Preventing Postcompletion Errors: How Much Cue Is Enough?

Michael D. Byrne (byrne@acm.org)
Departments of Psychology and Computer Science
Rice University, 6100 Main St., MS-25
Houston, TX 77005 USA

Abstract

Postcompletion error (Byrne & Bovair, 1997) is a robust form of routine procedural error that can be difficult to prevent. However, Chung and Byrne (2004, 2008) reported the complete elimination of this error with a highly aggressive cue, one which was just-in-time, highly specific, and highly visually salient. This paper reports three experiments designed to investigate the role of each one of these factors by using cues which have only two of the three key properties. Surprisingly, salience appears to be the least critical property, and being just-in-time seems to be the most vital. This has important implications for the mitigation of real-world postcompletion errors.

Keywords: human error; postcompletion error; routine procedures; visual cues; visual salience

Introduction

Even in the execution of known routine procedures, people make non-random errors. Everyone has had this experience, whether it is leaving one's bank card in an automated teller machine (ATM) or failing to attach a promised file to an email message. While many such errors have little or no real cost, many such errors have dire consequences, including loss of human life (Casey, 1998). A particularly robust type of routine procedural error is the *postcompletion error* (or PCE; Byrne & Bovair, 1997). Postcompletion errors occur when the primary goal of a task is fulfilled and terminal step (or steps) of the procedure, which occur after this, are omitted. Standard examples of this error include the aforementioned bank card left in an ATM, leaving the original document on a photocopier or flatbed scanner, and forgetting to replace the gas cap when filling up the car; most people are personally familiar with at least one postcompletion error.

While postcompletion errors have been discussed in the human-computer interaction literature for some time (e.g., Young, Barnard, Simon, & Whittington, 1989; Polson, Lewis, Reiman, & Wharton, 1992), it is only in the last few years that they have begun to receive more substantial attention, both theoretical (e.g., Altmann & Trafton, 2002) and empirical (Byrne & Davis, 2006; Li, Blandford, Cairns, & Young, 2005; Li, Cox, Blandford, Cairns, & Abeles, 2006; Ratwani, McCurry, & Trafton, 2008). This makes postcompletion errors unique among routine procedural errors, as errors in such tasks are generally not robust enough to lend themselves to controlled experimentation. Postcompletion errors, however, are both pervasive and robust, making them both tractable to study in the laboratory as well as significant in the real world. For example, one form of postcompletion error is reportedly common in electronic voting machines: voters forget to press the final "cast vote" button after making all their selections, thus failing to actually vote (see Everett, Greene, Byrne, Wallach, Derr, Sandler, & Torous, 2008). Legitimate practical consequences naturally lead to questions about what can be done to mitigate such errors.

In many cases, the solution is obvious: re-design the task so that there are no postcompletion steps. This is the strategy now used in many ATMs: the cash is not dispensed until the user has retrieved his or her card. While this is almost certainly the optimal solution in many cases, there are many other cases where this approach is not possible, such as already-deployed hardware-based systems.

The next most obvious mitigation strategy is to provide some kind of cue to let the user know that such an error has occurred (or is possible). Again, this is the solution adopted by many ATMs; they emit a loud beeping sound and/or flash a light near the card slot after dispensing cash. But how effective are such cues?

This is exactly the question asked by Chung and Byrne (2004; 2008). In one of their experimental conditions, they introduced a cue which *completely* eliminated postcompletion errors by their experimental subjects. The cue in this case was a pair of colored, blinking arrows which pointed to the button which needed to be pressed at the time it needed to be pressed. This cue has three properties that Chung and Byrne considered critical:

- *Salience*. The display used was a grayscale display, so the blinking colored arrows had a high level of perceptual salience.
- *Specificity*. Unlike the beeping of the ATM, which is non-specific, the cue used made it clear to the user exactly what to do: the arrows pointed directly to the relevant button on the display.
- *Just-in-time*. The cue appeared right at the moment that action on the critical button needed to be taken. Thus, it did not rely on the user to have to remember to take action at a specific time.

