score was decremented 200 points. No points were accumulated for successfully completing a recall trial. The large weight placed on the recall task was due to an observation made during a previous experiment (Serig, 2001) of participants tending to neglect the working memory task. At the end of each trial, participants were informed of their task completion time, number of errors committed, and score. Accumulated points were used in competition for prizes.

Since the current work is focused on the effect of an automated cue and downstream cost on PCE, the experiment program allowed participants to complete a trial at testing without executing the postcompletion step (Tactical task), although a warning beep was emitted. The cued version of the same task featured a simple red visual cue appearing adjacent to the "Tracking" button at the postcompletion step on every trial. In the mode error condition, the Tactical console stayed at the postcompletion step if it was forgotten on a previous trial and prevented the operator from proceeding until it was first completed. Finally, in the combined mode error and cued condition, the system combined both the downstream cost of the mode error condition with the visual cue. Analysis of the results focused on the data collected from the testing day.

2.3. Results

The mean frequency (out of all trials) of PCEs across groups was of primary interest in our analysis. Participants who made an omission at the postcompletion step at more than 50% frequency were removed. This was based on the rule that PCEs occur when the operator has the correct plan. If errors were found to occur in more than 50% of the trials, it is most likely that the participant had not correctly remembered the task correctly from training. Fifty percent was chosen as the point of delineation, as participants tended to pool into two groups: those with PCEs ranging from 38% and below and those with postcompletion frequency at 70% frequency and above. Thirteen participants were found in the second category and removed: five from the control group, three from the mode error group, two from the cued group and three from the combined group. This left 13 participants for the mode error and cued groups, 12 for the combined, and 11 for the control, combining for a total of 49 participants included in the final sample.

2.3.1. Postcompletion frequency

Postcompletion frequency, or the number of PCEs committed out of the total number of opportune steps (frequency = number of errors at a step X_i/total number of opportunities for error at step X_i) was analyzed by condition (Fig. 2). Contrary to the hypothesis, no reliable effect of group on PCE frequency was found, $F(3, 45) = 1.22$, $p = 0.31$. Mean frequencies across groups were low, in comparison to the results of Serig (2001), once data for participants with greater than 50% PCE frequency were removed. A power calculation for the main effect of condition showed the effect size to be 0.25, yielding a somewhat low power of 0.34 at the 0.05 level. Power to detect a medium-sized effect (i.e., effect size of 0.4), however, was a somewhat more respectable 0.61, but this is hardly conclusive.

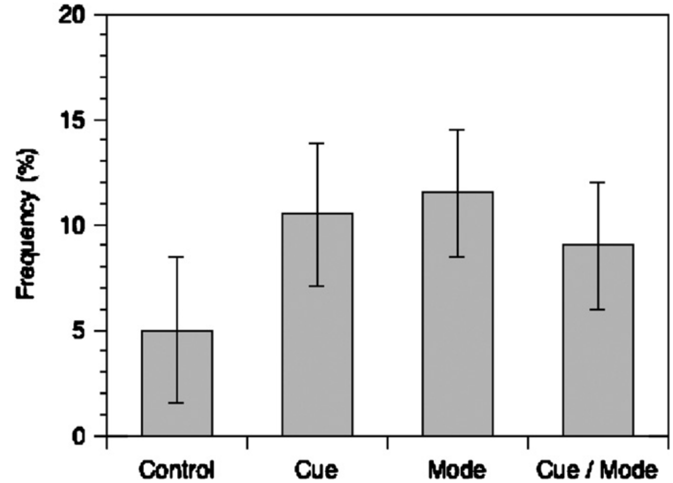

Fig. 2. Postcompletion error frequencies by condition. Error bars indicate standard error of the mean.

2.3.2. Postcompletion step times

The average initial postcompletion step times reflect the difficulty faced by most of the participants on the first trial: 5736 ms (Control), 6911 ms (Cued), 8040 ms (Mode), and 5174 ms (Combined). This is the time taken from the postcompletion, or the penultimate (second-to-last) step to the last step in the task. Times declined overall for the rest of the trials, falling within the range found by Serig (2001; ~4500 ms, Day 2a). This is reflected by participant responses to the questionnaires, in which many stated that it was not very difficult to recall how to perform the tasks after the first trial.

A repeated-measures ANOVA revealed a reliable main effect of trial, $F(12, 504) = 7.323$, $p < 0.01$ and linear trend, $F(1, 31) = 28.15$, $p < 0.001$. There was also a reliable main effect of condition, $F(3, 42) = 8.23$, $p < 0.001$, although Tukey's HSD test showed none of the intervention conditions to be significantly different from the control, $p > 0.05$. Finally, there was no significant trial by condition interaction, $F(36, 504) = 1.23$, $p = 0.17$. Differences in task completion times were also unreliable between conditions, $p > 0.05$.

2.4. Discussion of Experiment 1

Despite the lack of reliable differences in reaction times or error commission at the postcompletion step, the results have valuable implications. First, the fact that the visual cue did not significantly reduce the number of PCEs committed by the participants demonstrates that simply following a design heuristic and placing a contiguous and proximal reminder can be ineffective. While adding a large red cue to onset beside the button to be pressed seemed like an intuitive solution, participants made errors at this

Please cite this article as: Chung, P.H., Byrne, M.D., Cue effectiveness in mitigating postcompletion errors in a routine procedural.... International Journal of Human Computer Studies (2007), doi:10.1016/j.ijhcs.2007.09.001

postcompletion step regardless. Second, returning to the issue of salience, it remains to be determined whether this is a problem at the level of vision, attention, or memory. Either participants who made errors did not see the cue, did not pay attention to it, or forgot its association with the postcompletion action of pressing the "Tracking" button. The added working memory load and speed-accuracy tradeoff, encouraged by the time and performance pressures, were likely contributing factors.

