Interruptions Lead to Improved Confidence-Accuracy Calibration: Response Time as an Internal Cue for Confidence

Nathan Aguiar (naguiar@gmu.edu)
George Mason University
4400 University Dr. Fairfax, VA 22030

Kevin Zish (kzish@gmu.edu)
George Mason University
4400 University Dr. Fairfax, VA 22030

Malcolm McCurry (malcolm.mccurry.ctr@nrl.navy.mil)
Harris
4555 Overlook Ave SW Washington, DC 20375

J. Gregory Trafton (greg.trafton@nrl.navy.mil)
U.S. Naval Research Laboratory
4555 Overlook Ave SW Washington, DC 20375

Abstract

Past research has found that interruptions change the relationship between confidence and accuracy. However, it is unclear how interruptions affect confidence-accuracy calibration. In this study, we used a rule-based procedural task called UNRAVEL and compared confidence-accuracy calibration between interrupted and uninterrupted trials. Results showed that participants were better calibrated in the interruption condition than in the no interruption condition. We interpret this novel effect as a result of changes in the validity of internal cues for confidence between conditions. Specifically, we explore response time as one potential mediating factor.

Keywords: Interruptions, Response Time, Confidence, Accuracy, Calibration

Introduction

The utility of confidence as a valid indicator of memory accuracy has long been a topic of interest to cognitive scientists and forensic professionals. Historically, research conducted on the confidence-accuracy (CA) relationship has yielded mixed results. Earlier studies, for example, concluded that only a modest correlation exists between confidence and accuracy (Sporer 1993; Bothwell, Deffenbacher, & Brigham, 1987). Later research, however, suggested that the weak CA relationship found in earlier studies may be misleading (Juslin, Olsson, & Winman, 1996). Indeed, recent studies using calibration analyses instead of point-biserial correlations have found stronger relationships between confidence and accuracy (Brewer, Keast, & Rishworth, 2002).

While there exist conflicting evidence on the strength of confidence as an indicator of accuracy, there seems to be more consensus on the topic among lay people and professionals: the CA relationship is assumed to be both positive and strong (Roediger III & DeSoto, 2014). Indeed, the Supreme Court even ruled in Neil vs Biggers (1972) that highly confident eyewitness testimony is more likely to be accurate. The disconnect between the existing empirical data and the practices of applied fields signals a need for a more thorough investigation of the predictors of confidence and the factors that moderate the CA relationship.

Identifying the cognitive mechanisms that underlie confidence judgments has been a central mission of confidence researchers. One theory suggests that confidence judgments are inferential in nature and are based on feedback gathered during cognitive processes (Koriat, 1993). Indeed, research has identified several inferential cues for confidence. For example, the strength (Koriat, 1993) and vividness (Brewer, Sampaio, & Barlow, 2005) of a retrieved memory have been shown to be positively related to confidence judgments. In addition, response-latency, or task response time (RT), has been shown to have an inverse relationship with confidence (Weber & Brewer, 2006). It has been proposed that these internal cues contribute to an unconscious feeling-of-knowing which is then translated into a confidence judgment (Koriat, 2000). In this study, we sought to investigate how manipulations of internal cues might affect how people render confidence estimations. Specifically, we sought to identify how within-task changes in RT affect the CA relationship.

Although RT has been shown to predict confidence, the validity of RT as an internal cue for confidence is dependent upon the degree to which it also predicts accuracy (Ackerman & Zalmanov, 2012). Generally, RT is negatively associated with accuracy (Koriat, 2008). There are, however, some situations where RT and accuracy are

positively correlated (e.g. successful tip of the tongue memory), leading to a more mixed relationship.

A common task manipulation that has been shown to affect RTs are task interruptions. Research has shown that interruptions often lead to slower RTs on trials immediately following the interruption (Altmann & Trafton, 2007). The time it takes to generate a task-related response at the end of an interruption is called a *resumption lag* (Trafton, Altmann, Brock, & Mintz, 2003). Memory for Goals (MFG), an activation-based model, provides an explanation for why interruptions lead to slower RTs on trials immediately following an interruption. As implemented in the ACT-R cognitive architecture, when a memory retrieval is made, the memory item that is most active at that time is returned (Anderson, 1982; Altmann & Trafton, 2002). However, activation is subject to decay over time. Therefore, interruptions are disruptive because they lead to the decay of memory items. Activation decay in turn leads to more retrieval failures and longer RTs.