Chung and Byrne (2008) also describe an ACT-R model of the postcompletion error which demonstrates the effectiveness of the cue. As the model was constructed, it depended critically on all three of these properties to prevent the error. The model had to notice the cue (salience), had to have knowledge about what the cue meant (specificity), and had to be able to act on that knowledge immediately (just-

in-time). Thus, the model predicts that all three properties are necessary in order for the cue to fully mitigate the error.

The goal-activation account of Altmann and Trafton (2002) is the other most obviously relevant account of cognitive control; the Chung and Byrne model was built with this account in mind. Both agree that the cue must be just-in-time to be effective. Both these accounts posit that if the external cue is noticed, if it meets the other conditions, it should be able to prevent the error. Thus, a reduced-salience cue should be partially effective as long as it is noticed at least some of the time. To be 100% effective the cue would have to be noticed all the time, so high salience should be required for perfect mitigation, but both agree that partial mitigation should be possible.

The most complex property with respect to these two accounts is specificity. Neither are entirely clear about how much specificity is enough. The Altmann and Trafton account says that to the extent that the environmental cue primes the correct action, it should be effective. That is, if the cue primes enough to overcome background activation noise, then the cue should be enough. If not, it shouldn't. Cues which prime a little might then be partially effective. However, predictions about the degree to which that priming should occur in specific situations is unclear.

The Chung and Byrne ACT-R model has a similar scope problem with less specific cues, though it is in some sense less probabilistic than the Altmann and Trafton account because it does not rely on a noisy priming process, but rather a single production rule to associate the cue and the appropriate action. If the model has that rule, then the cue should work every time. If not, then the cue should fail every time. This seems a bit too discrete to be accurate. However, one has to consider how such a rule would be formed by ACT-R's learning mechanisms. To form such a rule, essentially the same information must be encoded in declarative memory. That is, the model has to have a chunk which says "What do I do when I see the cue? Take action x." When the model first sees the cue, this should be retrieved, and after enough retrievals, a production rule will be compiled. However, until that rule is compiled, the model will have a chance to fail when it tries to retrieve the relevant chunk. The failure rate for this retrieval will be a function of the strength of association between the cue and the relevant chunk—essentially the same unknown and non-predicted parameter as in the Altmann and Trafton account.

It is hard to say what other accounts of cognitive control for routine procedural tasks (e.g., Cooper & Shallice, 2000; Botvinick & Plaut, 2002) would predict, since these accounts do not cover postcompletion errors. This kind of phenomenon is particularly problematic for the Botvinick and Plaut account which argues that subgoals are epiphenomenal, in which case there should be no postcompletion errors at all. Whether the Cooper and Shallice account can be extended to postcompletion errors remains an open question.

So, how do the predictions hold up? How effective would a cue be if it had only some of those properties? And which ones are most important? These are exactly the questions the following series of experiments were designed to investigate. Cues with only two of the three properties were implemented and compared with a non-cued control condition. To the extent that the weaker cue is effective, it is reasonable to conclude that whatever property that cue is missing is less important to mitigating postcompletion errors.

Experiment 1

In Experiment 1, the non-cued (control) condition was compared with two conditions: one a cue that was just-in-time and visually salient, but nonspecific, and one that was just-in-time and specific, but less salient.

Methods

Participants. Participants were 47 Rice University students ranging in age from 17-24 who were compensated either with $25 or credit towards a course requirement**.**

Design. The design was a one-way between-subjects design with three conditions, to which participants were randomly assigned: Control (no cue), Reduced Specificity, and Reduced Salience. See the Stimuli section for details of how the cues were implemented.

The primary dependent measure was error frequency on the postcompletion step. On each trial, errors were binary; that is, either the participant did or did not make an error on the postcompletion step. An error consisted of clicking on any button other than the correct button.

