

Elevated Stress Impairs the Accuracy of Eyewitness Memory but Not the Confidence–Accuracy Relationship

Kathy Pezdek, Erica Abed, and Anne Cormia
Claremont Graduate University

Although numerous studies have identified factors that affect eyewitness identification accuracy, recent studies report that many of these factors do not affect the accuracy of high-confidence identifications. This is critical because legal cases are more likely to be prosecuted if they involve high-confidence eyewitnesses. Using a confidence–accuracy characteristic (CAC) analysis, we explored whether stress affects the accuracy of high-confidence identifications. In two experiments, people viewed faces followed by an old/new recognition-memory test and provided confidence ratings. Stress was manipulated by pairing a low- or high-valence image with each studied face. Identification accuracy was higher in the low- than high-stress condition, yet the proportion correct for high-confidence positive identifications was similar in the two stress conditions. Elevated stress impairs eyewitness identification accuracy overall. However, the results of this study suggest that confidence is a better predictor of recognition-memory accuracy than is stress even though confidence alone is still an imperfect predictor.

Public Significance Statement

The results of this study suggest that although elevated stress impairs eyewitness identification accuracy *overall*, eyewitnesses may be equally likely to be correct in their high-confidence identifications regardless of their stress level at the time of the event. Stressed eyewitnesses do not make good eyewitnesses, but they are generally aware of this and adjust their confidence ratings accordingly.


Keywords: stress, eyewitness memory, confidence–accuracy relationship, face recognition memory, metamemory

Many variables that influence the accuracy of eyewitness memory do not, in fact, affect the accuracy of memory judgments made with high levels of confidence. Variables reported not to affect the accuracy of eyewitness identifications at high levels of confidence include same- compared to cross-race faces (Nguyen, Pezdek, & Wixted, 2017); physical distance (Lindsay, Semmler, Weber, Brewer, & Lindsay, 2008); retention interval, exposure duration, and divided attention (Palmer, Brewer, Weber, & Nagesh, 2013;

Wixted, Read, & Lindsay, 2016); and the presence of a weapon (Carlson, Dias, Weatherford, & Carlson, 2017). Thus, although overall eyewitness identification accuracy is poor under these specific conditions (Deffenbacher, Bornstein, McGorty, & Penrod, 2008; Wells, Memon, & Penrod, 2006; for a review of this research also see National Research Council, 2014), the results of recent research suggest that observers who do make an identification are metacognitively able to calibrate their subjective ratings of confidence to account for these poorer memory conditions. This research suggests that identification accuracy is not affected by these variables when recognition confidence is high and collected appropriately (Wixted & Wells, 2017).

Although it is important to understand the cognitive conditions under which eyewitness identifications are generally more or less likely to be accurate, in fact, from a forensic point of view, the eyewitnesses who are most likely to testify in court are those who make high-confidence identifications (Wells, Ferguson, & Lindsay, 1981), and high-confidence identifications are more likely to persuade a jury (Cutler, Penrod, & Stuve, 1988). Thus, looking at factors that specifically affect *high-confidence* identifications is critically important. An experimental technique that facilitates this type of comparison relies on a confidence–accuracy characteristic (CAC) analysis (cf. Juslin, Olsson, & Winman, 1996; Mickes, 2015), in which the accuracy rate at each level of confidence is

This article was published Online First July 30, 2020.

 Kathy Pezdek, Erica Abed, and Anne Cormia, Department of Psychology, Claremont Graduate University.

Kathy Pezdek developed the study concept and drafted the research design in consultation with Erica Abed and Anne Cormia. Erica Abed analyzed and helped interpret the data. Kathy Pezdek drafted the manuscript, and Erica Abed participated significantly in the revision process. Anne Cormia built the Qualtrics survey, collected the data, and conducted some of the analyses. All authors read and approved the final manuscript. This experiment was formally preregistered and can be accessed at <https://osf.io/7j4nd/>. Upon acceptance of this article, anonymized data will be made publicly available on the Open Science Framework.

Correspondence concerning this article should be addressed to Kathy Pezdek, Department of Psychology, Claremont Graduate University, Claremont, CA 91711. E-mail: Kathy.Pezdek@cgu.edu

assessed. With a CAC analysis, the accuracy of high-confidence judgments can be compared between experimental conditions.

In this study, we focused on the effects of stress on eyewitness identification accuracy and metamemory judgments of confidence. It is important to understand the effects of stress on eyewitness memory primarily because of the forensic relevance of this variable. When an eyewitness observes a perpetrator in a criminal setting—whether the eyewitness is a victim or a bystander—this frequently occurs under a high level of stress. Although an extensive research literature has generally supported the detrimental effect of stress on memory (for a review see Wolf, 2009), this complex effect is moderated by a number of task-specific variables (Bergmann, Rijpkema, Fernández, & Kessels, 2012). Nonetheless, in a meta-analytic review, Deffenbacher, Bornstein, Penrod, and McGorty (2004) reported that high stress consistently impairs eyewitness identification accuracy. However, there has been scant research on the effect of stress on metamemory judgments of confidence, and the relevant findings have been inconsistent. In this research, we specifically examine how stress affects the accuracy of eyewitness identifications made with a high level of confidence.