Neither did the downstream cost in the form of a mode error present a significant change in behavior at the postcompletion step. It was hypothesized that this manipulation would act as negative feedback in response to an error, bringing about a change in behavior on subsequent trials. However, as the results showed, this did not reliably occur. On some occasions the mode error incurred a considerable cost in additional time at the initial step of the posterror trial. Still, participants' error rates did not decrease on subsequent trials, contrary to our expectations. This seems to follow findings by Serig (2001) that demonstrated participants' error commission to be relatively independent of negative or positive feedback about task performance.

The performance of the 10 participants who were unable to recall the postcompletion step at above 50% frequency was also unexpected. Fortunately, the removal of these participants does not substantially alter the pattern of results. The fact that ten of the initial 49 participants were so poor at recalling this step is rather remarkable considering that they had all completed the extensive training session successfully, as in previous work by Serig (2001) and Byrne and Bovair (1997). Nonetheless, despite a generally positive relationship between the number PCEs and other errors, some participants with a high frequency of PCEs made relatively few other errors. This seems to support the notion that the postcompletion step is particularly difficult to remember. In subjective reports, participants (including those who committed the error at over 50% frequency) reported that they did not have much trouble recalling the task as a whole. Some of those who did commit PCEs at over 50% frequency claimed that the task seemed to change or that they did not understand why the program kept beeping at the last step. Again, this suggests that they had distinct problems recalling the postcompletion step and most likely lost the concept of the tracking system (a mode that must be turned off) explained to them at training.

Among those participants removed from the data were several who failed to recall the postcompletion step altogether. Since the system no longer offered error messages telling the participant what the correct action was at each step, it is likely that they continued unaware of their mistakes as the trials progressed. Driven by performance and time pressures, these participants proceeded through all 13 trials, with the incorrect rule to ignore the cue or skip the postcompletion step gaining strength with repetition. As aforementioned, a likely explanation for the unreliable effect of the interventions was that they were insufficiently salient at the level of vision and attention or participants were unable to remember the association between the cue and correct postcompletion action. This issue was examined more closely in a follow-up experiment.

3. Experiment 2

The primary purpose of Experiment 2 was to address the question of the ineffective cue in Experiment 1 using other types of cues. Specifically, cues varying in appearance and function and based on existing research in the field were investigated. In addition, previous issues with training, train-test delay length, salience of interventions and strength of association, and the number of trials at testing were all considered in devising a follow-up study. A second Medical task and interface was introduced under the fictional scenario of a Starfleet Chief Medical Officer training program to examine differences in the effects of the cues across tasks.

3.1. Intervention implementation

Despite the supporting theories and evidence suggesting that a simple visual cue would be effective as a reminder, the red onset used in Experiment 1 was not found to be effective. This may be explained by Hollnagel's (1993, p. 299) assertion that the strength of a cue is relative to its specificity. Hence, it is the cue's strength relative to the other elements of the task that is important when assessing its potential as a reminder. This claim is based on the observation that when a task is considered trivial, attention is more easily diverted. Performance becomes controlled by more error-prone generic functions such as "look for cue which indicates a turn," rather than exact intentions such as "look for cue-X, then turn to the right." For this reason, it is important to design and train participants to visually specific (i.e., meaningful) cues demanding explicit actions. Generic cues can potentially lead to errors caused by multiple cues with ambiguous specification of the required action or state, if used within complex systems (Norman, 1988). A lack of specificity may have been the problem with the cue used previously.

In a study by Monk (1986), auditory cues were used to drastically reduce the occurrence of mode errors, such as those introduced in the downstream cost condition of Experiment 1. Keying-contingent sounds were used on a keyboard-based computer game to enhance draw the user's attention to a change in the system's mode. This worked well because the nature of mode errors is such that they generally occur when the user is unaware of the system's current mode and its consequences. Monk (1986) observed that display changes, however, are effective when the user is required to look at the relevant parts of the display at the appropriate moment in the dialogue. Pointing devices such as the mouse force users to focus on the screen, making

Please cite this article as: Chung, P.H., Byrne, M.D., Cue effectiveness in mitigating postcompletion errors in a routine procedural.... International Journal of Human Computer Studies (2007), doi:10.1016/j.ijhcs.2007.09.001

small visual changes or cues, which may go unnoticed with other types of interaction, more likely to be effective.

3.2. Cue attributes

Evidence suggests that the visual attributes most effective for attracting attention on a computer interface in order are as follows (Sutcliffe, 1995):

(1) Movement (blinking or change of position).
(2) Shape and size (character font, shape of symbols, text size, size of symbols).
(3) Color.
(4) Brightness.
(5) Shading and texture (different texture or pattern).
(6) Surroundings (borders, background color).

Sutcliffe (1995) suggests that care be taken to ensure that the user population interprets the warning icon or cue as the designer expects. Furthermore, such attributes should only be applied sparingly, as the presence of many conflicting stimuli can essentially dull their individual effectiveness. These guidelines directly relate back to Hollnagel's (1993) idea of cue strength and specificity.

For color, red, green, and yellow are recommended as status indicators, each corresponding to its meaning on a traffic light. To draw attention against a dark background, white, yellow, and red are most effective, although yellow offers the best visibility (Sutcliffe, 1995). Based on these recommendations from the literature and the failure of the cue in Experiment 1 to reduce PCEs against the control condition, alternating red and yellow blinking arrows (see Fig. 3) were used in both the Tactical task and a new Medical task. The directional shape of the cue pointing towards the button made it visually specific, while the blinking and colors made it salient to an extreme against the black on the interface. To ensure these assumptions were reasonable, the cue was presented to several people in a small pilot study to ensure that people associated the cue with the required step, as noted by Sutcliffe (1995).