As a secondary dependent measure, the time taken to complete each step was also recorded, primarily to assess speed-accuracy tradeoff effects.

Stimuli. The task used in this experiment was the same Phaser task used in previous work (i.e., Byrne & Davis, 2006; Chung & Byrne 2004, 2008) based on the Phaser task from Byrne and Boviar (1997). The experiment was set in a fictional "Star Trek"-themed universe, and the task was to fire the phasers to destroy the enemy Romulan ship. The display is presented in Figure 1. Participants learned and performed multiple tasks in this setting; only the Phaser task is relevant here.

When a participant completed any task, they clicked on a "Main Control" button to return to the display where they received feedback about their performance and received the next task. The Phaser task procedure had 11 steps. After the 10th step, the participant had fired the phaser and knew whether or not the enemy had been destroyed. However, before returning to Main Control, the task required that the "Tracking" button be clicked to de-activate the tracking system. This was the postcompletion step.

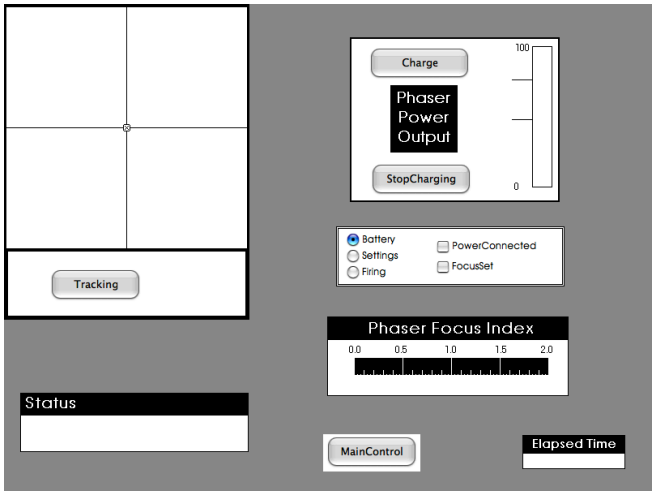

Figure 1. Display for the Phaser task.

The Control condition was identical to the one used in previous research. That is, there was no special reminder, instruction, or cue to the participants to remember the postcompletion step. In the Reduced Specificity condition, participants saw a cue that consisted of a blinking, alternating red and yellow, dot which appeared near the "Tracking" button. This cue appeared as soon as the 10th step was complete, and continued to blink until the participant clicked on the "Tracking" button. This cue is less specific than the successful cue in the Chung and Byrne experiments because it does not clearly indicate which control is to be clicked.

In the Reduced Salience condition, two red arrows appeared, one on either side of the "Tracking" button. These also appeared immediately after completion of the 10th step and also remained visible until the "Tracking" button was clicked. However, these arrows neither blinked nor changed color. Since the display was almost entirely gray and white, the arrows were a color singleton.

Procedure. Participants were run in two sessions spaced approximately one week apart. The first session was a training session. In this session, participants read about each of the tasks and then performed them once, during which time they could refer to the manual. In the cued conditions, the manual made specific mention of the appearance of the cue and the purpose of cue. After the first trial on a task, the manual was returned to the experimenter and participants trained until they achieved the criterion of three error-free performances of the task. When an error was made during training, the trial was immediately ended and the participant was given feedback about the correct step, and a new trial was begun.

Participants trained on three different tasks, only one of which is reported here. Order of training was randomized between subjects. Data on performance during training are not reported here.

The second session was the test session. Participants received a total of 39 trials, 15 of which were with the Phaser task. Order of trials were randomized for each participant. During the second session, the experiment software emitted a warning beep whenever the participant made an error. However, there was no other error message and the participant was allowed to continue the trial after making an error, unlike during training.