Because interruptions affect RT and RT has been shown to be a cue for confidence, it is reasonable to predict that interruptions may lead to changes in the CA relationship. Indeed, past research has shown that interruptions affect the CA relationship by decreasing accuracy at the highest level of confidence (Zish, Hassanzadeh, McMurry, & Trafton, 2015; Aguiar, Zish, McCurry, & Trafton, 2016). Although interruptions have been shown to change the relationship between confidence and accuracy, it is unclear how interruptions affect confidence-accuracy calibration.

Calibration can be defined as the degree to which confidence ratings match objective probabilities. For example, a person is deemed to be perfectly calibrated when their responses with 100% confidence are 100% accurate, their responses with 90% confidence are 90% accurate, and so forth. A strong CA relationship is characterized by strong calibration.

In confidence research, calibration is the preferred measure for studying the CA relationship (e.g. Bjorkman, 1994; Luna & Martín-Luengo, 2012; Brewer, Keast, Rishworth, & Ackerman, 2002; Baranski & Petrusic, 1994). Calibration is most often described using three measures: the calibration statistic (C), calibration curves, and over/underconfidence (O/U). The calibration statistic ranges from 0 to 1 and is calculated as the weighted square difference between accuracy and confidence at each level of confidence. Perfect calibration is achieved when C equals 0, thus the lower the C statistic, the better the calibration. A calibration curve is a visual representation of calibration and is created by plotting accuracy across each level of confidence. Over/ underconfidence is calculated simply by taking the difference between mean confidence and mean accuracy. A positive result is interpreted as overconfidence, a negative result as underconfidence, and perfect calibration is achieved when O/U equals zero (for a review of these measures, see Baranski and Petrusic (1994)).

The purpose of this study was twofold: to describe how interruptions affect calibration and to investigate RT as a potential internal cue for confidence. We hypothesized that confidence would be negatively correlated with RT regardless of whether an interruption occurs. However, we expected that calibration would be best in the condition that yielded the strongest relationship between accuracy and RT.

Methods

Participants

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Sixty-three George Mason University undergraduate psychology students participated for course credit. After analyzing the first 50 participants, we added an additional 13 participants. Of the 63 participants, one was excluded due to an outlying accuracy score. An outlying accuracy score was defined as one differing by more than 3.5 standard deviations from the mean accuracy across all participants over at least two of the four blocks of the task. An accuracy score differing by more than 3.5 standard deviations from the mean was deemed to be signal that the participant either misunderstood or disengaged from the task.

Primary Task Calibration analyses require a large number of observations per participant. As such, the primary task needed to result in a high volume of trials over a single session. The chosen primary task (adapted from Altmann, Trafton, & Hambrick, 2015) is defined by the acronym UNRAVEL and yields a substantial number of trials per session.

Each letter of UNRAVEL represents a step in a cyclical procedure. The letters indicate the order in which steps should be performed. For example, the U step is to be completed first, followed by the N step, then the R step and so on. Once the participant completes the L step he or she returns to the U step and continues the sequence. The goal of the task is to correctly complete each step of the task in the prescribed order and avoid skipping or repeating steps. A trial was defined as the completion single step of the UNRAVEL task.

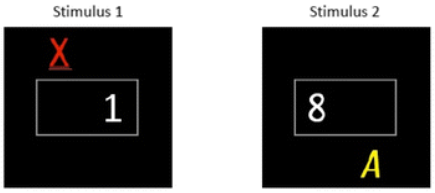
The task rules and candidate responses are illustrated in Figure 1. Participants could access the choice rules for the UNRAVEL task any time the stimulus was present by holding down a prescribed keyboard shortcut.

Each stimulus includes two characters: a digit and a letter. Each character possesses unique characteristics. For example, one character is colored red or yellow, one is either underlined or italicized, and one is located either above or below the gray box.

Each letter of UNRAVEL mnemonically relates to the choice rule for that step and each step requires a two-alternative forced choice related to one characteristic of the stimulus. For example, the U step asks whether the character is underlined or in italics, the N step whether the letter is near to or far from the start of the alphabet, and so forth.

