Interruptions can Change the Perceived Relationship between Accuracy and Confidence

Kevin Zish[1], Shiva Hassanzadeh[1], J. Malcolm McCurry[2], J. Gregory Trafton[3]

George Mason University[1]
Fairfax, VA

Harris Corporation[2]
Herndon, VA

Naval Research Laboratory[3]
Washington, DC

Interruptions are disruptive in that they can decrease accuracy and the time taken to complete a task. In fields such as aviation and medicine, interruptions can not only reduce performance but lead to egregious outcomes. In such situations, confidence in whether a procedure has been completed may become a crucial aspect of judging where to resume a task. This paper demonstrates that interruptions both decrease accuracy and reduce confidence. More importantly, interruptions change the relationship between accuracy and confidence, reducing the likelihood that participants can determine where to resume appropriately.

INTRODUCTION

Researchers in cognitive science and human factors can accurately measure how accurate someone's memory is in different situations and different tasks (for a review see: Dunlosky & Metcalfe, 2008). One of the assumptions that most people in applied areas (both lay people and practitioners) assume is that people can use confidence to accurately judge that their memory is correct (Dunlosky & Metcalfe, 2008). The experience that confidence and accuracy are positively related is so pervasive that even the U.S. Supreme Court has ruled in the case of Neil v. Biggers (1972) that highly confident eyewitness identifications are likely to be accurate.

However, there are several interesting caveats to the 'obvious' positive relationship between confidence and accuracy. In some cases, confidence and accuracy can be unrelated (N. Brewer & Wells, 2006; Odinot, Wolters, & van Koppen, 2009; Odinot et al., 2009; Smith, Kassin, & Ellsworth, 1989) and sometimes it can even be negative (Roediger III & DeSoto, 2014; Sampaio & Brewer, 2009).

These unusual relationships typically occur when researchers manipulate perceived confidence: if someone believes that their memory for a specific instance is better than it actually is, the correlation or relationship can become unrelated or negative. For example, using a well-known finding in geography alignment (that participants from North America respond to geography questions as if cities in the south of Europe were to the south of most cities in the United States when, in fact, many of these southern European cities are to the north of many of the cities in the United States), Brewer & Sampaio (2012) were able to find that participants responded with high confidence but less accuracy to these deceptive questions.

However, when researchers manipulate memory itself (e.g., providing different amounts of time to rehearse a to-be-remembered item), the findings are remarkably and consistently positive. For example, Busey, Tunnicliff, Loftus, & Loftus (2000) manipulated how long participants could rehearse faces that had been studied and found that participants' memory and confidence both increased the longer faces could be rehearsed. Roediger & Desoto (2012) suggest that when memory is manipulated, the relationship between confidence and memory almost always have the same relationship.

This study explores the relationship between accuracy and confidence for interruptions. Interruptions occur quite frequently in applied domains (Grundgeiger, Sanderson, MacDougall, & Venkatesh, 2010; Li, Magrabi, & Coiera, 2011) and there is a theoretical description of why interruptions are disruptive (Memory for Goals (MFG); Altmann & Trafton, 2002). According to MFG, interruptions are disruptive because primary-task memories have had a chance to decay during the interruption, and interference with other memories can cause errors (Trafton, Altmann, & Ratwani, 2011; Trafton, Jacobs, & Harrison, 2012). MFG suggests that memories should decay during an interruption. However, people's confidence in their memories after being interrupted should not be impacted; people should be less accurate, but when they are accurate, they should be just as confident as when they are not interrupted. Similarly, when they are inaccurate after an interruption, they should have the same amount of (lower) confidence as when they were not interrupted.

Another aspect of this study was to explore interruption type. One way to make interruptions more disruptive is to make the interruption similar to the primary task (Cutrell, Czerwinski, & Horvitz, 2001). When people have a disruptive interruption, they should be less accurate but also less confident in their answer.