Since postcompletion error frequency is linked to memory load (Byrne & Bovair, 1997), during testing participants also performed a concurrent memory loading task. Participants heard a letter presented through headphones at a rate of one letter per 3 seconds. At random intervals ranging from 9 to 45 seconds, a tone replaced the letter and a dialog box appeared on the display, prompting participants to enter the last three letters they had heard, in order.

Finally, participants were encouraged to work quickly and accurately by a scoring system. Participants earned points for correct steps and for rapid performance, and were penalized for errors in task execution or letter recall. The top three scorers among the participants earned additional compensation.

Materials. Stimuli were presented on Apple eMac computers with 17" displays at 1024 x 768 resolution; the Phaser task was implemented in Macintosh Common Lisp. Participants wore Sony MDR-201 headphones.

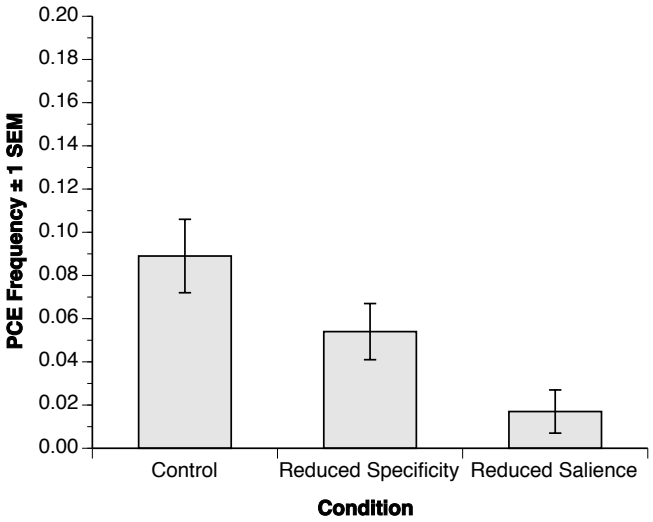

Figure 2. Postcompletion error frequency by condition for Experiment 1.

Results

The primary result for Experiment 1 is presented in Figure 2. Obviously, neither of the cues were as effective as Chung and Byrne's cue, which completely eliminated postcompletion errors. However, the the cued conditions did lead to fewer PCEs than the control condition, contrast on control vs. cues $F(1, 44) = 9.92$, $p = .003$. While the Reduced Salience cue did produce fewer errors than the Reduced Specificity cue, this difference did not quite reach

the conventional significance level, contrast of cue conditions vs. each other $F(1, 44) = 3.74$, $p = .06$.

The reduction in errors in the cued conditions was not due to a speed-accuracy tradeoff, as the mean step completion times for the cued conditions were actually slightly smaller for the cued conditions. ($M_{control} = 4.2$ sec, $M_{reduced\ specificity} = 3.9$ sec, $M_{reduced\ salience} = 3.3$ sec; $F(2, 44) = 2.76$, $p = .07$). If anything, the cues helped encourage more rapid performance while also reducing errors.

Discussion

Both cues produced an intermediate level of mitigation. That is, neither cue eliminated postcompletion errors, but both cues reduced PCEs considerably. This suggests that all three properties are indeed necessary to be sure to mitigate the error, but that weakened cues can at least provide some degree of protection against postcompletion errors.

However, this experiment is limited because it did not examine a cue lacking the just-in-time property. This was the purpose of Experiment 2.

Experiment 2

Experiment 1 examined two cues, each one missing one of the three properties considered critical by Chung and Byrne (2004; 2008). Experiment 2 tried a third cue, this one lacking the just-in-time property and instead providing a cue in advance of the time needed. Providing a cue prior to the appropriate time sets up a prospective memory situation; that is, the cue needs to be remembered and acted up on later, which is a notoriously unreliable mitigation strategy (e.g., Guynn, McDaniel, & Einstein, 1998). The situation is particularly bad for postcompletion errors, which are more likely when memory load is high. Thus, mitigation with a prospective cue seems particularly unlikely to be effective in this case. However, it is important to verify this empirically.