Two predictions can be made regarding whether the proportion correct for high-confidence identifications would differ between faces encoded under high- versus low-stress conditions. First, if people are aware that high stress impairs the accuracy of their memory and are metacognitively able to calibrate their subjective ratings of confidence to account for this poorer memory condition, then when a high rating of confidence is provided, the proportion correct for identifications should not differ between the high- and low-stress conditions. If this calibration occurs, it would be likely that high-confidence judgments would be provided more cautiously in the high-stress than the low-stress condition, such that when high-confidence judgments are provided, the proportion correct is likely to be high. This is the same interpretation offered in previous studies that have reported similar levels of accuracy for high-confidence identifications across conditions known to affect eyewitness memory, including, for example, same- compared to cross-race faces (Nguyen et al., 2017) and retention interval, exposure duration, and divided attention (Palmer et al., 2013).

Alternatively, if people are *not* aware that high stress impairs the accuracy of their memory and are not metacognitively able to calibrate their subjective ratings of confidence to account for this relatively poorer memory condition, then when a high rating of confidence is provided, the proportion correct for identifications would be predicted to differ between the high- and low-stress conditions. This, in fact, is the finding that was recently reported by Pezdek, Abed, and Reisberg (2020) regarding the effect of marijuana on the confidence–accuracy (CA) relationship. In this study, marijuana users were randomly assigned to a marijuana or control condition and participated in an old/new face-recognition-memory test with confidence ratings provided for each judgment. Consuming marijuana reduced discrimination accuracy (d' ; Cohen's $d = .47$), and the proportion correct for positive identifications, even at the high-confidence level, was significantly lower in the marijuana than the control condition. And as predicted, although marijuana significantly impaired face-recognition memory, 37% of participants responded erroneously that marijuana had either no effect or a beneficial effect on their memory. It might generally be the case that people are less aware of the effect of

state variables (e.g., drugs, alcohol, stress, sleep) on their memory than the effect of situational variables (cross-race faces, brief exposure time, retention interval, etc.). The current study tested this hypothesis as well.

Two recent studies have examined the effect of stress on eyewitness memory and metamemory judgments of confidence, and their findings are inconclusive. Sauerland et al. (2016) had participants view a mock crime after having been exposed to a high- or low-stress condition for 15 min. The high-stress condition was the Maastricht Acute Stress Test (MAST), in which people immerse one hand in an ice-water bath and perform difficult mental calculations. Lukewarm water and an easy counting task were used in the low-stress condition. They reported that in a face-recognition-memory test administered 1 week later, accuracy in identifying the perpetrator from a lineup did not differ between the high- and low-stress conditions. In an exploratory analysis, the CA relationship did not differ between conditions either. However, given the absence of an effect of stress on recognition memory, no conclusions about the effect of stress on the CA relationship can be made from this study. Another limitation of this study is that the researchers did not conduct a CAC analysis to examine accuracy at each level of confidence.

More recently, Davis, Peterson, Wissman, and Slater (2019) had participants view a sequence of faces while experiencing high stress (a cold pressor task in which a hand is submerged in ice water) or low stress (the same task but with room-temperature water). In two experiments, recognition memory was more accurate (higher d' values) in the low-stress than the high-stress condition. However, although the authors concluded that they observed the same CA relationship for identifications in both stress conditions, in fact, this is not true when focused on high-confidence judgments. Whereas the proportion correct did not differ between the high- and low-stress conditions for the highest-confidence identifications in Experiment 1 (see their Figure 1), this difference was reported to be significantly different in Experiment 2 (see their Figure 2). Specifically, in Experiment 2, the proportion correct for high-confidence identifications was significantly greater in the low-stress than the high-stress condition. This makes the results of their study inconclusive. Another issue with their study, and one that they pointed out, is that the source of the stress in their study was not intricately linked to the faces being studied, and this may have mitigated the effect of stress on cognitive processing of the faces. We addressed these concerns in our study.

In this study, we assessed whether stress affects eyewitness discrimination accuracy as well as confidence-specific accuracy, and we specifically focused on the proportion correct for high-confidence judgments. Note that confidence-specific accuracy is independent of discrimination accuracy. For reference, see Nguyen et al. (2017, Table 1), where the difference between discrimination accuracy (receiver operating characteristic [ROC] or d') and confidence-specific accuracy (CAC) is illustrated. It is also important to note that as in previous research on the CA relationship, in this study, we focused on positive identification accuracy (i.e., accuracy of “old” responses) because (a) only witnesses who identify a suspect from a lineup are likely to testify in court, and (b) researchers have reported a strong relationship between confidence and accuracy for “choosers” (e.g., “old” responses) but a weaker relationship between confidence and accuracy for “non-choosers” (e.g., “new” responses; Nguyen et al., 2017; Sporer,

Penrod, Read, & Cutler, 1995; Wixted, Mickes, Clark, Gronlund, & Roediger, 2015).