METHOD

Participants

Fifty-six George Mason University psychology undergraduates participated for course credit.

Tasks

Primary Task. The primary task consisted of a simulated stock exchange where participants filled out Buy and Sell orders (Figure 1). This Financial Management Task required participants to select one of four orders to complete based on the rule of buy low and sell high. The four orders at the bottom showed whether a client wanted to buy or sell for any one of three different stocks and their requested price. Participants then needed to compare the requested price to the current price of a stock in the market and determine if the order should be executed at that moment or at a later time.

Once that decision was made the participant would click "Go" to activate that order. They would then fill out 12 widgets on the screen in the order of left to right and top to bottom.

To activate a widget, participants selected a "Start" button on the side of each widget. This would teleport the widget to the bottom middle of the screen so that it became the main focus of the task. Participants would use the active order or the stock market information in the center of the task to complete the widget. When the correct answer was selected the widget would return to its place on the screen.

When all of the widgets were filled correctly the participant would select "Complete" on the bottom right which would "submit" the order. The screen was then occluded by a white mask and a fixation cross in the center and then replaced by a new trial.

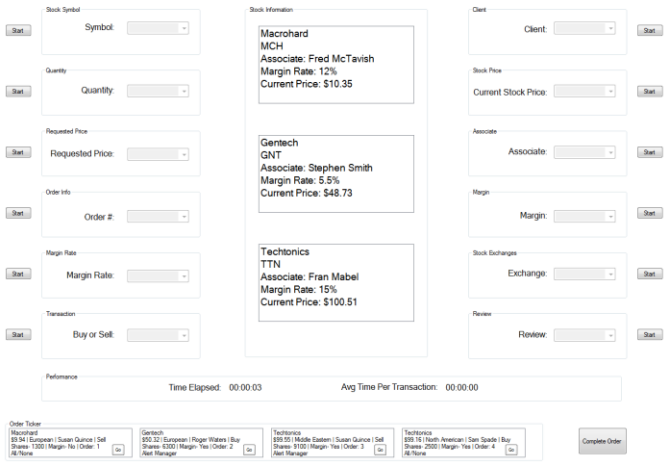

Figure 1. Screenshot of the financial management task

Likert Scale. On occasion participants were presented with a one to six Likert scale question after they clicked the "Start" button for a widget. The question read: "You selected the (name of widget) start button. How certain are you that this was correct?" with 1 being "Uncertain" and 6 being "Certain".

Secondary Task. Participants were also given one of two secondary tasks that would be triggered after correctly selecting information for a widget. The secondary task was either a series of addition problems (Figure 2) or another financial management task (Figure 3) with a different order from the primary task. Each secondary task acted as an interruption that lasted 20 seconds and completely occluded the primary task. Participants were encouraged to complete each math problem or stock order as quickly and as accurately as possible. The background colors of the secondary task were turned a pale yellow to distinguish them as separate from the primary task.

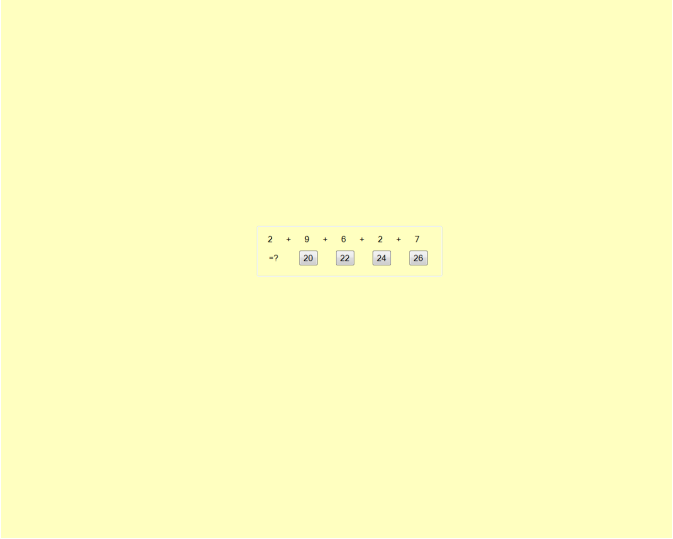

Figure 2. Screenshot of the addition problem secondary task.

