Task Structure and Postcompletion Error in the Execution of a Routine Procedure

Michael D. Byrne and **Elizabeth M. Davis,** Rice University, Houston, Texas

Objective: To replicate a successful laboratory slip-class error paradigm and, more importantly, to further understand the underlying causes of errors made in that paradigm. **Background:** Routine procedural errors are facts of everyday life but have received limited controlled empirical study, despite the sometimes severe consequences associated with such errors. This research concerns one such error, postcompletion error (M. D. Byrne & S. Bovair, 1997), which is a lapse that occurs after the main goal of a task has been satisfied. **Method:** In the two experiments conducted, participants were trained to criterion on a routine procedural task and were then brought back to the lab for a later session or sessions in which performance on task execution was measured. In the second experiment, a variety of motivational manipulations, retraining, and task redesign were compared. **Results:** Experiment 1 demonstrated a substantial reduction of error rate generated by a simple design change (alteration of when feedback about goal completion occurred). Furthermore, the reduction in error rate came with no penalty in terms of overall speed of performance. Experiment 2 showed that this more appropriate design is superior to motivationally oriented interventions, retraining, and even midtask redesign. As in Experiment 1, Experiment 2 revealed no speed-accuracy tradeoff. **Conclusion:** These experiments provide evidence that controlled laboratory studies of slip-class errors can be meaningful and highlight the centrality of cognitive factors (particularly goal structure) in such errors. **Application:** Potential applications include design of interfaces and their related procedures as well as error-mitigation techniques.

INTRODUCTION

Many tasks, ranging from those found in everyday life to massively engineered safety-critical systems, are what Ohlsson (1996) called “sequential choice tasks” or Card, Moran, and Newell (1983) termed “routine procedural tasks.” These are tasks performed to produce a product or bring about a specific state of affairs. The resolution to these tasks requires that an individual follow a certain sequence of actions. For each action, there may be several potential paths an individual might follow. Generally, however, only one or a few of these pathways are valid. An error in a sequential choice task typically manifests itself as the selection of an inappropriate choice for a particular action. However, failures in the execution of routine procedures are key components of errors in numerous safety-critical domains, such as aviation (e.g., Sarter & Alexander, 2000) and medicine (e.g., Xiao & Mackenzie, 1995). In addition, errors in such tasks have been discussed at some length by Norman (1981) and Gray (2000). Obviously, there are many other kinds of tasks in which errors are crucial, but this clearly represents a significant error class.

Postcompletion errors are an example of a type of error that can occur during a sequential choice task. These errors occur when the sequential choice 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). Some examples of postcompletion error are failing to collect the original from the photocopier glass, leaving one’s bank card in an automated teller machine after collecting the withdrawal, and forgetting to switch a clock alarm to “on” after setting the alarm time. In terms of a sequential choice

Address correspondence to Michael D. Byrne, Department of Psychology, Rice University, 6100 Main St., MS-25, Houston, TX 77005; byrne@acm.org. ***HUMAN FACTORS,*** Vol. 48, No. 4, Winter 2006, pp. 627–638. Copyright © 2006, Human Factors and Ergonomics Society. All rights reserved.

task, these errors occur when the incorrect action is selected at the end of the task procedure. For example, with the photocopier, the user has at least three choices: (a) collect the copies and leave the copy center (a postcompletion error), (b) collect the original *before* collecting the copies and leaving the copy center (no error), and (c) collect the copies and then the original before leaving the center (no error).

Byrne and Bovair (1997) attributed postcompletion errors to “goal forgetting” caused by an excessive working memory load. They demonstrated that students playing a designed computer game were more likely to commit postcompletion errors as working memory demands increased. This working memory load can lead a person into selecting the inappropriate action at the end of the task sequence if the main goal of the task procedure has been satisfied. Because the primary goal has been satisfied (e.g., make copies), the load on working memory (e.g., “I’m late for my meeting and still need to do x, y, and z”) may make it more probable that the inappropriate action (e.g., collect the copies and go) is activated than the “special” instruction corresponding to the postcompletion action (e.g., collect the original).

