Visual Cues to Reduce Errors in a Routine Procedural Task

Phillip H. Chung (Pchung@Rice.Edu)
Department of Psychology, 6100 Main Street
Houston, TX 77005 USA

Michael D. Byrne (Byrne@Rice.Edu)
Department of Psychology, 6100 Main Street
Houston, TX 77005 USA

Abstract

This paper reports an experiment designed to evaluate the efficacy of visual cues as error "interventions" in computer-based routine procedural tasks. Using two separate tasks with a well-documented error prone step, the effects of several visual cues were compared. The findings provide support for goal selection driven by environmental cues in routine procedural tasks. The importance of cue timing and movement and meaningfulness characteristics, particularly in dynamic tasks with external pressures, is demonstrated.

Introduction

History and Motivation

With the introduction of automation and computers, an outstanding arena for human error has been established. Subsequently, much effort has been given to categorize errors occurring in such situations, yet for the most part understanding of this very human phenomenon remains fairly nebulous. As John and Kieras (1996) stated in the 90s, "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…even the theoretical analysis of human error is still in its infancy."

Over the years, several elaborate models and taxonomies of human error have been developed for the purpose of qualitative diagnosis (e.g., Reason, 1990). Although useful for post hoc explanations, the predictive power of these is quite limited. Results from controlled studies with various error "interventions" may extend our understanding of why such cognitive errors arise, helping us not only evaluate (post hoc) but also design (predict) safer machines. Application areas to benefit abound, from aerospace to medicine.

Postcompletion Errors

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, 1994) omission error, postcompletion error. Postcompletion errors can be broadly 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 & Bovair, 1997, p. 32). With this particular class of error, the actor possesses the correct knowledge necessary to execute the task, usually performed frequently and correctly. Yet, for even operators highly familiar with the task, the isolation of a postcompletion step within the task structure makes omissions there not unlikely. This is particularly true when the actor is further affected by external factors such as a working memory load and/or fatigue, as well as internal tendencies such as hillclimbing (Gray, 2000; Polson & Lewis, 1990).

Some commonplace examples include forgetting to remove the original after making a photocopy, leaving a card in the ATM after withdrawing cash, and failing to replace the gas cap after filling up a car. 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 postcompletion errors the actor omits a specific subgoal rather than forgetting what to do altogether (the overlying 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 postcompletion errors to be the most common type of omission in that task (Figure 1).

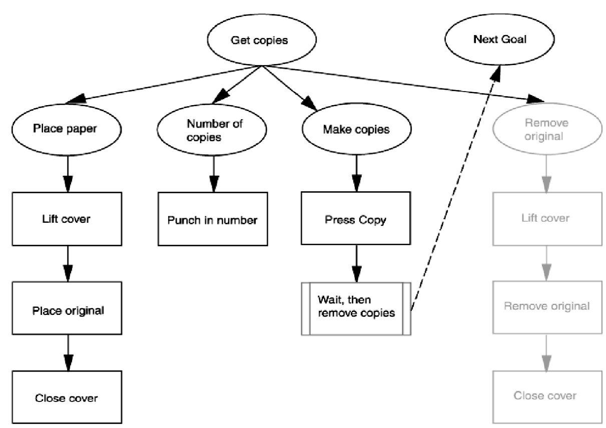

Figure 1: Photocopy task structure.

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*.

Hierarchical Control Structures and Goal Management

Many of the assumptions behind the current theory of postcompletion error reside on the foundational 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 subtasks as well as within the larger task, in keeping with the idea of a hierarchical task structure. Cognitive modeling work by Kieras, Wood, and Meyer (1997) has also provided strong evidence to suggest that even well practiced experts, such as telephone assistance operators, do not abandon such task hierarchies.

As Altmann and Trafton (1999) propose, the 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, their activation-based model of memory for goals (MAGS) 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 & Young, 1996; internal and external representations, Zhang & 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 (e.g., Figure 1, in grey) with perceptual heuristics acting as a substitute for the stack-native last in, first out rule. This model makes several predictions about postcompletion errors and the characteristics of a successful cue:

1. Any *salient* cue (e.g., a loud beep) should be sufficient to prime a postcompletion action (suspended goal).
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 (masked by other goals).
4. Just-in-time priming from environmental cues are the *only* reliable reminder.