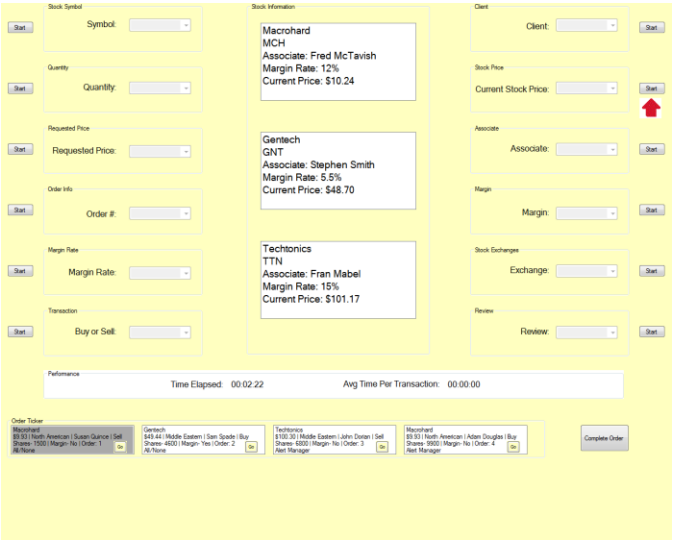

Figure 3. Screenshot of the financial management secondary task. The red arrow indicates the pre-selected widget the participant should begin on.

Design

Participants were assigned to one of two conditions in a 2 (between: math/stock interruption type) x 2 (within: interruption/non-interruption) design. Each participant had a total of 24 interruptions across 16 orders. Of the 16 trials 4 were control trials, 4 had one interruption, 4 had two interruptions, and 4 had three interruptions. This was to reduce the likelihood that participants could prepare for an interruption.

Participants were also presented with 24 Likert questions regarding their confidence at the moment. Half

occurred upon clicking the "Start" button after an interrupted widget and half after clicking the "Start" button after completing a non-interrupted widget. This was to reduce the likelihood that participants would prepare a response to their confidence after every interruption.

Procedure

Participants filled out an IRB approved consent form as well as biographical information. Participants were seated approximately 47 cm from the computer monitor. The task was described using screenshots of the primary and secondary task as well as the Likert scale question. Three additional practice trials were completed with the experimenter present to provide the opportunity for participants to ask clarifying questions about the behavior of the financial management task. Participants could begin once the experimenter left the room and were debriefed and dismissed once finished.

Measures

Behavioral data based on mouse clicks was collected for all participants in addition to screen recordings. Only the behavioral data was analyzed for this study. An accurate response was defined as clicking the next appropriate "Start" or "Complete" button in the task.

RESULTS

Fifty-six participants made a total of 1,344 responses regarding their confidence on the task and 174 procedural errors.

A mixed model ANOVA of interruption type (math/stock interruption) x interruption (interruption / non-interruption) showed a significant difference in accuracy rate and confidence score only for the interruption condition. As a result we focus our analyses on the presence or absence of interruptions and not interruption type.

Accuracy rates were higher when participants were not interrupted (M = 99.25%) than when they were interrupted (M = 74.85%), $F(1,54) = 91.07$, MSE = 16677, $p < .05$, $\eta 2 = .44$. Confidence scores were higher when participants were not interrupted (M = 5.89) than when they were interrupted (M = 5.17), $F(1,55) = 77.75$, MSE 14.71, $p < .05$, $\eta 2 = .36$.

Participants were often accurate and highly confident of their responses. This reduced the variability in the data that could be used to calculate analyses of variance for skewness, kurtosis, and correlations. As a result we will be describing the data at a high-level to show how the two distributions of interrupted and non-interrupted widgets differ.