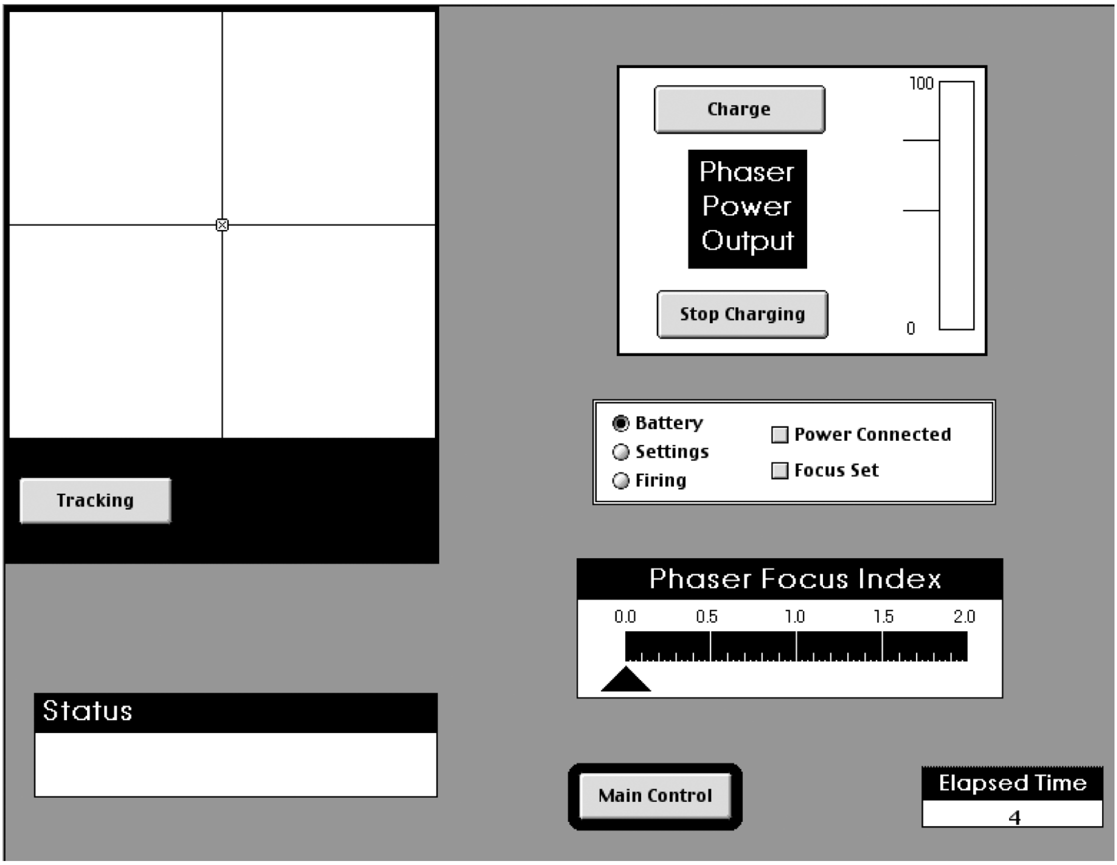
EXPERIMENTS

Two experiments were conducted to assess task performance on a routine sequential choice task. The first experiment was designed to replicate and extend the findings of Byrne and Bovair (1997) regarding the occurrence of postcompletion error during routine task performance. The second experiment was designed to demonstrate and assess the potential effects of “blame-and-train” style methods on task performance of a routine sequential choice task.

The phaser task from Byrne and Bovair (1997) was used to assess the occurrence of postcompletion error. To complete the task, participants executed a sequence of actions that culminated in destroying a target. There were two versions of this task: a control version and a postcompletion version. Figure 1 outlines the sequence of actions followed by participants. Every action that requires input to the system (e.g., button clicks and adjustment of controls using the mouse) is counted as an action step in the procedure. Places where judgments are made (e.g., determining whether the target was destroyed) do not count as action steps in the sequential procedure because they produce no overt response. In the control version, after firing at the target, participants had to click the *tracking* button to determine whether or not they had destroyed the target. If the target was destroyed, the task goal was met and participants ended the trial by clicking the *main control* button. In the postcompletion version, after firing at the target, participants received immediate feedback on whether or not the target had been destroyed. If the target was destroyed, participants had to turn off the tracking mechanism (click the *tracking* button) before terminating the trial by clicking the *main control* button. The “turn off tracking” step is the postcompletion step. Note that this step still exists, and can still be omitted, in the control version of the task; the step simply occurs before participants receive feedback about the status of the main goal, rather than after this feedback.

For both experiments, the primary dependent measure was the frequency of error committed by participants at the postcompletion step in the phaser task. For any given action step in the task procedure (see Figure 1), a participant could make an indeterminate number of inappropriate actions (i.e., click the incorrect button). Therefore, rather than each inappropriate click at a particular step being counted as an error, an error was counted if *any* inappropriate action was taken at a particular step. Because participants were completing a sequential choice task, interest was focused on the action *steps* that produced errors, not necessarily on the pattern of incorrect actions at any given action step. Therefore, the overall error count for a trial was the number of action steps in the trial that had inappropriate actions, not the total number of incorrect actions. A postcompletion error was counted if participants tested with the postcompletion version of the phaser task clicked the *main control* button rather than the *tracking* button after destroying the target. It was possible for participants to make other inappropriate actions at this step in the procedure (any click other than the correct click on the *tracking* button or the postcompletion error of clicking on the *main control* button), although this did not occur in the Byrne and Bovair (1997) results.

Error at the postcompletion step was assessed in two different ways. Because participants might not destroy the target on the first try (though they often did), necessitating a repeat of the sequential task procedure until the target was destroyed, it



Postcompletion Version

- Action 1: Click "Power Connected"
- Action 2: Click "Charge"
- Wait until phaser charges the appropriate amount
- Action 3: Click "Stop Charging"
- Action 4: Click "Power Connected"
- Action 5: Click "Settings"
- Action 6: Adjust location of slider to desired focus
- Action 7: Click "Focus Set"
- Action 8: Click "Firing"
- Action 9: Click "Tracking"
- Use arrow keys to adjust location of target indicator
- Action 10: Press the space bar to fire
- **Determine if the target has been destroyed**
  – If not, go back to Action 1
  – **If yes, Action 11: Click "Tracking"**
- Action 12: Click "Main Control"

Control Version

- Action 1: Click "Power Connected"
- Action 2: Click "Charge"
- Wait until phaser charges the appropriate amount
- Action 3: Click "Stop Charging"
- Action 4: Click "Power Connected"
- Action 5: Click "Settings"
- Action 6: Adjust location of slider to desired focus
- Action 7: Click "Focus Set"
- Action 8: Click "Firing"
- Action 9: Click "Tracking"
- Use arrow keys to adjust location of target indicator
- Action 10: Press the space bar to fire
- Action 11: Click "Tracking"
- **Determine if the target has been destroyed**
  – If not, go back to Action 1
  – **If yes, Action 12: Click "Main Control"**

Figure 1. Phaser task screen and step summary.

was possible to have several opportunities within a single trial to commit an error at each action step. Therefore, one measure of error assessed the occurrence of error at a particular action step in the procedure in relation to the number of opportunities there were to commit an error: *Error frequency* is defined as the number of errors at step X_i divided by the number of opportunities for error at step X_i. Each step can be considered a sequential choice (Ohlsson, 1996), so the definition was based on the step, not the action. That is, each step was counted as either containing an error or not containing an error, regardless of the number of incorrect actions taken. For example, if a participant was at Step 3 of the procedure and clicked one incorrect button, that means an error was made at Step 3. Further incorrect clicks made there do not advance the state of the procedure, so they were not counted as additional errors. This focuses the analysis on where in the procedure the errors occurred.