Previous Study

A previous experiment (Chung, 2004) examined the effects of a simple visual cue (red singleton onset) and a downstream error cost (in the form of a resultant mode error) on postcompletion error commission. Although neither was found to cause significant change in reaction times or error commission at the postcompletion step, the results did generate some valuable implications. First, the fact that the visual cue did not significantly reduce the number of postcompletion errors committed by the participants suggested that the cue lacked sufficient salience to prime the suspended postcompletion goal. While the sudden onset of a large red dot (against a black and white console) next to the button that needed to be pressed seemed informative enough, omissions were made regardless. Undoubtedly, participants had sufficient understanding of the task and scenario, since they could not proceed to testing without completing extensive training.

Neither did the downstream error cost (resulting in a mode error) bring about a significant change in behavior at the postcompletion step for participants. While it was expected that the visual cue would be more effective between the two treatments, it was also hypothesized that the downstream error cost would cause a change in behavior. However, the number of postcompletion errors in this condition was not significantly different from the control group, suggesting that it did not provide any significant advantage (or disadvantage) for participants. This perhaps follows findings by Serig (2001) that demonstrated error commission to be relatively independent of negative or positive feedback.

Experiment

Two tasks

Along with the original Tactical task introduced in the study by Byrne and Bovair (1997), a separate Medical task was added to distinguish the effects of the interventions. The task also included a potential postcompletion step (determined via task analysis) where the interventions could be implemented, as with the Tactical task.

Intervention Implementation

Two different interventions were introduced in this experiment: an enhanced visual cue (see Figure 2) and a mode indicator (Figure 3). These were developed specifically to address issues brought up by the findings in previous work (Chung, 2004), help pinpoint the characteristics of a successful intervention, and evaluate the predictions of MAGS (Altmann & Trafton, 1999), should one or both have a significant effect.

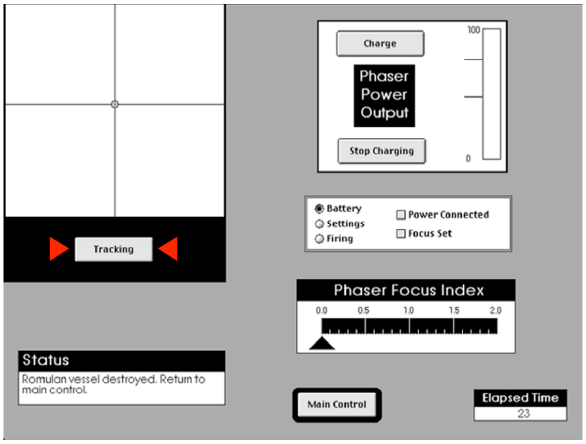

Figure 2: Tactical interface with cue (two arrows).

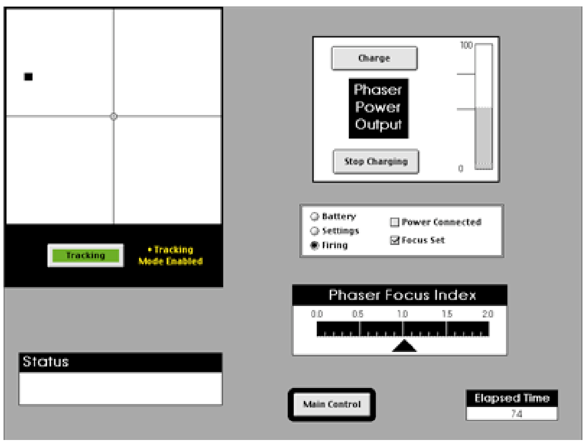

Figure 3: Mode indicator (highlighting on "Tracking" and "Tracking Mode Enabled" message).

Intervention Attributes

In a study by Monk (1986), auditory cues were employed to drastically reduce the occurrence of mode errors. Monk (1986) notes that display changes are similarly effective when the person is required to look at the relevant parts of the display at the appropriate moment in the dialogue. The related automatic processing of novel peripheral cues regardless of whether or not they are informative has been well documented (e.g., Remington, Johnston, & Yantis, 1992). Colorful visual cues are known to be effective and necessary in guiding individuals to select points of activity, such as the push plate on a door (Norman, 1988).