A chi-square test showed that the distribution of confidence responses was significantly different; $\chi^2(5, N = 56) = 191.24$, $p < .05$). Figure 4 shows the frequency of responses for interruption and no interruption widgets.

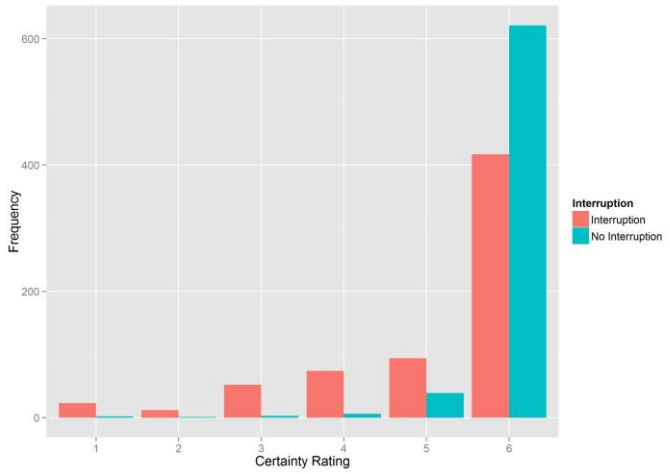

Figure 4. The frequency of certainty scores chosen from the Likert scale question.

To determine the nature of the distribution, the skewness and kurtosis were calculated. The data for both the interruption and no interruption widgets were skewed to the right (Interruption = -1.61; No Interruption = -6.22). Additionally, kurtosis measures suggest that both conditions had a non-Gaussian distribution (Interruption = 1.81; No Interruption = 48.58).

A Pearson's R correlation was calculated to investigate the relationship between confidence and accuracy. Performance was coded as a one for an accurate response and zero for an inaccurate response. Confidence was coded as a six for certain through a one for uncertain. The correlation between confidence and error was .40 for interruption widgets and .50 for no interruption widgets.

To investigate behavior when highly confident, the mean accuracy rate was taken from when participants were highly certain (responded with a "6") for both the interruption and no interruption condition (Figure 5). Participants had significantly higher accuracy when they were not interrupted (M= 99.82%) than when they were interrupted (M = 88.06 %), $F(1,55) = 28.25$, MSE = 3872, $p < .05$, $\eta 2 = .34$.

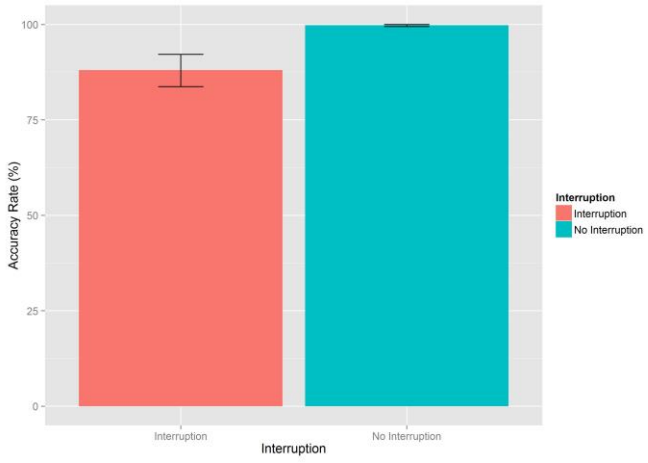

Figure 5. The accuracy rate for interrupted and non-interrupted widgets when participants were certain of their choice. Error bars are 95% confidence intervals.

DISCUSSION

Somewhat surprisingly, we did not find a difference between interruption types in this experiment. We are exploring this effect further in a future study. However, consistent with MFG, we found that interruptions increased errors compared to non-interrupted widgets.

Theoretical

When participants were not interrupted, they were frequently correct and they were quite confident they were correct. In contrast to previous work on memory manipulations, the relationship between confidence and accuracy changed when participants were interrupted.