Experiment 1

In Experiment 1 of the current study, participants viewed 24 target faces followed by an old/new recognition-memory test, with confidence ratings (1–5) collected immediately after each old/new recognition judgment. This provided optimal conditions for evaluating the CA relationship. The old/new recognition procedure is one in which faces are presented one at a time. This test is thus more similar to a showup than a lineup. A showup is a standard police eyewitness identification procedure in which an eyewitness is presented with a suspect who has been apprehended shortly after a crime. The police want to know if the witness recognizes the suspect, yes or no. Thus, a showup is a real-world old/new recognition procedure. Although lineups are reported to result in higher discrimination accuracy than showups (Wixted & Mickes, 2015), both procedures are commonly implemented and are thus forensically relevant.

Moreover, given that researchers have reported similarly strong relationships between confidence and accuracy using both old/new (e.g., showups; Nguyen et al., 2017) and lineup procedures (Wixted, Mickes, Dunn, Clark, & Wells, 2016), the type of procedure per se is unlikely to influence how stress affects the CA relationship.

In Experiment 1, we manipulated stress by presenting each target face alongside an image from the International Affective Picture System (IAPS). The IAPS is a database of images developed by the National Institute of Mental Health Center for Emotion and Attention at the University of Florida (Lang, Bradley, & Cuthbert, 2008). All images have been normed for valence (on a continuum from positive to negative) and arousal. The IAPS has been reliably used to induce stress responses in psychological research (Bradley, Greenwald, Petry, & Lang, 1992; Erk et al., 2003), including face processing (Xie & Zhang, 2016). In Experiment 1, the images selected for the high-stress condition were negative in valence and high in arousal; the images selected for the low-stress condition were neutral in valence and low in arousal. The level of stress was manipulated between subjects. Importantly, the source of the stress was intricately linked to each target face in this study; each target face was presented simultaneously alongside a unique IAPS image that was randomly paired with it.

Method

Participants and design. Participants were recruited using Amazon's Mechanical Turk (via TurkPrime; Litman, Robinson, & Abberbock, 2017) and qualified for participation if they lived in the United States and were at least 18 years old. A G*Power analysis (see Faul, Erdfelder, Lang, & Buchner, 2007) determined that for this design, the minimum sample size needed to have a 90% chance of detecting an effect that exists for the main effect of stress¹ is 98 (Cohen's $d = .60$, $\alpha = .05$). However, we recruited a total of 123 participants, anticipating that a number of participants would violate the exclusion criteria. The final sample ($N = 106$) included 56 participants in the high-stress condition and 50 in the low-stress condition.² The mean age did not differ between the high-stress (mean [M] = 38.1 years, $SD = 11.85$, range = 22–67)

and the low-stress condition ($M = 37.5$ years, $SD = 12.18$, range = 19–64). Both groups were primarily male (male = 57%) and were primarily Caucasian (76%).

This study used a two-independent-groups design; half of the participants were randomly assigned to the high-stress (negatively valenced) condition and half to the low-stress (neutral) condition. The research was reviewed according to the Claremont Graduate University Institutional Review Board (IRB) procedures for research involving human subjects.

Materials and procedure. The face stimuli were 48 White male faces selected from a database of faces used by Meissner, Brigham, and Butz (2005). This database includes two different headshots of each person: (a) smiling and wearing a casual shirt (used as study stimuli) and (b) neutral facial expression and dressed in a maroon-colored shirt (used as test stimuli). We selected an additional 10 study and 10 test faces for the practice presentation and practice test phases.

Forty-eight (24 high-stress and 24 low-stress)³ images were sourced from the IAPS (Lang et al., 2008). IAPS images are normed on arousal and valence. *Arousal* refers to physical and mental activation and varies from 1 (lowest activation) to 9 (highest activation). *Valence* refers to the intrinsic positivity or negativity of an image and varies from 1 (most negative) to 9 (most positive). Mean arousal was significantly greater for the high-stress (M arousal = 6.47, $SD = 0.33$) than the low-stress images (M arousal = 2.96, $SD = 0.65$), $t(66) = 28.10$, $p < .001$, $d = 6.81$. The mean valence was also significantly more negative for the high-stress (M valence = 2.34, $SD = 0.51$) than the low-stress images (M valence = 5.00, $SD = 0.08$), $t(66) = 29.80$, $p < .001$, $d = 7.23$. The low-stress IAPS images were specifically selected to be neutral in Experiment 1.