Research (Sutcliffe, 1995) and real-world practice indicates that the visual attributes most effective for attracting attention (warnings and indicators) on a computer interface, in order, are as follows: movement (blinking or change of position), shape and size (character font, shape of symbols, text size, size of symbols), color, brightness, shading and texture (different texture or pattern), and surroundings (borders, background color). Sutcliffe (1995) advises that these should be applied sparingly, however, as the presence of many conflicting stimuli can essentially dull their individual effectiveness. Red, green, and yellow are recommended as the optimal colors for status indicators, each corresponding to its meaning on a traffic light. To draw attention, white, yellow, and red are most effective, although yellow offers the best visibility.

Based on these recommendations and the characteristics of the failed cue from previous work (Chung, 2004), alternating red and yellow blinking arrows (Figure 2) were used for the visual cue. As per the Altmann and Trafton's (1999) predictions, the cue appeared "just-in-time" at the postcompletion step. In contrast, the mode indicator consisted of green and yellow highlighting on the "Tracking" button along with other contextual indication appearing *before* the postcompletion step. When combined with the given if-then rule at training (i.e., "If you see a mode indicator light and message, the system is on"), the mode indicator was expected to prime the corresponding goal (the postcompletion step) of turning off the Tracking system. Once the participant finishes the intermediate steps and hits the "Tracking" button a second time, the indicator disappears to indicate that the Tracking mode has ended. Exact placement of the interventions was determined through pilot studies.

Method

Participants

Ninety-one undergraduate and graduate students from Rice University aged 18 to 35 participated for course credit in a psychology course and additional cash prizes ranging from $10 to $40.

Materials

Materials consisted of a short paper-based quiz, paper-based instruction manuals for each of the four tasks (Tactical, Medical, and two filler tasks), Apple iMac computers running the Tactical and Medical applications written in Macintosh Common Lisp, stereo headphones, and a web-based general questionnaire for demographic information.

Design

This study used a two-factor between participants design with two independent variables, task and intervention. Task consisted of two conditions (Tactical and Medical) to compare the effectiveness of the interventions across task and interface. Intervention consisted of three conditions: control (no intervention), visual cue (alternating red and yellow blinking arrows), and mode indicator (mode indication for the system state change). Participants were randomly assigned to one of the six groups.

The primary dependent measure was the number of postcompletion errors made during the Tactical and Medical tasks. Other dependent measures of interest included the overall number of errors per task and performance on a concurrent working memory task (described in Byrne & Bovair, 1997).

Procedure

Participants were run in two sessions spaced two days apart. The first session served as a training session using written documentation for each of the tasks. The major steps of the two target tasks (Tactical or Medical) are outlined in Table 1 and essentially consist of a series of key presses and mouse clicks and movements. Order of training and group assignment were randomized for every participant. Once participants successfully completed the training trial with the manual and logged three subsequent error-free trials, they were allowed to move on to the next task.

Table 1: Steps in each task.

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 (1 substep)	

Errors resulted in warning beeps and messages and participants were returned to the main control to restart the task. This was to prevent participants from completing training without having gone through each of the tasks at least four times with all steps done correctly and completely. When training was complete, they were reminded that they would compete for prizes in two days and given a short quiz to ensure that they had accurate working knowledge of the tasks.

The second session consisted of the test trials for both the Tactical and Medical tasks. In random order, participants completed seventeen trials of their assigned postcompletion task (Tactical or Medical) and eleven trials for each of the two filler tasks, for a total of thirty-nine trials on the test day. At testing, the experiment program emitted beeps on error commission to warn individuals but did not immediately return them to the main control or provide warning messages, as in training. Participants were encouraged to work both accurately and quickly by means of a scoring system, prizes, and an onscreen timer. A three-letter span auditory working memory task was introduced in all task conditions at testing.

Results

Data from 82 of the original 91 participants was used in the final analysis. The primary reason for the loss of data was participant failure to show up at their assigned testing date. Only one participant was removed as an outlier (Medical, cue condition). Groups broke down as shown in Table 2 below.

Table 2: Participants per group.