After the completion of each step, a new stimulus would appear.

(a) Sample stimuli for UNRAVEL task:



(b) Choice rules and candidate responses for UNRAVEL task, and responses to the stimuli in (a):

Step	Candidate responses		Choice rules	Responses to sample stimuli Stimulus 1	Stimulus 2
U	u	i	character is Underlined or in Italics	u	i
N	n	f	letter is Near to or Far from start of alphabet	f	n
R	r	y	character is Red or Yellow	r	y
A	a	b	character is Above or Below the box	a	b
V	v	c	letter is Vowel or Consonant	c	v
E	e	o	digit is Even or Odd	o	e
L	l	m	digit is Less than or More than 5	l	m

Figure 1. (a) two stimuli from the UNRAVEL task. (b) choice rules and candidate responses for the unravel task.

Secondary Task To increase cognitive workload during the UNRAVEL task, participants were also required to retain a two or four-letter code in memory. Codes were presented at the beginning of each block and immediately following the output of a previous code. The codes were random, non-repeating permutations of the consonants in the candidate responses of the UNRAVEL task (i.e. "F", "R", "Y", "B", "V", "C", "L", and "M").

The procedure was as follows: occasionally after the completion of a step, the UNRAVEL stimulus would be masked by a screen containing an output box and instructions to enter the most recently presented two or four-letter code. After entering the code and pressing the return key, a new two or four-letter code would appear on the screen for four seconds before disappearing and revealing the UNRAVEL stimulus.

After returning to the UNRAVEL stimulus, participants would attempt to resume the UNRAVEL sequence from where they left off before the presentation of the output box screen. Because the output box screen and the new code presentation masked the UNRAVEL stimulus, this sequence was considered an interruption. Thus, the UNRAVEL step immediately following the interruption was considered the interruption trial. The frequency of interruptions was randomized with an average of six steps between each interruption. Half of the interruptions required outputting a two-letter code and the other half a four-letter code.

Thus, in addition to completing the steps of the UNRAVEL task, participants also needed to maintain the new code in memory until prompted to output the code several steps later. Pilot testing with a different interruption task had revealed a ceiling effect; thus, the purpose of the code memorization task was to increase participants' mental workload during the UNRAVEL task and decrease accuracy in the no interruption condition.

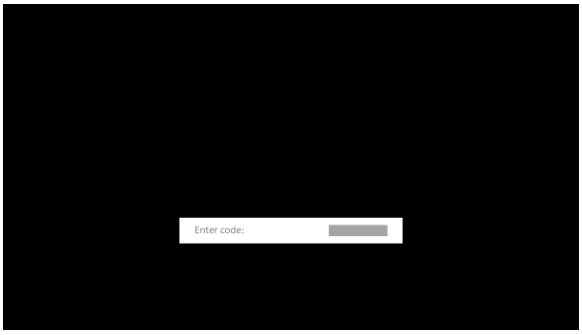


Figure 2. A screenshot of the code output box screen

Confidence Question Occasionally after the completion of an UNRAVEL step, participants would be prompted to rate their confidence in the accuracy of the last step they performed. Specifically, they were presented the prompt: "How confident were you that you just chose the correct step during the UNRAVEL task? Enter your choice on a scale from 1 to 6, with 1 being least confident, and 6 being most confident." Participants indicated their level of confidence by typing their response into a text field. During this time, the UNRAVEL stimulus was masked. After submitting their response, the mask would be removed and the UNRAVEL stimulus would be visible again. Participants would then pick up on the next UNRAVEL step from where they left off before the appearance of a confidence question.

Procedure

Practice Each session began with an introduction to the UNRAVEL task. Participants were given a step-by-step walkthrough of the UNRAVEL sequence and each choice rule was reviewed. Screenshots were used to explain the task and participants were encouraged to ask questions to ensure the participant understood all components of the task.

Each participant was required to successfully complete a practice session before beginning the main task. During the practice session, the participant was exposed to all aspects of the task. The experimenter was present during the practice session to ensure the participant understood all aspects of the task and to help if necessary. After the completion of the practice session the participant was then instructed to begin the main task as soon as the researcher left the room. When all trials of the main task were completed, the participant was debriefed and thanked.