Participants were randomly assigned to either the high-stress or the low-stress condition. In the presentation phase, participants were presented 24 target slides in a random order for 8 s each. Each target slide contained an IAPS image (high or low stress) presented alongside a White male face. Half of the target faces appeared on the right side, and half appeared on the left side of the screen, and the position of the faces was counterbalanced so that across participants, each face appeared equally often on the left and right side of the screen. Examples of two target slides are presented in Figure 1, with each face paired with a neutral stress image. Examples of the high-stress images included color photo-

¹ Davis et al. (2019, Experiment 1) reported a Cohen's d of .95 for the effect of stress on d' and a Cohen's d of .70 for the effect of stress on confidence. We estimated power based on a smaller effect size than they reported, Cohen's $d = .60$.

² OSF preregistered exclusion criteria for Experiment 1 that were violated were as following: (a) responded to more than 95% of the test items with either (i) a recognition response of only "old" or only "new" or (ii) the same confidence rating (3 excluded); (b) univariate (± 2 standard deviations from the mean) or multivariate outliers (p value associated with Mahalanobis distance $< .01$) on measures of criterion or d' (2 excluded); (c) completion time more than 2 standard deviations from the total mean completion time (8 excluded); (d) responded with less than a "4" to the question: "Complete this sentence by selecting the most appropriate option: "I gave this study _____ attention" (1 = "almost none of my" to 5 = "my full" (3 excluded); (e) indicated that they were interrupted during the experiment (1 excluded).

³ We selected an additional 10 high-stress images and 10 low-stress images for the practice presentation phase.

graphs of (a) a bloody severed clenched hand with the wrist bone visible and (b) an image of a human body, from chin to upper torso, with the throat slit.

Participants were instructed to try to remember both the image and the face because both would be important later in the study. An old/new recognition-memory test followed, in which participants viewed 48 test faces (24 old and 24 new). The test faces were presented in a different random order for each participant, and the 48 faces were counterbalanced so that across participants, each face served approximately equally often as a target (old) and foil (new) face. Participants were instructed to respond “old” or “new” to each test face and to indicate their confidence in each response on a scale of 1 (*completely guessing*) to 5 (*absolutely sure I’m correct*). To familiarize participants with the tasks, there were 10 practice presentation items prior to the presentation phase and 10 practice test items prior to the test phase.

Finally, at the beginning and end of the experiment, participants were asked, “At this moment, how stressed do you feel on a scale from 1 (*not at all stressed*) to 7 (*very high level of stress*)?”⁴ At the end of the experiment, participants were also asked to indicate whether they believed that viewing stressful materials affected their memory on a scale from 1 (*significant beneficial effect*) to 5 (*significant detrimental effect*).

Results and Discussion

As a manipulation check, we assessed self-report ratings of stress at the beginning and end of the experiment by asking, “At this moment, how stressed do you feel on a scale from 1 (*not at all stressed*) to 7 (*very high level of stress*)?” We conducted a 2 (high or low stress) \times 2 (time) mixed factorial analysis of variance (ANOVA) on these data. As predicted, the interaction of Stress \times Time was significant, $F(1, 104) = 4.56, p = .05, \eta_p^2 = .042$. At the beginning of the study, participants in the high-stress ($M = 2.41$, 95% confidence interval [CI; 2.04, 2.78]) and the low-stress ($M = 2.36$, 95% CI [1.97, 2.75]) conditions reported similar levels of stress. However, at the end of the study, participants in the high-stress condition reported higher levels of stress ($M = 3.66$, 95% CI [3.22, 4.10]) than those in the low-stress condition ($M = 3.02$, 95% CI [2.55, 3.49]), $t(104) = 1.97, p = .052, d = .38$. Together, these findings confirm that the stress manipulation in Experiment 1 was effective.

In the first critical analysis, as predicted, discrimination accuracy (as measured by the signal detection measure d') was signif-

icantly lower in the high-stress condition ($M = .46$, 95% CI [.32, .60]) than the low-stress condition ($M = .73$, 95% CI [.56, .91]), $t(104) = 2.46, p = .008, d = .48$. Note that this is a medium effect size. Given this significant effect of stress when examining old/new discrimination accuracy assessed with d' , subsequent analyses of confidence-specific accuracy cannot be attributed to a manipulation failure of the stress variable.⁵ The data for d' , hit rate, false-alarm rate, and response bias (as measured by the signal-detection measure of criterion, c) are presented in Table 1. As can be seen in Table 1, high stress affected the false-alarm rates more than the hit rates. Considering the criterion data, compared with participants in the low-stress condition, those in the high-stress condition were significantly less likely to respond “new,” $t(104) = 2.11, p = .037, d = .410$.