3.3. A mode indicator

In addition to the new cue, which appeared just-in-time with the postcompletion step, a mode indicator condition was introduced to examine the effect of a cue appearing prior to and remaining on through the postcompletion step. In contrast to the downstream cost (mode error) used in Experiment 1, this condition provided prior warning of the postcompletion step by highlighting the mode, as in Monk's (1986) work with auditory cues. The previously used Tactical interface of the Bridge Officer Qualification program was redesigned for this manipulation. As shown in Fig. 4, the mode indicator consisted of a green light appearing on the "Tracking" button, crosshairs showing in the targeting window, and the message "Tracking Mode Enabled" appearing in yellow against black. It was expected that the combination of these three novel features would be sufficient to inform the user that the system was in a distinct Tracking mode. Combined with the given IF–THEN rule at training (i.e., "If you see a mode indicator light, the system is on."), the presence of the mode indicator was a reminder to later shut down the Tracking system, indicated by the green light and message turning off after a second press of the "Tracking" button. Similar mode indicators are commonly found in real-world

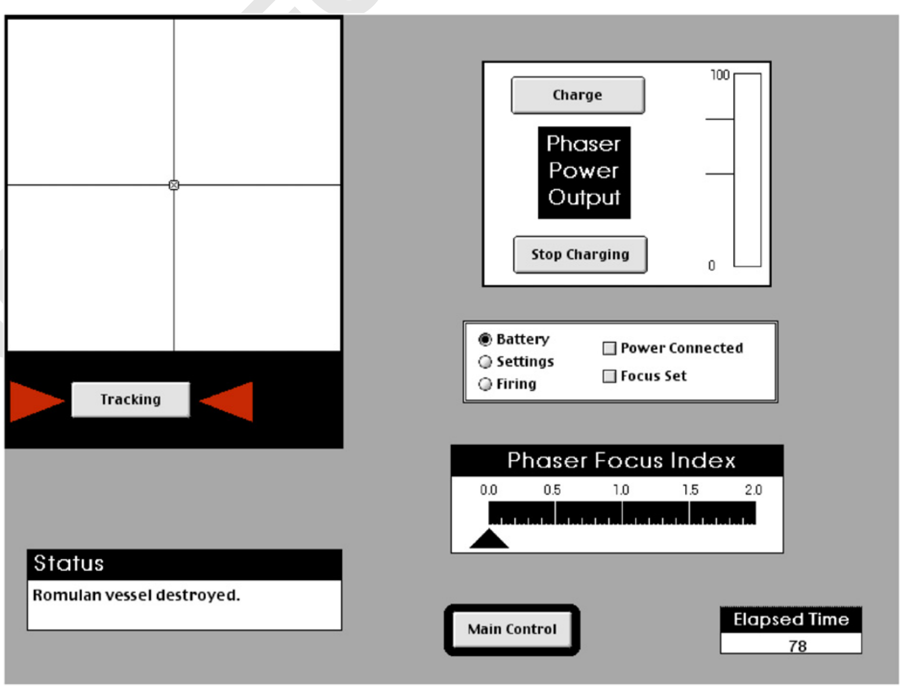

Fig. 3. Blinking (red and yellow) arrows by the "Tracking" button.

Please cite this article as: Chung, P.H., Byrne, M.D., Cue effectiveness in mitigating postcompletion errors in a routine procedural.... International Journal of Human Computer Studies (2007), doi:10.1016/j.ijhcs.2007.09.001

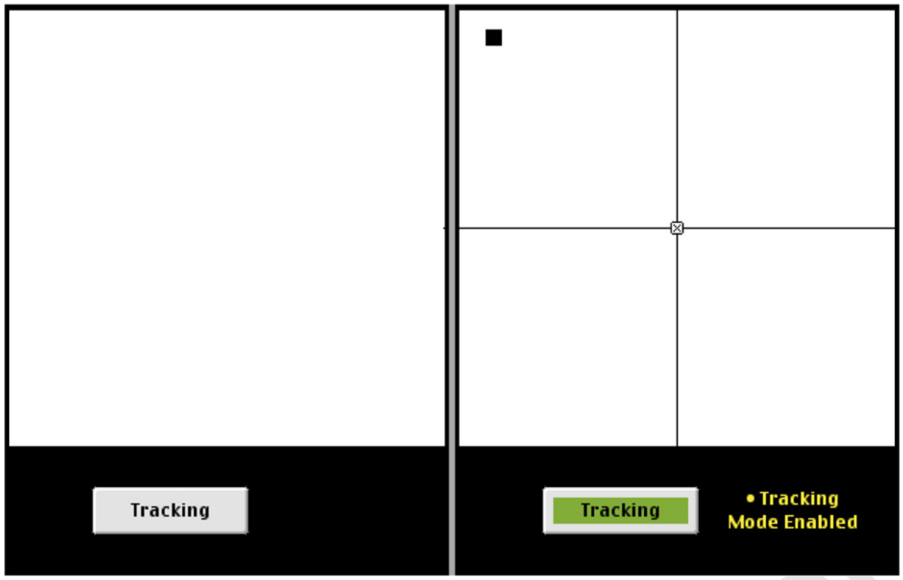

Fig. 4. Mode indicator in the Tracking window in off (left) and on (right) states.

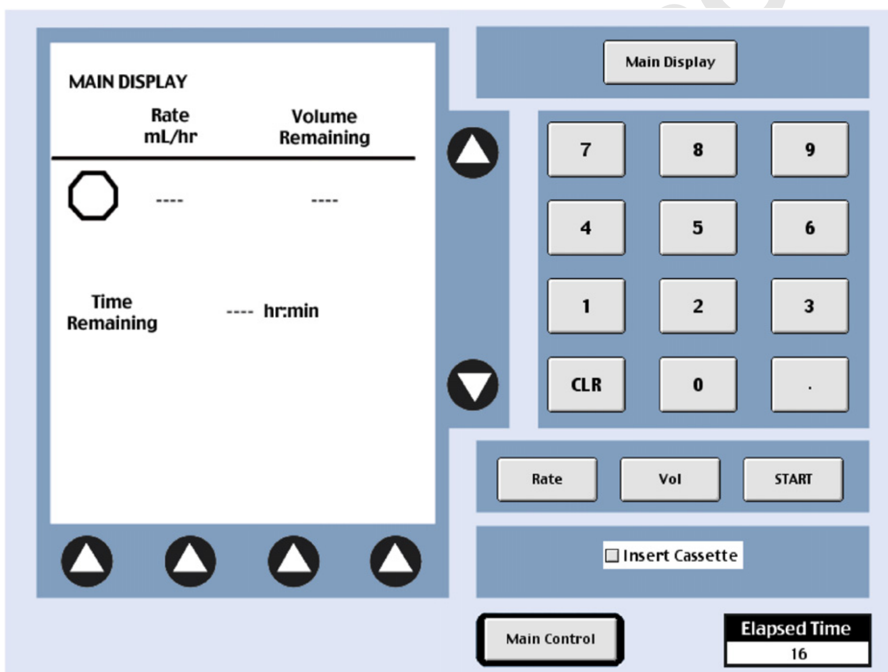

Fig. 5. Medical console.

devices, such as on automobile dashboards and television remote controls.