The second measure of error was more specific in scope: assessing the occurrence of error at a particular action step in relation to the total number of errors for that trial. This measure assesses whether errors tend to occur systematically at a particular step in the task procedure; *error proportion* is thus defined as the number of errors at step X_i in trial *n* divided by the total number of errors in trial *n*. This may appear to be an uninformative measure, but it allows for assessment of the systematicity of the error-generating mechanism. If errors are caused by a simple stochastic mechanism with a movable central tendency (i.e., "more errors" vs. "less errors," perhaps caused by high working memory load), then all errors should account for approximately the same proportion. Although falsifying this model is in some sense attacking a straw man, it is important that this explanation be empirically ruled out rather than simply dismissed out of hand.

Another performance variable of interest was the amount of time it took participants to complete the task procedure. The inverse relation between speed and accuracy (i.e., the speed-accuracy trade-off) raises interesting questions regarding performance discrepancies. In terms of industry, deficiencies in either speed or accuracy could represent unsatisfactory performance. Typically, the more rapidly people try to complete a task, the more likely they are to commit errors, and vice versa (Osman et al., 2000). Step completion time was measured as the time elapsed (in milliseconds) between the successful execution of the previous step and the successful execution of the current step – that is, as an interclick latency. Steps containing errors were omitted from this analysis.

Based on the Byrne and Bovair (1997) results, it is reasonable to expect postcompletion errors to be made in this task and for the errors to be more systematic than predicted by a simple stochastic model. However, those experiments did not compare a postcompletion version of the tasks used with a non-postcompletion version of the same task, assess speed-accuracy trade-off, or examine the effects of more extended exposure to the task. The first experiment here was designed primarily to explore those three issues. Other issues examined in the Byrne and Bovair (1997) experiments, such as individual differences in working memory and different postcompletion tasks, were not explored.

EXPERIMENT 1

Method

Participants. Twenty-two undergraduates from Rice University participated for course credit in a psychology course. There were 12 men and 10 women, with a mean age of 19.8 years and a standard deviation of 1.2 years.

Materials. A paper instruction manual was used for training participants on the phaser task. The phaser task was run through an application written in Macintosh Common Lisp version 4.3 on Apple iMac computers. Each participant wore lightweight stereo headphones that were connected to his or her computer.

Design. This experiment was a two-factor mixed between- and within-subjects design consisting, respectively, of task assignment and test session. Participants were randomly assigned to one of two conditions: control version of the phaser task or postcompletion version of the phaser task. Participants received one session of training and were then tested across three sessions. Participants also performed three other similarly themed tasks to prevent them from focusing entirely on errors committed in the phaser task. However, those tasks did not contain postcompletion steps and so are not included in this analysis.

Procedure. The experiment was conducted over four sessions, each 1 week apart from the next. During the first session (training), participants were trained on the phaser task procedure. Using the provided instruction manual as a guide, the participants followed the task sequence until they had successfully carried out the procedure one time without any errors. After this initial error-free trial, the instruction manual was removed and participants continued training on the procedure until they had reached the standard of three error-free trials. During training, on trials in which the participant committed an error, the trial was stopped, the participant was informed of the appropriate action that should have been taken, and a new training trial was initiated. Therefore, any given training trial could have one or zero errors.

The second, third, and fourth sessions consisted of test trials for the phaser task; these will be referred to as Sessions 1, 2, and 3 of testing (as opposed to training). Participants completed 10 trials of the test task (phaser task) per session. During these sessions, the application program signaled (buzzed) participants when an inappropriate action was taken. This allowed participants to know when they had selected the incorrect action for the particular action step.

Participants were timed on their performance and were encouraged to have both timely and accurate task completion. Upon successful completion of a trial, participants were informed of the time it took to complete the trial and the number of errors they had committed.

In order to impose a working memory load, we had participants complete a concurrent letter recall task during phaser task performance (as per Byrne & Bovair, 1997). Randomly ordered letters were spoken at the rate of one letter every 4 s. (Note that Byrne & Bovair, 1997, used a rate of once every 3 s. This slower rate used here was attributable to a programming error.) A tone was presented randomly at intervals ranging from 12 to 60 s. At the tone, participants were to type the last three letters they had heard, in the order they heard them, into a text box that appeared on the computer screen. Participants were assessed for their accuracy on this task (percentage of trials correct). The interrupted task trial resumed after the letter recall task response was logged.