The strongest evidence for this comes from when participants were most confident: with no interruption they made almost no errors, but after being interrupted their accuracy dropped by over 10%! Thus, the relationship between confidence and accuracy changed after an interruption.

Applied

Interruptions can be particularly egregious in complex applied domains because of the time-critical nature of some tasks. Some examples from the medical field show that interruptions can increase the risk and severity of medication errors (Westbrook, Woods, Rob, Dunsmuir, & Day, 2010) and delay resumption in the critical care environment (Grundgeiger et al., 2010). In light of the finding that professionals in the medical domain can be interrupted nearly 9 times every hour, it seems important to understand the relationship between accuracy and confidence (Chisholm, Dornfeld, Nelson, & Cordell, 2001).

While both accuracy and resumption lag are important, this experiment suggests that the relationship between accuracy and confidence changes after an interruption. This is particularly important because it reduces the likelihood that professionals can accurately gauge whether or not they have resumed appropriately after an interruption. Even when highly confident, participants were nearly seven times more likely to make an error after an interruption, suggesting that confidence was a less reliable marker for accuracy when interrupted then when not interrupted.

ACKNOWLEDGEMENTS

This work was supported by the Office of Naval Research, funding document numbers N0001413WX00002 and N0001413WX20726l. The views and conclusions contained in this document do not represent the official policies of the US Navy.

REFERENCES

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, *26*(1), 39–83.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11.

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, *67*(1), 59–77.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*(1), 26–48.

Chisholm, C. D., Dornfeld, A. M., Nelson, D. R., & Cordell, W. H. (2001). Work interrupted: a comparison of workplace interruptions in emergency departments and primary care offices. *Annals of Emergency Medicine*, *38*(2), 146–151.

Cutrell, E., Czerwinski, M., & Horvitz, E. (2001). Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.418

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. SAGE Publications.

Grundgeiger, T., Sanderson, P., MacDougall, H. G., & Venkatesh, B. (2010). Interruption management in the intensive care unit: Predicting resumption times and assessing distributed support. *Journal of Experimental Psychology: Applied*, *16*(4), 317.

Li, S. Y., Magrabi, F., & Coiera, E. (2011). A systematic review of the psychological literature on interruption and its patient safety implications. *Journal of the American Medical Informatics Association*, amiajnl–2010.

Odinot, G., Wolters, G., & van Koppen, P. J. (2009). Eyewitness memory of a supermarket robbery: a case study of accuracy and confidence after 3 months. *Law and Human Behavior*, *33*(6), 506.

ROEDIGER III, H. L., & DESOTO, K. A. (2012). The Curious Complexity between Confidence and Accuracy in Reports from Memory. *Memory and Law*, 84.

Roediger III, H. L., & DeSoto, K. A. (2014). Confidence and memory: assessing positive and negative correlations. *Memory*, *22*(1), 76–91.

Sampaio, C., & Brewer, W. F. (2009). The role of unconscious memory errors in judgments of confidence for sentence recognition. *Memory & Cognition*, *37*(2), 158–163.

Smith, V. L., Kassin, S. M., & Ellsworth, P. C. (1989). Eyewitness accuracy and confidence: within-versus between-subjects correlations. *Journal of Applied Psychology*, *74*(2), 356.

Trafton, J. G., Altmann, E. M., & Ratwani, R. M. (2011). A memory for goals model of sequence errors. *Cognitive Systems Research*, *12*(2), 134–143.

Trafton, J. G., Jacobs, A., & Harrison, A. M. (2012). Building and verifying a predictive model of interruption resumption. *Proceedings of the IEEE*, *100*(3), 648–659.

Westbrook, J. I., Woods, A., Rob, M. I., Dunsmuir, W. T., & Day, R. O. (2010). Association of interruptions with an increased risk and severity of medication administration errors. *Archives of Internal Medicine*, *170*(8), 683–690.