All three conditions (Control, Cued, and Mode) with the cue and mode indicator appearing at the "Main Display" button instead of at "Tracking" were mirrored in the new Chief Medical Officer Qualification program (Fig. 5). This task was similar to the Tactical task (Table 1) in that it had a postcompletion step, as identified by HTA (Chung et al., 2003).

3.4. Predictions

Based on the results of Experiment 1 and research supporting the manipulations introduced in Experiment 2, there were three key predictions:

(1) The new cue was visually more specific and would therefore reliably reduce the PCE frequency versus the control, whereas the more generic cue used in Experiment 1 did not.

Table 1
Comparison of Tactical and Medical tasks

Tactical	Medical
Charge phaser (5 substeps)	Insert cassette (1 substep)
Set focus (3 substeps)	Program rate (2 substeps)
Track target (3 substeps)	Program vol (2 substeps)
Fire phaser (4 substeps)	Start flow (2 substeps)
Return to Main Control	

(2) The mode indicator would also reduce the PCE frequency, although not as much as the cue, since it did not onset just-in-time and relied on peripheral information.
(3) The Medical task would display the same trend of PCE frequencies as the Tactical task, but to a lesser magnitude overall, since it contained fewer steps.

3.5. Method

Training in the previous experiment was thorough and detailed, using manuals to promote mental models of the system (Norman, 1988). However, as noted, there were still problems in Experiment 1 with several participants who failed to recall the postcompletion step at greater than 70% frequency at testing. Hence, the association between the system change (pre–postcompletion) step and maintenance (postcompletion) steps was further emphasized in Experiment 2, with four main changes introduced across all three conditions:

(1) The training manuals were revised to be more specific and promote a stronger mental model, with more detailed pictures, diagrams, and instructions.
(2) The delay between training and testing was reduced to 2 days, in light of the numerous participants who had trouble recalling the task in Experiment 1.
(3) Paper-based quizzes were given at the end of training to reinforce participants' understanding of the tasks.
(4) The system reminded participants of the postcompletion step on the first trial at testing, if they forgot.

3.5.1. Participants

Ninety-one undergraduate (42 female and 49 male) students from Rice University aged 18–35 participated for credit toward a requirement in a psychology course as well as the possibility additional cash prizes ranging from \$10 to \$40.

3.5.2. Materials

The materials for this experiment consisted of a paper instruction manual for each of the four tasks (Navigation, Transporter, Tactical, and Medical), paper-based quizzes for the first day, Apple iMac computers running the Bridge Officer Qualification and Chief Medical Officer Qualification applications written in Macintosh Common Lisp, Sony MDR-201 headphones, and a web-based general questionnaire.

3.5.3. Design

Experiment 2 used a two-factor design with task and intervention as variables. Task consisted of two conditions: Bridge Officer (Tactical) and Chief Medical Officer (Medical). Intervention consisted of three conditions: no intervention (Control), alternating red and yellow blinking arrows (Cue), and a mode indication for the system state change (Mode). Participants were randomly assigned to one of the six groups. The primary dependent measure was the number of PCEs made during the Tactical and Medical tasks. Other dependent measures of interest included response times at the postcompletion step and the overall number of errors per task.

3.5.4. Procedure

All procedures were essentially the same as in Experiment 1. However, when training was complete, participants were reminded this time that they would be tested for prizes in 2 days and given a short paper-based quiz to ensure that they had a correct understanding of the tasks. Also, in Experiment 2, participants completed 17 trials of their assigned postcompletion task (Tactical or Medical) and 11 trials for each of the two dummy tasks (Navigation and Transporter) for a total of 39 randomized trials on the test day. The number of postcompletion task trials was increased from 13 to provide greater statistical power. In Experiment 2, the program also reminded participants at testing of the correct postcompletion step, if they made an error on their first trial. All participants were encouraged to work accurately and quickly, by means of a scoring system, prizes, and an onscreen timer, and were subjected to the same working memory task at testing, as in the first experiment.

3.6. Results

Of the original 91 participants, data from 82 were used in the final analysis. The main reason for the loss of data was participant failure to show up at their assigned time at testing. Our primary measure of interest was again the frequency of errors at the postcompletion step in the target tasks (the step immediately following completion of the main task goal). In contrast to Experiment 1, there were no participants with greater than 50% PCE frequency, suggesting that participants had less trouble remembering the tasks at testing. Outliers in the response time data greater or less than three standard deviations from each participant's mean were removed and replaced with their mean.

For the Tactical task, mean PCE frequencies were 6.81%, 0% (exactly), and 6.21% for the Control, Cued, and Mode conditions, respectively (Fig. 6). Immediately apparent is that the new cue *completely* eliminated PCEs across all participants in this condition. This was a significant effect versus the control, $t(76) = 3.14$, $p = 0.002$, and Mode indicator group, $t(76) = 2.81$, $p = 0.006$. In comparison, the mode indicator failed to produce a reliable difference against the control group, $t(76) = 0.26$, $p = 0.80$.

In the simpler Medical task, mean errors at the postcompletion step were very low overall: 0.82%, 0%, and 1.99% for the Control, Cue and Mode indicator conditions, respectively. Again, none of the 12 participants in the Medical cued condition made a single PCE in all 17 of their trials. The same planned comparisons done on the Tactical task revealed no reliable differences across intervention.