Results

Error proportion. One of the goals of the current research was to assess whether the postcompletion error (the "turn off tracking" step) occurred systematically during task performance or if error commission reflected a uniform stochastic process. Such a hypothesis would predict that in this task, any particular step would be responsible for 1/12 (12 steps in the procedure) or 8.3% of the errors made. (This is, in fact, a conservative overestimate because some participants occasionally had to do the first several steps in the procedure more than once if they missed the target.) As can be seen in Figure 2, participants in the postcompletion condition on the first two sessions had error proportions for this step (as defined previously) that were substantially higher than this; fully 40% of the total errors made were postcompletion errors. This difference is reliable (or statistically significant; *t* tests vs. null hypothesis of 1/12 yielded *p* values < .05 for both days), suggesting a clearly systematic error. That is, the high memory load imposed by the secondary task did not yield a generally high level of error across all steps; it produced a disproportionate number of postcompletion errors, emphasizing the importance of task structure in this kind of error. This replicates the Byrne and Bovair (1997) results. However, the case is not as clear on the third session of testing; it appears that more extensive exposure to the task may reduce the systematicity of the error.

Effects on error frequency. Figure 3 shows the results for error frequency (as previously defined) for the "turn off tracking" step as a function of condition and session. Clearly, the postcompletion condition had a substantially higher error frequency, $F(1, 20) = 14.81, p < .001$. This result replicates the Byrne and Bovair (1997) findings. Extending those findings, the effect of test session was reliable as well, with lower frequency as participants got more experience with the task, $F(1, 20) = 7.06$, $p = .015$ for the linear contrast on session. It also appears that the drop in frequency was larger for the postcompletion condition, as the interaction between condition and session was also reliable, $F(1, 20) = 4.61, p = .044$ for linear contrast on session. However, too much should probably not be read into that interaction, as participants in the control condition were near floor anyway. It should also be noted that 27% of the participants still committed at least one postcompletion error on the third testing session. So, although exposure does appear to reduce the incidence of this type of error, it may not eliminate it; future work would have to include further sessions to determine such effects.

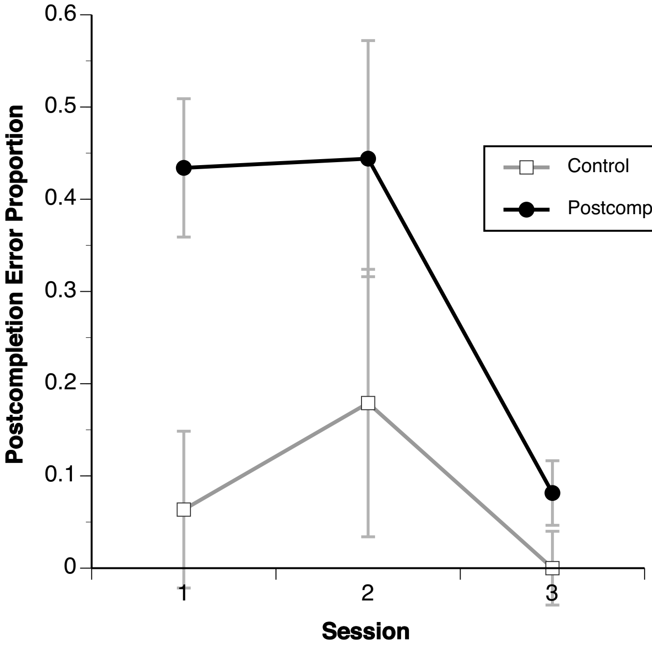


Figure 2. Postcompletion error proportion as a function of condition and session for Experiment 1. Error bars show 1 standard error of the mean.

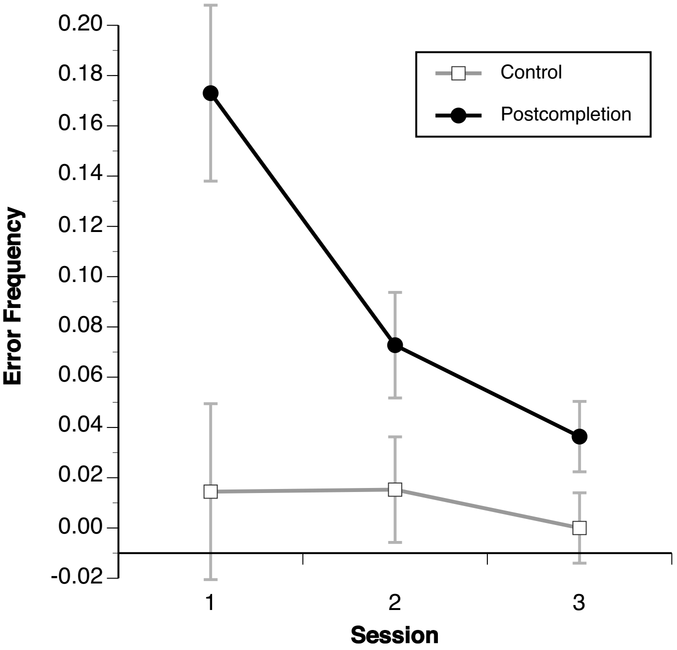


Figure 3. Error frequency at the "turn off tracking" step as a function of condition and session for Experiment 1. Error bars show 1 standard error of the mean.

Effects on step completion time. Another extension of the Byrne and Bovair (1997) findings, which did not include timing information, are the results for step completion time. The difference between control and postcompletion conditions at the postcompletion step is quite large, $F(1, 20) = 63.09, p < .001$; however, it is critical to note that participants in the control condition were substantially faster on this step, not slower. That is, the two conditions were clearly not differentially trading off speed and accuracy, as both measures were superior in the control condition. However, the different requirements of the two versions of the task at this step in the procedure is likely the source of this effect. Participants in the postcompletion condition had to read a feedback message at this step (e.g., "Romulan ship destroyed") and decide whether or not to restart the task, whereas participants in the control condition do not do this until the subsequent step. The appropriate comparison for speed-accuracy purposes is to combine the step completion times for the two steps so that it includes the reading and decision time for both groups; these data are presented in Figure 4. By this measure, there is no reliable difference between the conditions, although there is some improvement in performance over the three test sessions, $F(1, 20) = 10.84, p = .004$ for linear contrast on session. So, the advantage for the control condition disappears when measured this way. However, it is important to note that there is no evidence that the reduction of errors associated with the control version of the task cost the participants any time.