Methods

Except where noted below, methods for Experiment 2 were the same as those in Experiment 1.

Participants. Participants were 45 Rice University undergraduate students ranging in age from 18-22 who were compensated either with $25 or credit towards a course requirement**.**

Design and Stimuli. There were three between-subjects conditions to which participants were randomly assigned. Two of these were non-cued control conditions where other aspects of the display not related to the postcompletion error were manipulated. In one condition the buttons were changed so that all of them were push button-style buttons (that is, the radio buttons and checkboxes were replaced with pushbuttons). In another, extra unused buttons were added to the display. (Effects of these other manipulations on the earlier steps in the procedure were small and are reported in Byrne, Maurier, Fick, & Chung, 2004.) The third condition was a cued condition. In this condition, the red and yellow flashing arrows from Chung and Byrne were used; however, they were not presented at the time the "Tracking" button needed to be clicked. Instead, they were presented after the 9th step of the procedure. This step initiated a tracking subtask where the participant had to track a moving blip to fire the phaser at it. This subtask took participants on average a little more than seven seconds to complete. In this cued condition, the red and yellow blinking arrows blinked for the first two seconds of the tracking subtask, and were then removed from the display. Thus, the cue was highly salient and specific, but it was not just-in-time, and will thus be referred to as a *precue*.

Results

Because the two control conditions were not different at the postcompletion step and showed no reliable difference in error frequency, $t(26) = 0.70$, $p = .49$, the two control conditions were combined.

Figure 3 shows postcompletion error frequency for the control and precue conditions. Obviously, this difference is not reliable, $t(43) = 0.14$, $p = .89$, as it is hardly a difference at all. This is unlikely to be an issue of statistical power, as the power to detect a difference as large as the one obtained between the control and the cued conditions in Experiment 1 was just over 85%. (Power calculations were performed with G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007.) If the Experiment 2 cue had been as effective as those cues, it most likely would have been detected.

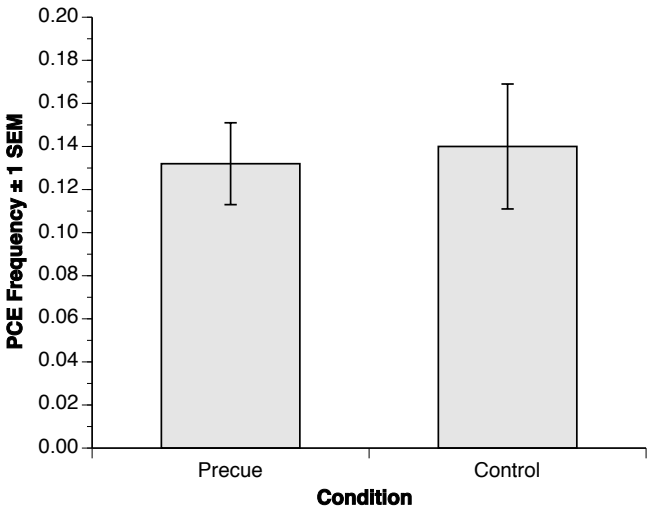

Figure 3. Postcompletion error frequency by condition for Experiment 2.

Again, there was no issue with speed-accuracy trade-off here, as the mean step completion time for the precue group was 3.6 sec and the mean for the control group was 4.1 sec, $t(43) = 0.74$, $p = .46$.

Discussion

Despite high salience and specificity, the cue used in Experiment 2 was completely ineffective at mitigating postcompletion errors. Given the prospective/working memory issues in play, this is not particularly surprising. However, it is interesting that cues missing one of the other two key properties were still partially effective, yet this cue was not. This suggests the just-in-time nature of the cue is the most critical of the three key properties.

Experiment 3

The previous experiments suggest that just-in-time is the most important property and that a cue with somewhat reduced salience can still be at least helpful. Just how important is salience? How low can cue salience be while still maintaining some effectiveness for the cue? This is the key question that motivated Experiment 3.