Whether due to the number or nature of the steps, the mean postcompletion step completion time was drastically shorter in the Medical task compared to the Tactical, 4168 ms (Tactical) versus 1053 ms (Medical). An ANOVA showed this effect of task to be reliable, $F(1, 76) = 305.41$, $p < 0.001$. The overall effect of intervention on postcompletion step time was also reliable, $F(2, 76) = 4.32$, $p = 0.017$. However, the intervention by task interaction did not quite reach conventional criteria for rejection, $F(2, 76) = 2.86$, $p = 0.06$.

The average total number of errors (at any step) was also found to be higher for the Tactical task than the Medical: 0.67 in the Tactical versus 0.28 in the simpler Medical task, $F(1, 76) = 14.60$, $p < 0.001$, as expected. Differences across intervention were not reliable, $F(2, 76) = 2.24$, $p = 0.11$, although it should be noted that the total number of errors was slightly higher for both the cue and mode indicator conditions in both tasks. Participants showed no reliable differences in working memory performance regardless of task $F(1, 76) = 3.47$, $p = 0.07$ or intervention, $F(2, 76) = 1.09$, $p = 0.30$.

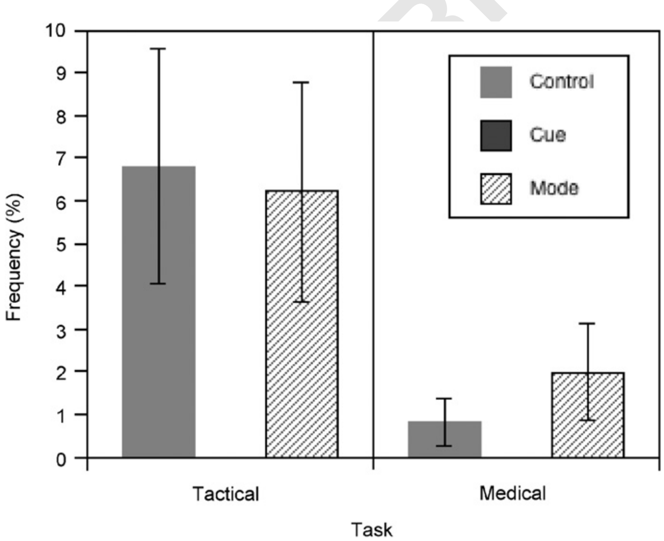

Fig. 6. PCE frequency by condition and task. Bars indicate standard error. Note that Cue is exactly 0% in both tasks.

3.7. Discussion of Experiment 2

As reported, all 16 participants in the cued condition of the Tactical task exhibited error-free performance at the postcompletion step. The new intervention was strikingly successful, completely eliminating errors at that step for those participants across all trials. Adding specificity to the cue with blinking and directional arrows (versus a simple red onset) made a major difference. In contrast, the control and mode indicator groups showed mean PCE frequencies between 6% and 7%. The unreliable effect of the mode indicator was somewhat unexpected, yet it supports the abundant recommendations in the human factors literature against introducing modes into a system (e.g., Norman, 1988). Since the initial graphical change occurs prior to the postcompletion step, the mode indicator places demands on prospective memory: the remembering and execution of delayed plans with no additional prompts at the time of intended retrieval (Guynn et al., 1998). According to Marsh and Hicks (1998), prospective memory performance decreases with increasing load on the executive resources, such as working memory. Hence, mode indicators, which are sometimes used as memory aids and reminders, are in fact susceptible to the same stressors they are meant to alleviate. In contrast, the effective cue appeared just-in-time, a necessary condition for effective reminders, according to the Altmann and Trafton (2002) model.

The Medical task failed to generate sufficient error rates to truly prove useful for comparing the effects of the interventions. There are several possible explanations for this. First, it was substantially shorter in length, taking participants nearly one quarter of the time taken to complete the Tactical task on average. Following the Byrne and Bovair (1997) account, the increased susceptibility for PCEs in the Tactical task may be explained by its greater demand on working memory. Nonetheless, the cue again completely eliminated errors in the Medical task (versus 0.82% and 1.99% for the control and mode indicator conditions, respectively), suggesting that its effect is robust across different tasks and interfaces.

4. A computational model

Cognitive architectures may help overcome the existing weaknesses in error prediction methods in two ways. First, they can provide a highly developed representation of both the human perceptual system and the external environment via either a model of the system or the actual system itself (e.g., Byrne and Kirlik, 2005). Recent inquiries into human error occurring in interactive tasks have frequently focused on the goal structure (e.g., Gray, 2000) for explanations. However, errors rooted in the perceptual mechanisms must also be addressed, since most computer-based tasks today are highly visual. Just as failing to see a stop sign or a red

Please cite this article as: Chung, P.H., Byrne, M.D., Cue effectiveness in mitigating postcompletion errors in a routine procedural.... International Journal of Human Computer Studies (2007), doi:10.1016/j.ijhcs.2007.09.001

light while driving may result in a car accident, similarly failing to see or misinterpreting some component of an interface can lead to serious consequences as well. To account for perceptual factors at the level of the computer interface, a systematic approach integrating all aspects of human cognition is required. The ACT-R architecture (Anderson et al., 2004) is an architecture supporting this kind of integrated view.

Although both Experiments 1 and 2 utilized two other manipulations apart from the cue (i.e., downstream cost and mode indicator), the present model focused on the cued and control conditions from Experiment 2. The main independent variable in this line of work has been the error intervention. Byrne and Bovair (1997) have demonstrated that the working memory load imposed by the digit span task in these experiments affects performance, leading to PCEs. The model takes into consideration their account, which is in keeping with general findings of task performance degradation under situations of high cognitive load (e.g., Ruffel-Smith, 1979). The model did not, however, account for skill learning occurring at the beginning of the experiments.