Blocks The task was composed of four blocks. Each block contained fifteen confidence questions for a total of sixty confidence questions per study session. Confidence questions were presented in three situations: immediately following the first UNRAVEL step after the output of a two-letter code (short interruption condition), immediately following the first UNRAVEL step after the output of a four-letter code (long interruption condition), and occasionally after UNRAVEL steps not preceded by an interruption (no interruption condition). In total, each

participant answered twenty confidence questions in each condition (i.e. short interruption, long interruption, no interruption).

Feedback After the completion of each block, participants were presented with feedback on their performance over that block. Accuracy was computed as the percentage of UNRAVEL trials during which the response and step were both answered correctly. To prevent condition-specific adjustments, participants were provided a single accuracy score averaged across all trials. If the score was above 90%, the participant was asked to go faster. If the score was below 70%, the participant was asked to be more accurate. The 90% and 70% thresholds were based on those used in Altmann, Trafton, and Hambrick (2015) and were used to deter participants from adopting a bias toward accuracy over speed or vice versa.

Participants were also presented with feedback on their performance on the secondary task. Secondary task feedback was computed as the percentage of typed codes that correctly matched the most recently presented code. Like the primary task feedback, participants were given a single accuracy score averaged across all interruption trials. Participants were encouraged to maximize accuracy on both the UNRAVEL task and the secondary code memorization task.

Measures

Behavioral data and confidence ratings were analyzed in this study. Behavior was measured as accuracy and response time on UNRAVEL steps.

Results

From 62 participants, 21,699 UNRAVEL trials were completed and 3,566 confidence responses were collected. Of the 21,699 UNRAVEL trials, 308 trials (about 1% of total trials) were excluded due to outlying RT scores. An outlying RT score was determined to be any RT score that exceeded 3.5 standard deviations from the mean RT. RTs that exceeded 3.5 standard deviations from the mean were deemed to be indicative of temporary task disengagement. The remaining 21,391 UNRAVEL trials were analyzed.

No significant difference in accuracy was found between short (two-letter codes) and long interruptions (four-letter codes). As such, the two interruption types were combined into a single interruption condition in the following analyses.

To assess the effect of interruptions on performance, a one way repeated-measures ANOVA was conducted to compare the effect of condition (interruption/no interruption) on response accuracy. As expected, participants were less accurate on trials immediately following an interruption (M = 60.42%) than on non-interrupted trials (M = 96.42%), $F(1,60) = 459.80$, MSE =.01, $p < .05$, $\eta 2 = .75$. The high accuracy observed in the no interruption condition suggests that participants knew the task well.

Confidence-Accuracy Calibration

Due to its strong presence in the confidence and decision-making literature, our analysis focused primarily on calibration measures. To assess calibration, we first converted the six-point confidence scale to probabilities. Thus, a confidence response of 1 was converted to 1/6 = .16 or a 16%, a response of 2 was converted to 2/6 or .33 or 33%, and so on.

To measure calibration and assess the difference between the interruption and no interruption condition, we computed three calibration assessments: the calibration curve, the calibration statistic (C), and over/underconfidence (O/U).
The calibration curve is plotted in Figure 3 and shows accuracy across confidence levels for interruption and no interruption trials. Calibration curves are used to visualize calibration across confidence levels. The dashed line represents perfect calibration; thus, the nearer points are to the dashed line, the better the calibration. Here, the interruption condition appears to more closely track the perfect calibration line than the no interruption condition.

The C statistic represents overall calibration with a score of 0 representing perfect calibration and a score of 1 representing the worst possible calibration. Two C scores were calculated for each participant, one for interruption trials and one for no interruption trials. A one way repeated-measures ANOVA was conducted to assess the difference between C statistic scores for interruption and no interruption trials. Results indicated that participants were better calibrated on interruption trials (M = .05), $F(1,60) = 6.28$, MSE =.005, $p < .05$, $\eta 2 = .04$. than on no interruption trials (M = .09) (Table 1).