Concurrent memory task. One possible explanation for the superiority of the control group is that they differentially traded off performance in the phaser task with performance on the concurrent memory task. However, results for the concurrent memory task do not support this explanation. The mean percentage correct for the control condition was 42% and for the postcompletion condition it was 37.9%, a nonreliable difference, $F(1, 20) = 0.30, p = .60$. Participants actually got slightly worse on the letter recall as a function of session, but this did not interact with condition, $F(2, 40) = 1.14, p = .33$. Thus, there is no reason to suspect differential handling of the multiple-task situation. (No such analysis was reported by Byrne & Bovair, 1997.)

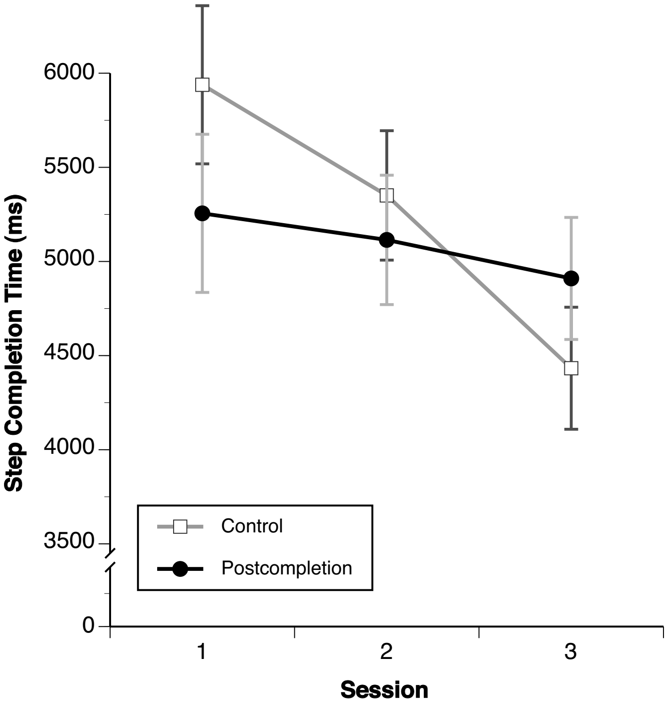


Figure 4. Combined step completion time for "turn off tracking" and subsequent step as a function of condition and session for Experiment 1. Error bars show one standard error of the mean.

Discussion

This experiment replicated Byrne and Bovair (1997) in demonstrating that the postcompletion error in the phaser task (clicking the *tracking* button after destroying the target) is most likely not the result of a simple uniform stochastic mechanism, occurring more often than such an account would predict. At this point, we believe this straw man can be put to rest; postcompletion errors are conspicuously more common than such an account would suggest, at least early in task performance.

Furthermore, participants showed a marked reduction in postcompletion errors over time. As there were no changes to the tasks themselves during the three test sessions, such a decrease in error suggests that some learning took place, resulting in error reduction. This suggests that it may be possible to affect the frequency of this error through training or possibly other interventions. However, it should be noted that for many more naturally occurring postcompletion errors, practice regimens such as the one used here are the exception and not the rule.

Finally, there does not appear to be any time cost associated with the lower error version of the task, suggesting that speed and accuracy need not be traded off against one another. This is particularly interesting because the two versions of the task have the same number of steps executed in the same order. Many error remediation techniques rely on making changes to the task by adding or removing steps and/or subtasks. Such measures appear unnecessary in the case of postcompletion errors.

EXPERIMENT 2

Given a successful replication of postcompletion error as a sequential choice task error in an experimental situation and the finding that the reduced error condition did not come with a time cost, Experiment 2 was designed to assess the potential effects of "blame-and-train" style methods on task performance. A blame-and-train method is characterized by rewarding (or ignoring) error-free performance and punishing error-prone performance, as if people had direct and conscious control of which type of performance they will express. This model arises from the fundamental misconception that individuals are solely responsible for the occurrence of an error and that punishing the individual will improve that particular person's performance while deterring the expression of that same behavior in others. With constructs such as hindsight bias, in which one is led to believe that individuals could have chosen a different (error-free) path from the one that led to failure, it is easy to see why errors are often attributed to lazy, unmotivated, careless, and negligent employees (Woods, Johannesen, Cook, & Sarter, 1994).

The modus operandi in many different types of organizations is to take actions against these hapless employees, not to address the cause of their errors, and this has been called the "blame trap" or "blame and train" (Reason, 1994; Tullo, 2001). Punishment in these cases often takes several forms, including social censure, peer disapproval, disciplinary actions against the employee, and sending the employee to remedial training on the task or tasks in question (Leape, 1994; Reason, 1994). Leape (1994), in describing the culture of blame in health care, tied employee performance to what could be called "train-then-blame." In health care, the approach to error prevention is what Leape (1994) called a "perfectibility model": "If physicians and nurses could be properly trained and motivated, then they would make no mistakes" (p. 1853). According to this way of thinking, if employees are properly trained, they will not make mistakes; if employees make mistakes, then they were not properly trained in the first place and should undergo "proper" retraining to address this deficiency. The situation is perhaps improving in health care, but it is likely that such practices are still widespread there and in other industries.