Condition	Control	Cue	Mode
Tactical	14	16	13
Medical	14	12	13

Postcompletion Error Frequency

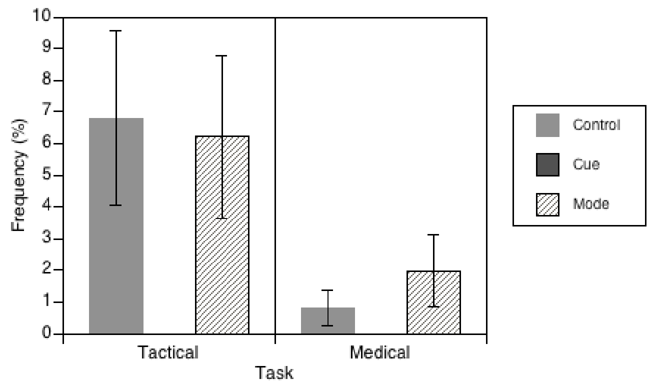

Figure 4: Postcompletion error frequency (std. error bars). Cue condition is 0% for both Tactical and Medical tasks.

Our primary measure of interest was the frequency of errors at the postcompletion step (out of seventeen trials) in both tasks. This is the step immediately following completion of the main task goal. For the Tactical task, mean postcompletion error frequencies were 6.81%, 0%, and 6.21% for the control, cued, and mode indicator conditions, respectively (Figure 4).

Analysis of variance showed the effect of intervention to be reliable, $F(2, 76) = 4.061$, $p = .021$, but not the interaction of intervention by task, $F(2, 76) = 1.86$, $p = .162$. Planned comparisons confirmed our hypothesis, as participants made significantly less errors in the cued condition versus the control, $t(76) = 3.14$, p = .002, and even versus the mode indicator group, $t(76) = 2.81$, $p = .006$. In comparison, the mode indicator failed to produce reliable differences with the control group, $t(76) = .263$, $p = .793$.

In the simpler Medical task, mean errors at the postcompletion step were very low: 0.82%, 0%, and 1.99% for the control, cue and mode indicator conditions, respectively. Again, none of the twelve participants in the Medical cued condition made a single postcompletion error in all seventeen of their trials. The same planned comparisons done on the Tactical task revealed no reliable differences across intervention and task.

Total Errors

The average number of total errors (out of all possible steps) was found to be higher for the Tactical task than the

Medical: 0.67 in the Tactical versus 0.28 in the simpler Medical task, $F(1, 76) = 14.60$, $p < .001$. Differences across intervention were not reliable, $F(2, 76) = 2.24$, $p = .113$, 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.

Working Memory Task

Participants showed no reliable differences in working memory task performance regardless of task $F(1, 76) = 3.47$, $p = .07$ or intervention, $F(2, 76) = 1.09$, $p = .342$.

Discussion

Our findings generally corroborate our hypothesis for the visual cue. As reported, all sixteen participants in the just-in-time cue condition of the Tactical task exhibited error-free performance at the postcompletion step on all seventeen of their trials. In contrast, the control and mode indicator groups showed mean postcompletion error frequencies between six and seven percent. Given the lack of reliable differences across intervention for overall error rates and performance on the working memory task, there seems to be no reason not to attribute the difference in postcompletion error frequency to the success of the intervention.

Nevertheless, our expectations for the mode indicator were not met. While the just-in-time cue reduced the postcompletion error mean to nil, the mode indicator had hardly any effect relative to the control. This was despite the fact that all participants were given equal training and the mode indicator was made as large, if not larger, than the flashing arrows in the cued condition. With the additional novel appearance of the crosshairs (Tactical) and display information (Medical), the state change should have been noticeable. Thus, its failure does not seem attributable to a lack of knowledge or relative visibility.

The Medical task was ineffective as a parallel of the Tactical task, perhaps primarily due to its substantially shorter length (see Table 1). It took participants nearly one quarter of the time taken to finish the Tactical task and simply failed to generate sufficient error rates to prove useful for comparing the effects of the interventions. However, it is notable that the visual cue also completely eliminated postcompletion errors in the Medical task as in the Tactical task.