Methods

Except where noted below, the methods for Experiment 3 were identical to those for Experiments 1 and 2.

Participants. Participants were 45 Rice University students ranging in age from 18-31 who were compensated either with $25 or credit towards a course requirement**.**

Design and Stimuli. There were three between-subjects conditions to which participants were randomly assigned. Two of these were non-cued control conditions where other aspects of the display not related to the postcompletion error were manipulated. In one condition, the display was re-arranged to a different layout. In the other, the text on the buttons was replaced with a string of "x" characters to prevent semantic label-following.

The third condition was a cued condition. In this condition, arrows appeared at the same location as in the previous experiments, and appeared just-in-time. However, the arrows in this condition were a very light gray and did not blink on and off. Thus, salience was quite low, while specificity and just-in-time properties were maintained.

Results

Because the two control conditions were not different at the postcompletion step and showed no reliable difference in error frequency, $t(28) = 1.31$, $p = .20$, the two control conditions were combined.

Figure 4 shows postcompletion error frequency for the control and cued conditions. This difference is statistically reliable, $t(43) = 3.51$, $p = .001$. Clearly, this cue did reduce postcompletion errors, though it did not eliminate them. Again, this is not the result of a speed-accuracy tradeoff, as the trend was actually in the opposite direction, with the cued condition being faster than the control. The mean step completion time for the control condition was 4.2 sec and it was 3.7 sec for the cued condition, $t(43) = 1.78$, $p = .08$.

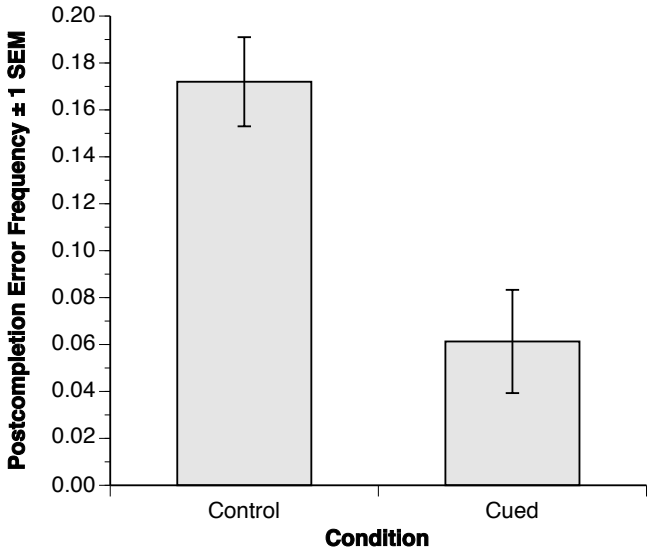

Figure 4. Postcompletion error frequency by condition for Experiment 3.

Discussion

Experiment 3 demonstrates that it is possible to have a cue which at least partially mitigates postcompletion errors even when that cue has modest salience. Obviously, the cue had some degree of salience, as there was still a visual onset when the cue first appeared, and onsets themselves tend to be salient (e.g., Yantis & Jonides, 1990). However, this cue was so light that it was barely visible; in fact, some participants complained that they had trouble seeing it at all. Nonetheless, the cue did substantially reduce, though not eliminate, postcompletion errors. This suggests that while salience plays a role in cue effectiveness, high levels of salience are not necessary to confer substantial benefit.

General Discussion

These results show that for a cue to be fully effective at mitigating postcompletion errors, that all three properties—salience, specificity, and just-in-time—are required. Removing any one of these weakens the cue; removing the just-in-time property makes the cue entirely useless. This has implications both for design of real-world systems and for the theories in this domain.