Next, we examined metamemory judgments for identifications by assessing confidence-specific accuracy for “old” responses. Consistent with previous research on the CA relationship, we defined accuracy at each level of confidence as Number of Hits_{*c*} / (Number of Hits_{*c*} + Number of False Alarms_{*c*}), where *c* indicates that the hits and false alarms were made with a specific level of confidence. This proportion can be interpreted as the proportion of “old” responses that are correct, which is similar to analyzing data from only “choosers” in an eyewitness identification paradigm. We did not conduct one overall 2 (stress condition) \times 4 (confidence: 1 and 2, 3, 4, 5) ANOVA on the proportion-correct data because there would have been too few participants who had nonmissing data (i.e., a calculable proportion correct with nonzero values) across all confidence levels. In addition, there were few observations at the two lowest levels of confidence, so we collapsed across Levels 1 and 2 for all analyses; this is a common practice to increase the stability of proportion-correct estimates (e.g., Wixted et al., 2015).

To explore whether the CA relationship differed as a function of the stress condition, we compared proportion-correct data at each of the four levels of confidence. These data are presented in the left panel of Figure 2. Independent-groups *t* tests were performed on the proportion correct in the high-stress versus the low-stress condition at each level of confidence. A Bonferroni correction of $\alpha = .013$ was used. The critical finding was that at Confidence Level 5, the proportion correct did not differ between the high-stress ($M = .78$, 95% CI [.68, .87]) and the low-stress conditions ($M = .82$, 95% CI [.73, .91]), $t(84) = .69, p = .494, d = .15$. In addition, at Confidence Level 4, again, the proportion correct did not differ between the high-stress ($M = .62$, 95% CI [.55, .68]) and the low-stress conditions ($M = .72$, 95% CI [.65, .79]), $t(98) = 2.14, p = .035, d = .43$. At Confidence Level 3, the proportion



Figure 1. Representative examples of Experiment 1 stimuli with a low-stress image presented alongside each face. The images presented here are similar to the low-stress International Affective Picture System (IAPS) images used in this study. However, because IAPS images cannot be posted publicly, we had to include representative images and not the actual images. These images are used with permission from the database. See the online article for the color version of this figure.

⁴ The instructions elaborated, “By stress, we mean a feeling of anxiety that comes from an increase in adrenaline that triggers your fight-or-flight response. Examples of this kind of stress would be giving a speech on a topic you are unprepared for, getting pulled over by a police officer, or watching a distressing scene in a movie or television show.”

⁵ One criticism of the signal-detection measure of discrimination accuracy, d' , is that this measure assumes that the standard deviations of the signal and noise distributions are equal. To address this limitation, we also computed d_a on the group-level data. This measure accounts for unequal variances for the signal and noise distributions (Gaetano, Lancaster, & Tindle, 2015). The mean d_a values computed at the group level were similar but slightly lower than the mean d' values for both the high- ($d' = .46, d_a = .42$) and the low-stress conditions ($d' = .73, d_a = .66$).

Table 1
Mean d' , c , Hit Rate, and False-Alarm Rate in Each Condition in Experiments 1 and 2

Experiment	d'	c	FAR	HR	N
1					
Low stress	.73 [.56, .91]	.23 [.13, .33]	.30 [.26, .34]	.55 [.51, .59]	106
High stress	.46 [.32, .60]	.08 [−.03, .18]	.39 [.35, .44]	.56 [.51, .60]	
2					
Low stress	.95 [.83, 1.08]	.23 [.13, .33]	.27 [.23, .31]	.59 [.55, .63]	133
High stress	.68 [.53, .82]	.15 [.06, .24]	.33 [.30, .37]	.57 [.53, .61]	

Note. FAR = false-alarm rate; HR = hit rate. The 95% confidence intervals are in brackets. Positive criterion values (c) represent a bias to respond “new,” and negative criterion values represent a bias to respond “old.”

correct did not significantly differ between the high-stress ($M = .49$, 95% CI [.42, .55]) and the low-stress conditions ($M = .61$, 95% CI [.54, .68]), $t(101) = 2.49$, $p = .015$, $d = .49$. Finally, at the lowest levels of confidence (1 and 2 combined), the proportion correct did not differ between the high-stress ($M = .50$, 95% CI [.40, .59]) and the low-stress conditions ($M = .51$, 95% CI [.41, .60]), $t(81) = .19$, $p = .853$, $d = .04$. These findings suggest that participants in the high- and low-stress conditions were similarly able to metacognitively calibrate subjective ratings of confidence and were more likely to provide high-confidence judgments when their identifications were likely to be accurate.