In some cases, what this comes down to is whether or not the appropriate response to an error is at the level of task and interface design (i.e., cognitive factors; e.g., Hollnagel & Woods, 1999) or at the organizational level in terms of training or reward systems (e.g., Reason, 1997). Obviously, in many cases the appropriate response is both kinds of intervention. An interesting question is whether or not situations exist that are amenable to task and interface interventions but are particularly less amenable to motivational manipulations.

The methods used in this demonstration included praising individuals for acceptable performance, issuing reprimands to a subset of individuals for less than acceptable performance, and reinstructing a subset of individuals on the "correct" task procedure. Because these methods did not directly address the actual task being performed, it was hypothesized that they would have little effect on the occurrence of error during future task performance. Note that this is in contrast to the predictions of the blame-and-train approach, which is focused on motivational aspects of performance. In order to attempt to address the actual performance discrepancy (postcompletion error occurrence), we introduced an intervention that addressed the underlying problem with the task itself. This method redesigned the primary task sequence so that it no longer had a postcompletion step; that is, participants were switched from the postcompletion version of the task to the control version of the task in midexperiment. It was hypothesized that this redesign would reduce the probability of participants selecting the inappropriate choice at that particular action point.

Method

Participants. Fifty-four undergraduate students from Rice University participated for course credit in a psychology course, and 6 psychology graduate students from Rice University received $40 for their participation, for an overall total of 60 participants. There were 41 women and 19 men, with a mean age of 19.6 years and a standard deviation of 1.6 years. None of the participants in Experiment 2 had taken part in Experiment 1.

Materials. Materials were identical to those in Experiment 1 except where noted as relevant to design.

Design. This experiment was a two-factor mixed between- and within-subjects design. Participants were randomly assigned to one of six conditions that represented a method for addressing performance discrepancies: control (control version of the phaser task), baseline, reprimand, reinstruction, praise, and redesign. The baseline, reprimand, reinstruction, and praise conditions used the postcompletion version of the phaser task for all sessions. The redesign condition started with the postcompletion version of the task and switched participants to the control version during Session 2.

Procedure. The overall procedure for Experiment 2 was nearly identical to the procedure for Experiment 1, with the only difference being the between-subjects manipulation. Interventions were introduced during Session 2 of testing to assess their impact on subsequent participant performance. The control and baseline conditions did not receive interventions; they served as comparisons for the other conditions. (Note that the baseline condition was identical to the postcompletion condition in Experiment 1.) Interventions were introduced after participants in the reprimand, reinstruction, praise, and redesign conditions completed six trials of the phaser task. These interventions are described in Table 1. As participants were randomly assigned to a condition, their actual task performance was not used to determine which intervention they received. As participants were tested independently and because they did not have an identifiable standard against which to compare their performance, it was deemed that the random assignment should not confound subsequent performance beyond the effects of the interventions.

For the concurrent letter recall task, the lag between letter presentation was reduced to 3 s match that in the original Byrne and Bovair (1997) study, and the recall cuing tone was presented at random intervals ranging from 9 to 45 s.

Results

Because the first half of the session on Session 2 was preintervention and the second was postintervention, trials for that session would have had to be split into two groups, yielding a low number of trials on that session. This would have made error frequency estimates somewhat unstable in the statistical sense, so results from that session

TABLE 1: Intervention for Reprimand, Reinstruction, Praise, and Redesign Conditions

Condition	Intervention
Reprimand	Report of poor performance and advised to improve.
Reinstruction	Report of poor performance, advised to reread manual and complete paper-and-pencil test demonstrating proficiency.
Praise	Report of good performance and advised to "keep up the good work."
Redesign	Report of poor performance, but advised was due to outdated phaser system. Informed of new version, read control version manual, completed one control test trial, and resumed testing on Days 2 and 3 with control version of phaser task.

were discarded. In addition, 1 participant had exceptionally deviant performance in terms of both speed and accuracy and was removed from those analyses for all days.

Effects on error frequency. Figure 5 depicts the results for error frequency at the "turn off tracking" step for the six conditions on the 2 relevant days of testing. All conditions showed a decline in errors, as demonstrated by the main effect of session, $F(1, 53) = 25.73, p < .001$. Overall, the control condition had a lower total overall rate of errors than the mean of the others, $F(1, 53) = 7.72, p = .008$, but there was no statistically reliable evidence that the amount of change in error rates was a function of condition, interaction $F(5, 53) = 1.04$, p = .40. That is, all conditions improved an approximately equal amount in terms of error frequency. Thus, there is no evidence that the instructional manipulations or the redesign improved error frequency more than participants did simply on their own (i.e., the baseline condition). This suggests a certain ineffectiveness of purely motivational manipulations for this type of error (and also of the redesign).

Effects on step completion time. There was once again a reliable effect of session, $F(1, 53) = 47.61$, $p < .001$, on step completion time, indicating that all conditions showed improvements over time. Of course, as shown in Experiment 1, the advantage for the control condition was also a function of the postponement of the reading and decision processes until the next step in the procedure. So, as in Experiment 1, we combined the completion time for the two steps, yielding the data shown in Figure 6. Here again we found a substantial speedup over the sessions, $F(1, 53) = 20.76, p < .001$, but no conclusive evidence for differential speedup between conditions, though there is a hint that the control condition experienced more speedup than the others, $F(1, 53) = 2.98, p = .09$. What is clear is that here, as in Experiment 1, the control condition showed superior performance in terms of errors without sacrificing speed.