Validation of MAGS

Our findings generally fell in line with the predictions of Altmann and Trafton (1999), given that the cue in the previous experiment (Chung, 2004) failed from a lack of salience. As claimed, the new ("just-in-time") cue was sufficient to prime the postcompletion step, making it unnecessary to place the postcompletion action on the critical path. Moreover, the mode indicator (a state change "at the start") did not sufficiently prime the postcompletion step that followed. It was likely "masked" by the intermediate steps or goals, as they explained.

These results support the idea of goal selection in tasks as a product of environmental cues. Hence, it follows that postcompletion errors are often generated by "false completion signals" (Reason, 2002), such as the emergence of the copy in the photocopy task. Likewise, the visual cue implemented here, in the form of two blinking arrows, was able to prime the postcompletion goal sufficiently to be correctly retrieved.

Conclusion

Several guidelines for the design of safe interfaces used in routine procedural tasks can be gleaned from this work. Interventions should be made to appear "just-in-time," as to reduce demands on memory. Asynchronous cues like the mode indicator place their own demands on memory, since there are steps intermediate to the step they are meant to prime. Even negative feedback (*after* the postcompletion step), as a downstream error cost (Chung, 2004) or as a reprimand from an "overseer" (Serig, 2001), has demonstrated no reliable reduction in the frequency of errors. Postcompletion errors cannot simply be willed away.

Also, it seems that movement and/or shape (meaning) are strong determinates of whether or not a cue is attended to (Sutcliffe, 1995). A cue (a simple red singleton) used in previous work (Chung, 2004) appeared at the same exact location as the just-in-time cue in this experiment, yet generated no significant reduction in error frequency. The mode indicator, which relied on static contextual cues, also had no reliable effect. It was made static (as in most real-world applications) since the nature of such cues is that there are intermediate steps between their onset and the step they are meant to prime. Blinking would unnecessarily attract visual attention to an inactive control.

Implementing a successful error intervention may not, however, be so simple as merely adding visual cues with these properties to the interface. Differences in task (e.g., length) and interface (e.g., background color) characteristics also attenuate the effectiveness of these cues, as demonstrated by the Medical task. Additionally, the fact that our participants had explicit training on the meaning of the cue must be considered. Simply placing blinking arrows or other novel cues on the interface would affect naïve users differently from those who had been trained.

The failure of a singleton onset (Chung, 2004) and subsequent success of two blinking arrows in this experiment may at least partially be explained by the speed at which our visual attention shifts in procedural tasks with medium to high level of skill and external pressures. Hence, while the cue used in the previous experiment also appeared in temporal conjunction with the completion of the previous step and in spatial proximity to the targeting window, it was overlooked. In contrast, the successful blinking cue continued to generate attention-capturing movement until the postcompletion step was satisfied. Moreover, it offered immediate information (arrows pointing to the correct button) about its meaning.

Hence, these findings come as further empirical evidence for cue-driven goal selection in procedural tasks. More specifically, they highlight the importance of cue timing and the visual properties of movement and meaningfulness. Follow up inquiry is underway to determine the individual strengths of these properties. Such data will be vital for any truly predictive theory of human error.

Acknowledgments

We would like to acknowledge the support of the Office of Naval Research under grant number N00014-03-1-0094. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ONR, the U.S. Government, or any other organization.

References

Altmann, E. M. & Trafton, J. G. (1999). Memory for goals: An architectural perspective. *Proceedings of the twenty-first annual conference of* the Cognitive Science Society (pp. 19-24). Hillsdale, NJ: Erlbaum.

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

Chung, P. H. (2004). Visual cues to reduce error in computer-based procedural tasks. Masters thesis, Rice University, Houston, TX.

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

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.

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

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.

Reason, J. (1990). Human Error. Cambridge, UK: Cambridge University Press.

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.

Serig, E. M. (2001). Evaluating Organizational Response to a Cognitive Problem: A Human Factors Approach. Doctoral dissertation, Rice University, Houston, TX.

Sutcliffe, A.G. (1995). Human-Computer Interface Design. London: Macmillan Press Ltd.

Young, R. M. (1994). The unselected window scenario: analysis based on the SOAR cognitive architecture. *Proceedings of the Computer Human Interaction (CHI) Conference*, 1994.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18,* 87-122.