A potential criticism of these parametric t tests is that participants could only be included in each analysis if they provided an identification (“old” response) at a given confidence level, and subsequently, the sample sizes varied across tests. To address this, four Mann–Whitney U tests were conducted to address the potential limitation of a small sample size. The results of these nonparametric tests replicated the results of the parametric t tests for Confidence Level 5 ($U = 836.0$, $p = .400$) and Confidence Level 1 and 2 ($U = 827.0$, $p = .775$), the highest and lowest levels of confidence. However, in contrast to the nonsignificant parametric tests, the proportion correct was significantly higher for high-versus low-stress participants at Confidence Level 4 ($U = 902.5$, $p = .016$) and Confidence Level 3 ($U = 975.0$, $p = .021$).

We next compared the effect sizes for the proportion-correct data (measured by Cohen’s d) for the stress variable at each level of confidence, with the effect size for d' data provided for comparison. These data are presented in the left panel of Figure 3. Based on this comparison of effect sizes, it is clear that the magnitude of the effect of stress was generally larger for discrimination than for confidence-specific accuracy. In fact, the nonsignificant effect of stress on the proportion correct was smallest at the extreme ends of the confidence scale, when participants were very sure that they were correct or just guessing. Most important, and consistent with the proportion-correct data in Figure 2, the small Cohen’s d at Confidence Level 5 suggests that even when stress had a medium-size effect on memory discrimination accuracy, participants were somewhat aware of this and were able to adjust their high-confidence judgments accordingly.

The CA relationship for “new” responses was also examined in the high- and low-stress conditions. These data are presented in the left panel of Figure 4. Using a Bonferroni correction of $\alpha = .013$, the results from independent t tests indicated that at each level of confidence, the proportion correct for items judged to be “new” did not differ between the high- and low-stress conditions.⁶ This result

suggests that the effect of stress on the CA relationship is greater for “old” than for “new” faces. This finding is consistent with the results of other studies (see, e.g., Nguyen et al., 2017).

Are people generally accurate in their assessment of the effect of stress on their own memory? Our results suggest that they are. At the end of the experiment, participants were asked, “How do you think viewing stressful materials affects your own memory?” The results for the high-stress participants are most telling because they had just experienced “being stressed.” The frequency of responses for each response option were as follows: *significant beneficial effect* ($N = 0$), *small beneficial effect* ($N = 6$), *no effect* ($N = 10$), *small detrimental effect* ($N = 23$), and *significant detrimental effect* ($N = 17$). Most of the participants in the high-stress condition (71%) responded that stress had a small or significant detrimental effect on their memory.

Experiment 1 assessed whether stress affects face-recognition accuracy as well as confidence-specific accuracy and specifically focused on the proportion correct for high-confidence judgments. A unique feature of this study in the eyewitness memory literature is that we manipulated stress using a well-established, valid procedure for manipulating stress (Bradley et al., 1992; Erk et al., 2003) with a procedure in which the induced stress was intricately linked to the face stimuli; each IAPS image was presented simultaneously with a face with which it was randomly paired, and a unique image was paired with each face. We found higher discrimination accuracy for faces in the low-stress than the high-stress condition, but even so, the proportion correct did not differ between stress conditions at the highest level of confidence. Together, these findings suggest, first, that eyewitness discrimination accuracy is impaired by elevated levels of stress. But further, eyewitnesses appear to be sensitive to this effect and metacognitively adjust their subjective confidence ratings to account for what they perceive to be relatively poorer encoding conditions. That is, in the high-stress condition, participants recognized that they observed each face under a high level of stress and thus did not rate their confidence to be high unless they were relatively sure.

Experiment 2

There are two potential limitations of the results of Experiment 1 regarding the effect of stress on the CA relationship. First, the overall discrimination accuracy (d') was relatively low (high

⁶ This pattern of nonsignificant results was replicated with nonparametric Mann–Whitney U tests.

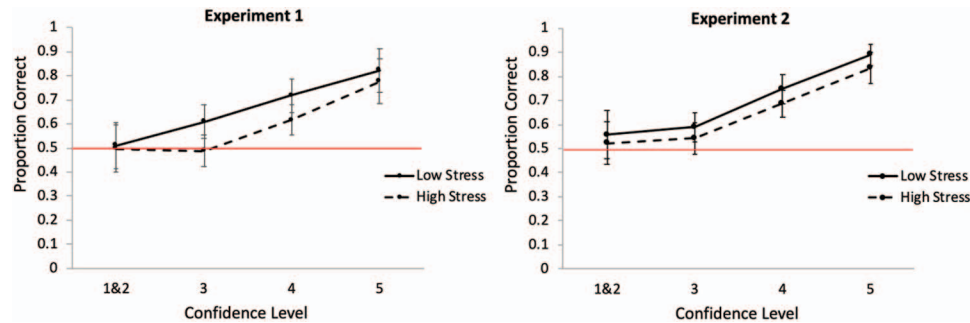


Figure 2. Confidence-specific accuracy for “old” responses assessed with confidence-accuracy characteristic (CAC) curves for Experiment 1 (left panel) and Experiment 2 (right panel). Proportion correct is computed as Number of Hits/(Number of Hits + Number of False Alarms) at each level of confidence. Chance performance is denoted by the horizontal line. Error bars represent 95% confidence intervals. See the online article for the color version of this figure.