4.1. Error modeling traditions

Traditionally, *symbolic* systems have modeled consistent errors and errors of commission by assuming certain rules are missing or fail to apply (Van Lehn, 1989). Symbolic systems have more difficulty, however, with occasional slips or intrusions (Norman, 1981). On the other hand, *connectionist* models, with their holistic computation style, are intended to reproduce human-like errors and graceful degradation of performance under noise or component failures. However, scaling up to computer-based procedural tasks like that used here has generally not been attempted with connectionist systems. ACT-R, being a *hybrid* system, is better suited than traditional symbolic systems in this case, because activation of chunks is spread through an association network (e.g., Altmann and Trafton, 2002).

The architecture consists of a set of perceptual-motor modules (e.g., motor and visual), declarative memory, procedural memory, buffers, and a pattern matcher that work together to model human-like cognition. Declarative memory in ACT-R is represented as *chunks* of knowledge, whereas procedural memory consists of IF–THEN condition–action pairs termed *productions*. The pattern matcher detects chunks placed into the buffers by the modules and production rules are selected to fire serially. The subsymbolic computations of ACT-R helps guide the selection of rules and the internal operations of the modules. Learning and errors, for example, depend heavily on these subsymbolic processes. By taking advantage of ACT-R's subsymbolic construct of activation, potential errors at each step in the task structure can be produced. The increased working memory load, in its model representation, "steals" activation required to make a retrieval of the postcompletion subgoal.

Lebiére et al. (1994) have already demonstrated ACT-R's capacity to model graded human error, traditionally considered a domain restricted to connectionist models. Their study required participants to dual-task, as in the experiments reported here. Participants performed a high-level cognitive task of solving simple linear algebra problems while concurrently memorizing a digit span. The ACT-R model reproduced errors of omission by utilizing a cutoff on the latency of memory retrievals—retrievals failed if a chunk did not have sufficient activation. Because chunk activation is noisy, the model was able to generate a pattern of error quantitatively similar to participant data. With the current model, a similar method was utilized to generate errors of omission at the postcompletion step.

4.2. Model specifications

The model was built only to perform the Tactical task. Although it has some abstractions, particularly in the non-postcompletion steps, the focus was the postcompletion step itself. When the model reaches the postcompletion step (as with every other step in the procedure), it attempts to retrieve a declarative representation of the next thing to do, including information like the visual coordinates of the relevant button. Unlike at other steps in the procedure, there are two chunks which could be retrieved by the model here: the "correct" chunk (indicating the *Tracking* step) and the incorrect chunk (indicating the *Main Control* step). The incorrect step can be retrieved here in place of the correct one via ACT-R's partial matching mechanism. Because clicking Main Control is appropriate when the task is complete and the participants know that the target was destroyed (therefore completing the main task), these two chunks are given a non-zero similarity, meaning when one is requested, the other can sometimes be retrieved in its place. This is even more likely when memory load is high, and additional working memory load was simulated using dummy chunks representing state information placed in the goal buffer. These chunks, which can be considered as the digits in the digit span task, "steal" activation available for the retrieval of the chunk that produces the postcompletion action. With activation noise enabled, random retrievals of the incorrect (last) step were generated, leading to PCEs.

In the Cue condition, additional processes are available to the model. ACT-R's visual buffer is automatically filled with a representation of the cue when it appears due its sudden appearance on the screen (a native property of ACT-R's vision module), allowing the model to act immediately on that information. This in turn triggers a production that attempts to retrieve the postcompletion goal (that is, the *Tracking* step) in response to the red cue. This was consistent with instructions given in the training manuals, which asked participants to complete the post-completion step when the cue appeared. In this case the automatic capture of visual attention by the cue's sudden onset led to the application of additional procedural knowledge, and that knowledge eliminates PCEs, as found with the participants in Experiment 2.

The postcompletion frequency found in Experiment 1 for the control condition was 4.9%. However, this was only after participants who committed the error with over 50% frequency were removed, in adherence to the definition of PCEs as knowledge-based. Postcompletion frequency was at nearly 25% with their data included. In Experiment 2 the baseline postcompletion frequency was slightly lower, although this time participants with 25% + frequency were absent. Thus, as a compromise, 5–15% was the target PCE frequency for this model. This seemed reasonable considering that the cued and downstream cost groups' participants exhibited PCEs at around 10%.

We did not do extensive parameter-fitting for this model. Activation noise (ACT-R's parameter) was set to 0.2 (which is in a fairly conventional range for ACT-R models), the retrieval threshold was set to −0.6 (meaning that nearly all retrievals succeed in returning something), and the similarity between the two-step chunks referred to earlier (that is, the Tracking and Main Control steps) was set to −0.8. All other parameters were left at system defaults. Running the model for 200 trials generated a 14.1% PCE frequency in the Control condition. In the Cued condition the model responded correctly every time (0% PCE frequency), as we found in Experiment 2. This was dependent on several factors: (1) visual attention being "captured" by the novel appearance of the cue and (2) successful retrieval of the correct knowledge (imparted at training) about what to do when the cue was detected. The model was thus able to demonstrate that if a cue is salient and designed with sufficient specificity (Hollnagel, 1993), it will lead to the successful retrieval of appropriate knowledge and therefore correct performance.

5. General discussion

These findings illustrate the basic challenges to error prediction and mitigation in highly visual interactive tasks. Using extant error identification methods, which tend to focus on the goal structure, and design guidelines, which are faced with a problem of subjectivity, it would be difficult to predict the disparity found between the two cues of Experiments 1 and 2 or even the mode indicator. The difference in effectiveness between the two visual cues would be especially challenging to predict, as the results suggest cue effectiveness was highly dependant on participant interpretation of the their visual features: both cues appeared in proximity to the button, onset just-in-time, and utilized color and novel appearance. Accurate knowledge, represented in the model by chunks of information or as rules learned at training in Experiments 1 and 2, is critical for a reminder to elicit a correct response (goal retrieval).