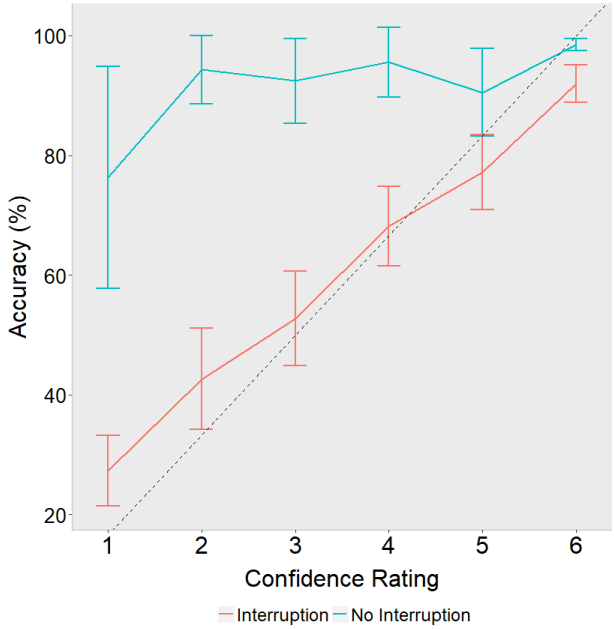


Figure 3. Accuracy for interruption and no interruption trials across each level of confidence. Dashed line represents perfect calibration. Error bars are 95% confidence intervals.

The final calibration measure calculated was over/underconfidence. Like the C statistic, an O/U score of 0 represents perfect calibration. However, unlike the C statistic O/U scores range from -1 to 1 with -1 indicating total underconfidence and 1 indicating total overconfidence. Two O/U scores were calculated for each participant, one for interruption trials and one for no interruption trials. Results from a one way repeated-measures ANOVA showed that participants were better calibrated during the interruption condition ($M = -.01$), $F(1,60) = 47.78$, $MSE = .01$, $p < .05$, $\eta 2 = .17$ than the no interruption condition ($M = -.16$). The negative O/U value observed in no interruption trials indicates that participants were biased toward underconfidence when they were not interrupted (Table 1).

Table 1: Calibration (C) and Over/Underconfidence (O/U)

Condition	C	O/U
No Interruption	.09	-.16
Interruption	.05	-.01

Response Time

To measure the relationship between RT and confidence, we calculated two point-biserial correlations for each participant, one for each condition. The mean correlation between RT and confidence across all participants in the interruption condition was -.28. A mean correlation of -.21 was found in the no interruption condition. A one way repeated-measures ANOVA revealed no significant difference in the strength of the confidence-RT relationship between conditions (Table 1).

A similar procedure was performed to assess the relationship between RT and accuracy. After calculating a point-biserial correlation for each condition across all participants, we conducted a one way repeated-measures ANOVA to compare mean correlations between the interruption and no interruption conditions. Results showed a non-significant, marginally stronger correlation between RT and accuracy in the interruption condition ($M = -.17$), $F(1,60) = 3.53$, $MSE = .03$, $p = .065$, $\eta 2 = .03$ than in the no interruption condition ($M = -.11$) (Table 2).

Table 2: Response Time Correlations

Condition	RT: Accuracy	RT: Confidence
No Interruption	- .11	-.21
Interruption	- .17	-.28

Discussion

The purpose of this study was to investigate the effect of interruptions on CA calibration and to assess how changes in task RT affect the validity of RT as an internal cue for confidence. Although interruptions were associated with lower accuracy overall, results indicated that participants were better calibrated after an interruption than on uninterrupted trials. When participants were not interrupted, they tended to be more accurate but less calibrated. As evident in the negative O/U value, when uninterrupted, participants tended to exhibit a clear bias towards underconfidence.

Notably, our results replicate the previous finding that interruptions lead to reduced accuracy at the highest level of confidence. When participants indicated that they were entirely confident in their choice, they were significantly more likely to have made an error after an interruption than when they were not interrupted. This is particularly concerning considering that highly confident eyewitness testimony is often considered more trustworthy (e.g. Cutler, Penrod, & Dexter, 1990).