stress: $M = .46$; low stress: $M = .73$). As reported by Nguyen et al. (2017) and Weber and Brewer (2003), a less reliable CA relationship results when recognition accuracy is low and approaches chance. Second, because each stress image was presented simultaneously with a face, it is not clear whether it was the stress of the image per se that affected face-recognition memory or if the high-stress images simply drew more attention away from the face stimuli than did the low-stress images. We conducted Experiment 2 to test the replicability of the results of Experiment 1 in an experiment that produces higher overall face-discrimination accuracy. To increase recognition accuracy, in Experiment 2, we reduced the number of target faces to 20 and the number of test faces to 40. In addition, we modified the procedure following that of Xie and Zhang (2016). Specifically, rather than simultaneously presenting each stress image with each face, in Experiment 2, each stress image immediately preceded the face with which it was randomly paired. This served to preclude the possibility that attention to faces was reduced by time spent attending to high- or low-stress images; they were not presented simultaneously.

Several additional procedural changes were introduced in Experiment 2. Rather than selecting negatively valenced and neutral stress images as in Experiment 1, in Experiment 2, following suggestions by Xie and Zhang (2016), the high-stress condition utilized negatively valenced images and the low-stress condition utilized positively valenced images to increase between-condition differences in the level of stress manipulated. Further, to test the generalizability of the findings to other stressful materials, images from the IAPS database used in Experiment 1 were replaced by images from the Open Affective Standardized Image Set (OASIS) database (Kurdi, Lozano, & Banaji, 2017). Except for the procedural changes just mentioned, Experiments 1 and 2 were essentially the same.

Method

Participants and design. A total of 163 volunteers participated on Amazon’s Mechanical Turk (via TurkPrime; Litman et al., 2017), following the same inclusion criteria as applied in Experiment 1. Although the power analysis (see Experiment 1) required $N = 49$ participants per condition, we recruited additional participants, anticipating that a number of them would violate the

exclusion criteria. The final sample ($N = 133$) included 66 participants in the high-stress condition (negatively valenced) and 67 participants in the low-stress (positively valenced) condition.⁷ The mean age did not differ between the low-stress condition (40.2 years, $SD = 11.89$, range = 22–70) and the high-stress condition (38.7 years, $SD = 11.54$, range = 21–68). There was an approximately equal gender split in both groups (female = 52%), and participants were primarily Caucasian (83%).

This study used a two-independent-groups design; approximately half of the participants were randomly assigned to the high-stress (negatively valenced) condition and half to the low-stress (positively valenced) condition.

Materials and procedure. In total, 40 target faces were selected from the same face database as was used in Experiment 1 (Meissner et al., 2005), with the same counterbalance conditions applied. In Experiment 2, the IAPS pictures were replaced with pictures from the OASIS database (Kurdi et al., 2017), an open-access online stimulus set containing 900 color images normed on valence and arousal. OASIS images, normed in 2015 on a Mechanical Turk population similar to that used in the present study, feature more current images and reflect more current valence and arousal ratings than do IAPS images. In Experiment 2, following suggestions by Xie and Zhang (2016), the high-stress condition utilized negatively valenced images and the low-stress condition utilized positively valenced images to increase the between-

⁷ Open Science Framework preregistered exclusion criteria for Experiment 2 that were violated were as follows: (a) responded to more than 95% of the test items with either (i) a recognition response of only “old” or only “new” or (ii) the same confidence rating (nine participants excluded); (b) univariate (± 2 standard deviations from the mean) or multivariate outliers (p value associated with Mahalanobis distance $< .01$) on measures of criterion or d' (two participants excluded); (c) completion time more than 2 SD from the total mean completion time (eight participants excluded); (d) responded with a rating of less than 4 to the question, “Complete this sentence by selecting the most appropriate option”: “I gave this study _____ attention” (1 = *almost none of my*, 5 = *my full*; one participant excluded); (e) responded “no” to the question, “. . . in your honest opinion, should we use your data in our analyses?” (three participants excluded); (f) reported that they completed the experiment on a device other than a laptop computer (three participants excluded); and (g) reported an interruption or technology issues while participating in the experiment (four participants excluded).