In light of this, one might argue that the cue in Experiment 2 was successful, whereas the cue in Experiment 1 was not, due to changes in the training procedures, which promoted better recall of the tasks at testing. Although better training may account for some of the difference, it should be noted that PCE frequencies in the control condition were similar in both experiments, and the more rigorous training procedures were shared across all conditions in Experiment 2. Likewise, the elimination of PCEs by the cue in Experiment 2 does not seem completely attributable to the addition of blinking (versus an onset), as capturing visual attention alone does not lead to a correct response with the model. Nevertheless, it is likely that this increase in visual salience was a contributing factor, as the Altmann and Trafton (2002) goal-activation model suggests. The very salient visual cue in Experiment 2 effectively primed the postcompletion goal. In contrast, the mode indicator, as an example, was unsuccessful as a cue, perhaps due to masking by the intermediate goal of firing the phaser.

The high visual specificity (Hollnagel, 1993) of the arrows and the easier-to-recall, non-experiment-specific knowledge they leveraged for interpretation were also contributing factors. The importance of specificity in graphical user interface design is largely recognized in the applied world, as its presence in some form on many usability guidelines and checklists indicates. It is an inherently variable property, however, since there is a contingency on the person's preexisting knowledge. Our model contained a specific rule to carry out the appropriate action at the postcompletion step in response to the cue's appearance. Without this, the cue would not have caused the model to act correctly. Thus, the cue in Experiment 2 may have been so effective in part because it leveraged preexisting, and therefore easily retrieved, knowledge that most people hold today, concerning its two defining visual features: arrow-like shapes and blinking. This is in contrast to the cue in Experiment 1, which was not as distinctive in its visual features (a round dot), and participants had difficulty recalling its meaning from training. From an applied perspective, this shows that it is imperative to consider what knowledge a person possesses, when trying to predict how they will interpret a cue. Placing blinking arrows or other novel cues on an interface may lead to different and unexpected outcomes, depending on a person's familiarity with computers, cultural background, education, etc.

The key contribution of the model to this effort is the lengths required to get the model to circumvent the error. The model essentially requires three things: a cue which will be noticed (salient), a cue which is just-in-time, and a cue which makes contact with knowledge held by the model, that is, a cue which actually cues something in particular. This advice is hardly novel; however, the model provides a clear explanation of why all three pieces are necessary; omitting any one of those features would certainly render the cue ineffective for the model. This is an empirically testable prediction which can and should be evaluated in future research; are cues with only two of the three critical features effective? The red dot in Experiment 1 was not, and it has two (salience and just-in-time). Are the other possible pairs equally ineffective? The model indicates they should be; we hope to test this in future research.

The Medical task in Experiment 2 plainly illustrated the difficulty of designing a task that can elicit sufficient error rates from participants to study human error in the laboratory. Although it has a task step after the main goal of the task, this supposed postcompletion step failed to generate significant error rates. For those conducting expert evaluations of systems (e.g., heuristic evaluation), this finding reveals the difficulty of identifying potential error-inducing steps or features on an interface. Simply conducting a task analysis to examine goal structures or attempting to classify potential errors with a list of usability guidelines, for instance, will not always suffice.

Interfaces must be designed to both reduce the frequency of human error and mitigate their effects, particularly in safety critical domains. This work was an initial step to extend our understanding of how visual cues may be used effectively to improve performance in interactive tasks. While human factors guidelines (e.g., US Department of Defense, 1999) and research (e.g., Wang et al., 1994) are available, they only suggest the appropriate visual properties of proper cues and reminders and cannot predict which will be effective in a specific situation. Most existing methods of task representation and error identification are also deficient, as they fail to account for the perceptual factors relevant to human performance in interactive tasks (see also Byrne et al., 2004). As our work demonstrates, understanding errors at the level of the interface requires consideration of all aspects of human cognition. For this reason, cognitive modeling has been suggested as a solution Q7 (see Byrne, 2003; Gray, 2004). Those studying eye movements in reading (e.g., Rayner, 1998) have converted massive collections of data into models that may be iteratively tested and validated. Pursuing a similar approach here will therefore necessitate further empirical studies.

Q8 Uncited references

Anderson and Lebiére, 1998; Kirwan and Ainsworth, 1992.

References

Altmann, E.M., Trafton, J.G., 1999. Memory for goals: an architectural perspective. In: Proceedings of the 21st Annual Conference of the Cognitive Science Society, Erlbaum, Hillsdale, NJ, pp. 19–24.

Altmann, E.M., Trafton, J.G., 2002. Memory for goals: an activation-based model. Cognitive Science 26, 39–83.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebière, C., Qin, Y., 2004. An integrated theory of the mind. Psychological Review 111, 1036–1060.

Anderson, J.R., Lebiére, C., 1998. The Atomic Components of Thought. Erlbaum, Mahway, NJ.

Annett, J., Duncan, K.D., 1967. Task analysis and training design. Journal of Occupational Psychology 41, 211–221.

Baber, C., 1996. Repertory grid and its application to product evaluation. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I. (Eds.), Usability Evaluation in Industry. Taylor & Francis, London, pp. 157–166.

Baber, C., Stanton, N.A., 1994. Task analysis for error identification. Ergonomics 37 (11), 1923–1942.

Byrne, M.D., Bovair, S., 1997. A working memory model of a common procedural error. Cognitive Science 21 (1), 31–61.

Byrne, M.D., Davis, E.M., 2006. Task structure and postcompletion error in the execution of a routine procedure. Human Factors 48, 627–638.

Byrne, M.D., Kirlik, A., 2005. Using computational cognitive modeling to diagnose possible sources of aviation error. International Journal of Aviation Psychology 15, 135–155.

Byrne, M.D., Maurier, D., Fick, C.S., Chung, P.H., 2004. Routine procedural isomorphs and cognitive control structures. In: Schunn, C.D., Lovett, M.C., Lebiere, C., Munro, P. (Eds.), Proceedings of the Sixth International Conference on Cognitive Modeling, pp. 52–57.

Chung, P.H., Zhang, J., Johnson, T.R., Patel, V.L., 2003. An extended hierarchical task analysis for error prediction in medical devices. In: Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium. American Medical Informatics Association, Washington, DC.

Gray, W.D., 2000. The nature and processing of errors in interactive behavior. Cognitive Science 24 (2), 205–248.