To our knowledge, this is the first study to show that interruptions lead to improved CA calibration. We hypothesized that the difference in calibration between interruption conditions is due to changes in the validity of internal cues for confidence. As suggested by Koriat (2000) internal cues manifest as an unconscious feeling-of-knowing which is then used to render a confidence judgment. Here, we investigated RT as one potential internal cue. We proposed that a portion of confidence judgments is determined by the amount of time it takes to retrieve an answer from memory, or in this case, the amount of time it took participants to determine where they left off on the UNRAVEL sequence. This hypothesis is supported by past research identifying RT as a predictor of confidence (e.g. Ackerman & Zalmanov; Weber & Brewer, 2006). Indeed, our results show a modest relationship between RT and confidence in both conditions. However, the effectiveness of RT as an inferential cue for confidence is dependent upon the degree to which it also predicts accuracy. Although the correlation between RT and accuracy was not significantly different between conditions, our results indicate a trend toward significance. The small effect size, however, leaves open the possibility that there may be other factors that account for some of the remaining variance between conditions. Future research should be conducted to identify other mediating factors in addition to RT.

In conclusion, our results replicate previous findings suggesting that interruptions change the relationship between confidence and accuracy. We extend upon previous research by showing that the change in the CA relationship after an interruption is characterized by better calibration. In all, our findings suggest that even momentary interruptions can impact the way in which confidence judgments are rendered.

Acknowledgments

This work was supported by a grant to GT from ONR. The views and conclusions contained in this document do not represent the official policies of the U.S. Navy.

References

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review, 19*(6), 1187-1192.

Aguiar, N., Zish, K., McCurry, JM., & Trafton, JG. (2016). Interruptions Reduce Performance across All Levels of Signal Detection When Estimations of Confidence are Highest. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*, 254-258.

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

Altmann, E., & Trafton, M. (2007). Timecourse of recovery from task interruption: Data and a model. *Psychonomic Bulletin & Review, 14*(6), 1079-1084.

Altmann, E. M., Trafton, J. G., & Hambrick, D. Z. (2014). Momentary interruptions can derail the train of thought. *Journal of Experimental Psychology: General, 143*, 215–226. doi: 10.1037/a0030986

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369.

Baranski, J., & Petrusic, V. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55*(4), 412-428.

Bjorkman, Mats. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior & Human Decision Processes, 58*(3), 386.

Bothwell, R., Deffenbacher, K., Brigham, J., & Guion, R. (1987). Correlation of Eyewitness Accuracy and Confidence: Optimality Hypothesis Revisited. *Journal of Applied Psychology, 72*(4), 691-695.

Brewer, N., Caon, A., Todd, C., Weber, N., & Wiener, Richard L. (2006). Eyewitness Identification Accuracy and Response Latency. *Law and Human Behavior, 30*(1), 31-50.

Brewer, N., Keast, A., Rishworth, A., & Ackerman, Phillip L. (2002). The Confidence-Accuracy Relationship in Eyewitness Identification: The Effects of Reflection and Disconfirmation on Correlation and Calibration. *Journal of Experimental Psychology: Applied, 8*(1), 44-56.

Brewer, Sampaio, & Barlow. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language, 52*(4), 618-627.

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior*, *14*, 185–191. doi: 10.1007/BF01062972

Juslin, Olsson, & Winman. (1998). The Calibration Issue: Theoretical Comments on Suantak, Bolger, and Ferrell (1996). *Organizational Behavior and Human Decision Processes, 73*(1), 3-26.

Koriat, Asher. (1993). How Do We Know That We Know? The Accessibility Model of the Feeling of Knowing. *Psychological Review, 100*(4), 609-39.

Koriat, A. (2000). The Feeling of Knowing: Some Metatheoretical Implications for Consciousness and Control. *Consciousness and Cognition, 9*(2), 149-171.

Koriat, Asher. (2008). Subjective Confidence in One's Answers: The Consensuality Principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 945-959.

Luna, K., & Martín-Luengo, B. (2012). Confidence–Accuracy Calibration with General Knowledge and Eyewitness Memory Cued Recall Questions. *Applied Cognitive Psychology, 26*(2), 289-295.

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

Sporer, S., & Schmitt, N. (1993). Eyewitness Identification Accuracy, Confidence, and Decision Times in Simultaneous and Sequential Lineups. *Journal of Applied Psychology, 78*(1), 22 33.

Trafton, Altmann, Brock, & Mintz. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human – Computer Studies, 58*(5), 583-603.

Zish, K., Hassanzadeh, S., McMurry, J. M., & Trafton J. G. (2015, September). Interruptions can Change the Perceived Relationship between Accuracy and Confidence. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 230-234. SAGE Publications.