Concurrent memory task. In this experiment, unlike Experiment 1, there was actually a mild increase in letter recall performance across the sessions. However, just as in Experiment 1, there was no reliable difference between conditions, $F(5, 53) = 1.64, p = .16$, nor was there a condition by session interaction, $F(5, 53) = 1.44, p = .23$. Thus, there is no reason to believe participants were differentially trading off with the working memory task.

Discussion

Redesigning this sequential choice task did not result in reliably fewer errors than did reprimands, reinstruction, or praise. Overall, the only group that did better on the whole was the control group,

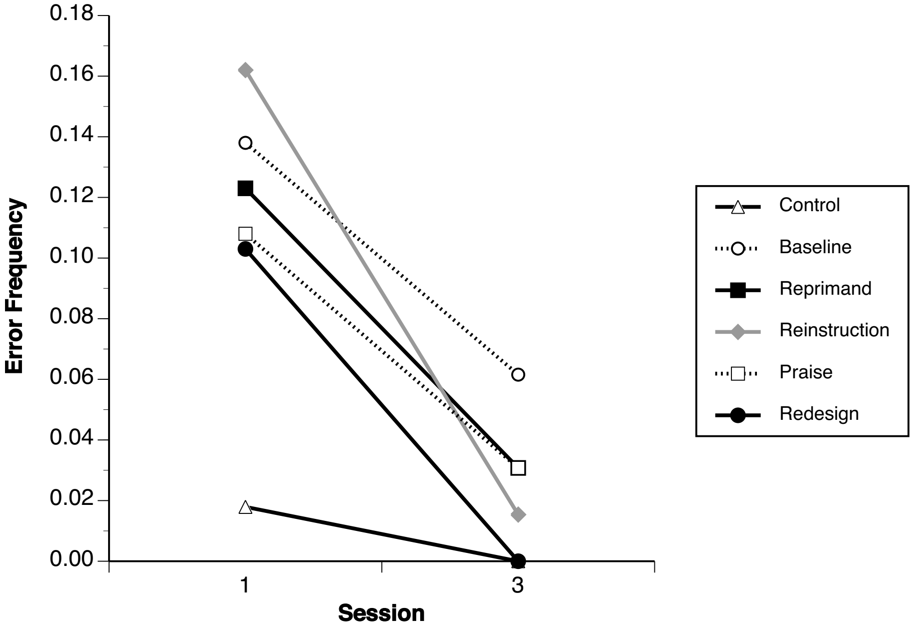


Figure 5. Error frequency at the "turn off tracking" step as a function of condition and session for Experiment 2.

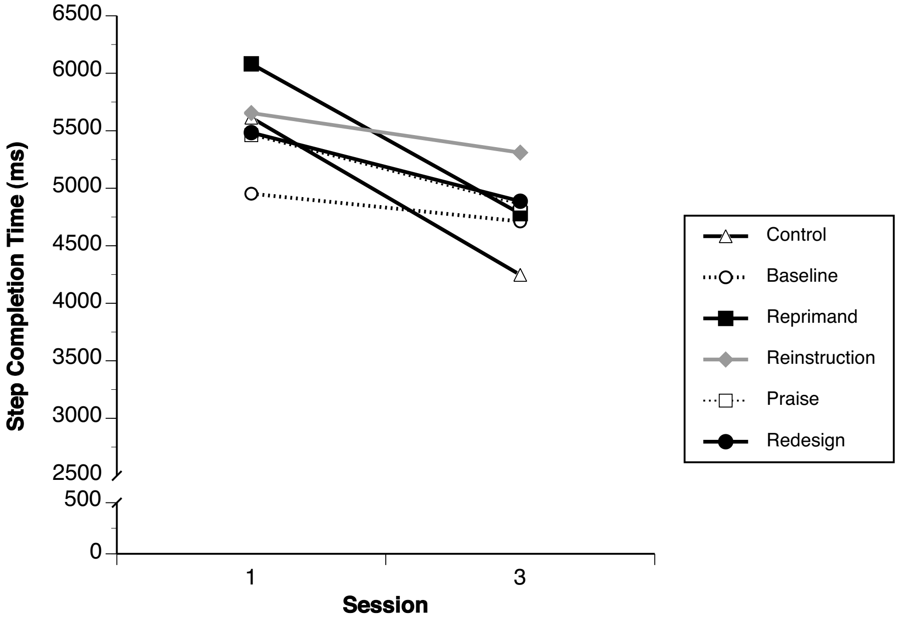


Figure 6. Combined step completion time for "turn off tracking" and subsequent step as a function of condition and session for Experiment 2.

who had a non-postcompletion version of the task from the outset. The fact that the motivational manipulations (reprimand, reinstruction, and praise) did not appear to have a reliable effect on altering task performance (for either task completion time or task accuracy) lends support to the idea that "blame-and-train" may not be a particularly effective paradigm for addressing performance discrepancies in routine procedural tasks. This suggests, at least for routine procedural tasks, that assessing the demands the task places on the operator's cognitive capabilities may be a more fruitful approach for error remediation.

The redesign condition was similarly ineffective; it is noteworthy, however, that although participants in this condition did not show significantly more performance improvement than those in the baseline and motivational conditions, they no longer made *any* errors at this step of the procedure in Session 3. In contrast, they did not show substantial overall speedup in task performance. Taken together, these hint that although redesigns after the fact may indeed be helpful, perhaps they are still inferior to getting the system design right in the first place. (For example, redesigns open the possibility of negative transfer effects.) Obviously, the data here are only suggestive at best, and further studies would need to be done to fully justify such a conclusion.