Gray, W.D., 2004. Errors in interactive behavior. In: Bainbridge, W.S. (Ed.), Encyclopedia of Human–Computer Interaction. Berkshire Publishing Group, pp. 230–235.

Green, M., Senders, J., 2004. Human error in road accidents. Visual Expert. Retrieved March 27, 2005 from 〈http://www.visualexpert.com/Resources/roadaccidents.html#Green91〉.

Guynn, M.J., McDaniel, M.A., Einstein, G.O., 1998. Prospective memory: when reminders fail. Memory & Cognition 26 (2), 287–298.

Hollnagel, E., 1993. Human Reliability Analysis, Context and Control. Academic Press, London.

Hollnagel, E., 1998. Cognitive Reliability and Error Analysis Method (CREAM). Elsevier Science Ltd., Oxford.

John, B.E., Kieras, D.E., 1996. The GOMS family of user interface analysis techniques: comparison and contrast. ACM Transactions on Computer–Human Interaction 3 (4), 320–351.

Jonides, J., 1981. Voluntary versus automatic control over the mind's eye's movement. In: Long, J.B., Baddeley, A.D. (Eds.), Attention and Performance IX. Erlbaum, Hillsdale, NJ, pp. 187–203.

Kieras, D.E., Wood, S.D., Meyer, D.E., 1997. Predictive engineering models based on the EPIC architecture for multimodal high-performance human–computer interaction task. ACM Transactions on Computer–Human Interaction 4, 230–275.

Kirwan, B., 1992. Human error identification in human reliability assessment. Applied Ergonomics 23 (5), 299–381.

Kirwan, B., Ainsworth, L.K. (Eds.), 1992. A Guide to Task Analysis. Taylor & Francis, London, UK.

Lebiére, C., Anderson, J.R., Reder, L.M., 1994. Error modeling in the ACT-R production system. In: Proceedings of the 16th Annual Meeting of the Cognitive Science Society, Erlbaum, Hillsdale, NJ, pp. 555–559.

Marsh, R.L., Hicks, J.L., 1998. Event-based prospective memory and executive control of working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 24 (2), 336–349.

Monk, A., 1986. Mode errors: a user-centered analysis and some preventative measures using keying-contingent sound. International Journal of Man–Machine Studies 24, 313–327.

Most, S.B., Simons, D.J., Scholl, B.J., Jimenez, R., Clifford, E., Chabris, C.F., 2000. How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. Psychological Science 12, 9–17.

Please cite this article as: Chung, P.H., Byrne, M.D., Cue effectiveness in mitigating postcompletion errors in a routine procedural.... International Journal of Human Computer Studies (2007), doi:10.1016/j.ijhcs.2007.09.001

Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall, Inc., Englewood Cliffs, NJ.
Norman, D.A., 1981. Categorization of action slips. Psychological Review 88, 1–15.
Norman, D.A., 1988. The Design of Everyday Things. Basic Books, New York.
Polson, P.G., Lewis, C.H., 1990. Theory-based design for easily learned interfaces. Human–Computer Interaction 5, 191–220.
Rasmussen, J., 1982. Human errors: a taxonomy for describing human malfunction in industrial installations. Journal of Occupational Accidents 4, 311–335.
Rasmussen, J., 1987. The definition of human error and a taxonomy for technical system design. In: Rasmussen, J., Duncan, K., Leplat, J. (Eds.), New Technology and Human Error. Wiley, Chichester, pp. 53–62.
Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124, 372–422.
Reason, J., 1990. Human Error. Cambridge University Press, Cambridge, UK.
Reason, J., 2002. Combating omission errors through task analysis and good reminders. Quality and Safety in Healthcare 11, 40–44.
Remington, R.W., Johnston, J.C., Yantis, S., 1992. Involuntary attentional capture by abrupt onsets. Perception & Psychophysics 51, 279–290.
Rieman, J., Young, R.M., Howes, A., 1996. A dual-space model of iteratively deepening exploratory learning. International Journal of Human–Computer Studies 44, 743–775.
Ruffel-Smith, H.P., 1979. A Simulator Study of the Interaction of Pilot Workload with Errors, Vigilance, and Decisions (Technical Memo 78482). NASA Ames Research Center, Moffett Field, CA.
Serig, E.M., 2001. Evaluating organizational response to a cognitive problem: a human factors approach. Doctoral Dissertation, Rice University, Houston, TX.
Stanton, N.A., 2004. Human error identification in human–computer interaction. In: Jacko, J. (Ed.), Handbook of Human Factors and Ergonomics, second ed. Wiley, New York.
Sutcliffe, A.G., 1995. Human–Computer Interface Design. Macmillan Press Ltd., London.
Swain, A.D., Guttman, H.E., 1983. Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications. NUREG/CR-1278, Washington, DC.
Treisman, A., 1986. Features and objects in visual processing. Scientific American 254, 114–124.
US Department of Defense, 1999. Department of Defense Design Criteria Standard: Human Engineering (MIL-STD-1472F). Government Printing Office, Washington, DC.
Van Lehn, K., 1989. Problem solving and cognitive skill acquisition. In: Posner, M.I. (Ed.), Foundations of Cognitive Science. MIT Press, Cambridge, MA.
Wallace, D.F., Huffman, D., 1990. Use of a visual cue to reduce errors in exiting a crash-bar type door. In: Proceedings of the Human Factors Society 34th Annual Meeting, pp. 567–569.
Wang, Q., Cavanagh, P., Green, M., 1994. Familiarity and pop-out in visual search. Perception & Psychophysics 56, 495–500.
Wood, S.D., 2000. Extending GOMS to human error and applying it to error-tolerant design. Doctoral Dissertation, University of Michigan.
Yantis, S., Jonides, J., 1988. Uniqueness of abrupt visual onset in capturing attention. Perception & Psychophysics 43 (4), 346–354.
Young, R.M., Barnard, P., Simon, T., Whittington, J., 1989. How would your favorite user model cope with these scenarios? ACM SIGCHI Bulletin 20, 51–55.
Zhang, J., Norman, D.A., 1994. Representations in distributed cognitive tasks. Cognitive Science 18, 87–122.