On the practical side, it is clear that designers of real-world warnings and cues often rely solely on cue salience. Aircraft have many bright red flashing lights which are non-specific and not just-in-time. ATMs which beep at the postcompletion step provide a cue which is possibly salient, sometimes a little late in terms of timing, and entirely non-specific. It is thus unsurprising that ATMs in many parts of the world no longer beep and instead use a task re-design to prevent the error.

The fact that all three properties are important also explains why it is that it is so difficult to mitigate PCEs via cueing. In order to provide a cue, the system needs to know

both what action the user should take and when they should take it. If the system knows what to do and when, why not simply automate the step in the first place? Many postcompletion errors will be impossible to mitigate exactly because the system cannot know either the appropriate time, the appropriate action, or both.

Another possible drawback of the cueing approach to error mitigation is that people may learn to depend on the cue and thus be particularly error-prone were the cue to fail. Further studies need to be done to determine the extent to which people develop such a reliance on the cue.

These results also have implications for theories of cognitive control and goal management. Both the Chung and Byrne model and the Altmann and Trafton account predict that a cue that is not just-in-time will have no effect, as found in Experiment 2. They also essentially agree about the role of salience. As both theories are somewhat vague with respect to the role of specificity, they both at least suggest some role, which is consistent with the partial mitigation found in Experiment 1.

One of the issues which is still open in the domain is more precise specification of specificity, salience, and just-in-time. There are many different definitions of salience in the visual attention literature; which is the most appropriate in this domain? Similarly, specificity is also probably not binary. And there is plenty of room for exploration in terms of just how temporally close the cue must be to the appropriate action in order to count as just-in-time.

Despite their importance in both everyday life and in safety-critical situations, errors in the execution of routine procedures are still not well-understood. This research is one of many steps which need to be taken to rectify that situation.

Acknowledgments

I would like to thank Phil Chung, Chris Fick, and Jennifer Tsai for their assistance in conducting the experiments.

This research was supported by the Office of Naval Research under grants #N00014-03-1-0094 and #N00014-06-1-0056. The views and conclusions expressed are those of the author and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of the ONR, the U.S. government, or any other organization.

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science, 26*, 39–83.

Botvinick, M., & Plaut, D. C. (2002). Representing task context: Proposals based on a connectionist model of action. *Psychological Research, 66*, 298–311.

Byrne, M. D., & Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science, 21*, 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., Maurier, D., Fick, C. S., & Chung, P. H. (2004). Routine procedural isomorphs and cognitive control structures. In C. D. Schunn, M. C. Lovett, C. Lebiere & P. Munro (Eds.), *Proceedings of the Sixth International Conference on Cognitive Modeling* (pp. 52-57). Mahwah, NJ: Erlbaum.

Casey, S. (1998). *Set phasers on stun* (2 ed.). Santa Barbara, CA: Aegean.

Chung, P. H., & Byrne, M. D. (2004). Visual cues to reduce errors in a routine procedural task. In *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Chung, P. H., & Byrne, M. D. (2008). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. To appear in *International Journal of Human-Computer Studies*, *66*, 217–232.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology, 17*(4), 297–338.

Everett, S. P., Greene, K. K., Byrne, M. D., Wallach, D. S., Derr, K., Sandler, D., & Torous, T. (2008). Electronic voting machines versus traditional methods: Improved preference, similar performance. In *Human Factors in Computing Systems: Proceedings of CHI 2008* (pp. 883–892). New York: ACM.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

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

Li., S. Y. W., Blandford, A., Cairns, P., & Young, R. M. (2005). Post-completion errors in problem solving: A study. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1278–1283). Austin, TX: Cognitive Science Society.

Li., S. Y. W., Cox, A. L., Blandford, A., Cairns, P., Young, R. M., & Abeles, A. (2006). Further investigations into post-completion error: The effects of interruption position and duration. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2008). Predicting postcompletion errors using eye movements. In *Human Factors in Computer Systems: Proceedings of CHI 2008* (pp. 539–542). New York: ACM.

Yantis, S., & Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary verses automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 121-134.