It is also worth noting that although participants in the various noncontrol conditions did not often make the postcompletion error at the "turn off tracking" action step by Session 3 of testing, this occurred after substantial practice with the task. In contrast to the participants in our experiments, people in many real-world situations are prone to this error as they do not have extensive and massed practice, so one cannot necessarily depend on this kind of learning to prevent the error.

Taken together, results from this experiment point to the value of formative human factors in error reduction, as the only group that outperformed the others was the one with the less error-prone task in the first place.

GENERAL DISCUSSION

There are several important points to be taken from this research. First, an anecdotally pervasive error type, postcompletion error, is resistant enough to self-monitoring that participants in laboratory situations still systematically make this error. Whereas systematic studies of mistake-class errors (i.e., those based on incomplete or inaccurate knowledge of procedures) are not entirely unusual, the vast majority of research on slip-class errors is based on anecdotes and postincident analysis (e.g., Reason, 1990). Our successful replication of the Byrne and Bovair (1997) findings makes it clear that slip-class errors can be viably studied under laboratory conditions, although the particular laboratory conditions involved moderately complex procedures and a highly demanding multitasking situation. Still open is the extent to which these measures are actually necessary to produce such errors, but a fruitful experimental paradigm for the study of slip-class errors is in itself significant.

Furthermore, the findings here go beyond simple replication of a successful methodology. One of the key findings in this research is that none of the after-the-fact manipulations yielded performance that was superior to, or even the overall equal of, the control condition. This is in contrast to what would be expected according to the blame-and-train account; instead, motivational manipulations had no discernible effect. The effective intervention, the control version of the task, was focused on the task, not the user.

That control version includes what might be considered a soft form of a forcing function (Norman, 1988): Participants had to click the *tracking* button after firing at the target in order to find out if their main goal had been completed. Omission of this step was still possible in this version of the task (and thus not a strict forcing function), but it was extremely rare. This highlights the role of goal structure in determining error (similar to Gray, 2000) because the two versions of the task had the same overt actions performed in the same order, but they differed greatly in error frequency at a particular step.

Additionally, this "forced" version of the task essentially breaks the speed-accuracy trade-off, producing superior performance in terms of errors with no overall cost in terms of execution time. Whereas one would expect that errors might be eliminated by forcing operators to complete the task in a slow, controlled fashion, such measures do not appear necessary to eliminate this class of error.

Finally, our results suggest that getting the initial design of the task correct (i.e., by not allowing a task with a postcompletion step or goal to be inflicted on operators in the first place) is of paramount importance, particularly in safety-critical applications. This is hardly a novel lesson in the human factors community, but it may be the case that certain classes of error, particularly slips that result from high demands on memory or attention, are particularly hard to eliminate once systems prone to such errors are introduced. Whether other after-the-fact remediations such as cuing or enforcing some kind of time cost perform better is an open question for future research.

ACKNOWLEDGMENTS

The research presented in this article is based in part on the doctoral dissertation of Elizabeth M. Davis. We would like to thank Bryan Blauvelt and Peter Centgraf for their assistance in running participants. Michael Byrne's work on this research was supported in part by Office of Naval Research Grant #N00014-03-1-0094.

REFERENCES

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

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Hollnagel, E., & Woods, D. D. (1999). Cognitive systems engineering: New wine in new bottles. *International Journal of Human-Computer Studies, 51,* 339–356.

Leape, L. L. (1994). Error in medicine. *Journal of the American Medical Association, 272,* 1851–1857.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88,* 1–15.

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

Ohlsson, S. (1996). Learning from error and the design of task environments. *International Journal of Educational Research, 25,* 419–448.

Osman, A., Lou, L., Muller-Gethmann, H., Rinkenauer, G., Mattes, S., & Ulrich, R. (2000). Mechanisms of speed-accuracy tradeoff: Evidence from covert motor processes. *Biological Psychology, 51,* 173–199.

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

Reason, J. (1994). Foreword. In M. S. Bogner (Ed.), *Human error in medicine* (pp. vii–xv). Hillsdale, NJ: Erlbaum.

Reason, J. (1997). *Managing the risks of organizational accidents.* Hampshire, UK: Ashgate.

Sarter, N. B., & Alexander, H. M. (2000). Error types and related error detection mechanisms in the aviation domain: An analysis of ASRS incident reports. *International Journal of Aviation Psychology, 10,* 189–206.

Tullo, F. J. (2001). Viewpoint: Responses to mistakes reveal more than perfect rides. *Aviation Week & Space Technology, 154*(21), 106.

Woods, D. D., Johannesen, L. J., Cook, R. I., & Sarter, N. B. (1994). *Behind human error: Cognitive systems, computers and hindsight.* Dayton, OH: CSERIAC.

Xiao, Y., & Mackenzie, C. F. (1995). Decision making in dynamic environments: Fixation errors and their causes. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 469–473). Santa Monica, CA: Human Factors and Ergonomics Society.

Michael D. Byrne is an associate professor in the Department of Psychology at Rice University. He received his Ph.D. in experimental psychology in 1996 at the Georgia Institute of Technology.

Elizabeth M. Davis is a systems improvement specialist at the Institute for Healthcare Excellence at the University of Texas M. D. Anderson Cancer Center, Houston. She received her Ph.D. in psychology (human factors) in 2001 at Rice University.

Date received: September 14, 2002
Date accepted: October 24, 2005