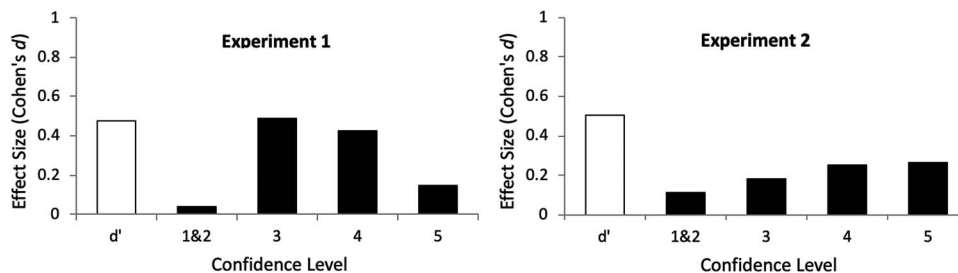


Figure 3. A comparison of the effect size of the difference in discrimination accuracy (as measured by d') and confidence-specific accuracy (as measured by the proportion correct for each level of confidence) for “old” responses between the high-stress condition and the low-stress condition in Experiment 1 (left panel) and Experiment 2 (right panel).

condition difference in the level of stress manipulated. Twenty-eight images from the OASIS database were selected in each stress condition (20 target stimuli and 8 practice stimuli). The images in the two conditions varied in valence but were matched on arousal. The normed mean valence rating on a 7-point scale significantly differed between the images selected for the positively valenced ($M = 6.16$, $SD = 0.03$) and the negatively valenced conditions ($M = 1.81$, $SD = 0.07$), $t(54) = 55.2$, $p < .001$. The normed mean arousal rating on a 7-point scale did not significantly differ between the images selected for the positively valenced ($M = 4.32$, $SD = 0.08$) and the negatively valenced conditions ($M = 4.44$, $SD = 0.62$).

The same basic procedure was used in Experiments 1 and 2, with a few changes for Experiment 2. The major change, following suggestions by Xie and Zhang (2016), was that rather than presenting the image and the face simultaneously as was done in Experiment 1, participants in Experiment 2 (a) viewed each image alone (for 4 s) and were told to think about how pleasant the image was on a 1–9 scale; (b) then (to intricately link the image to the face) were told to hold this rating in mind while they viewed the face that followed (for 6 s); and (3) finally, immediately after viewing the face, were told to check off their pleasantness rating of the image using the portrait version of the 9-point Self-Assessment Manikin (SAM) scale (Suk, 2006). Additional procedural changes in Experiment 2

included the following: (a) The number of target faces was reduced to 20, with 8 practice image-face pairs presented first, and (b) the number of test faces was reduced to 40, with 4 practice test faces presented first. Finally, as in Experiment 1, at the beginning and end of the study, participants were asked, “At this moment, how stressed do you feel on a scale from 1 (*not at all stressed*) to 7 (*very high level of stress*)?” In Experiment 2, participants responded to this question one additional time, at the end of the presentation phase immediately after they had viewed the last image–face pair, more proximate to when they experienced the stress condition.

Results and Discussion

As a manipulation check, we assessed self-report ratings of stress (on a scale from 1 [*not at all stressed*] to 7 [*very high level of stress*]) at the beginning of the study (Time 1), at the end of the presentation phase proximate to when they had experienced the stress condition (Time 2), and at the end of the study (Time 3). We conducted a 2 (High or Low Stress) \times 3 (Time) mixed factorial ANOVA on these data. As predicted, higher self-reported levels of stress were reported in the high-stress ($M = 3.14$, 95% CI [2.86, 3.41]) than the low-stress condition ($M = 2.20$, 95% CI [1.93, 2.47]), $F(1, 131) = 23.37$, $p < .001$, $\eta_p^2 = .15$. In addition, there was a significant main effect of time. Lower levels of stress were

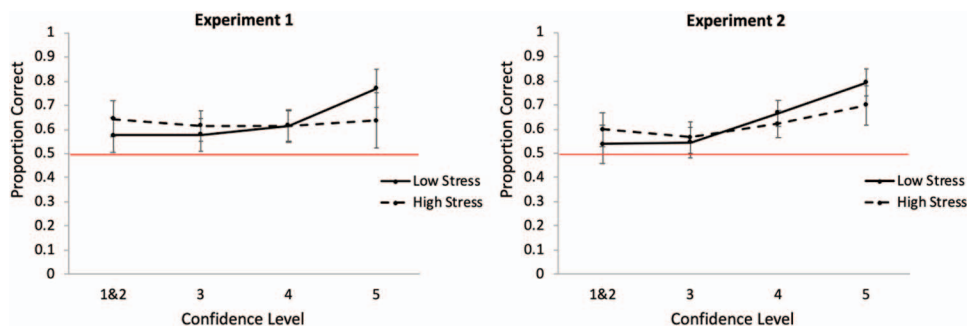


Figure 4. Confidence-specific accuracy for “new” responses assessed with confidence–accuracy characteristic (CAC) curves for Experiment 1 (left panel) and Experiment 2 (right panel). Proportion correct for new responses is computed as Number of Correct Rejections/(Number of Correct Rejections + Number of Misses) at each level of confidence. Chance performance is denoted by the horizontal line. Error bars represent 95% confidence intervals. See the online article for the color version of this figure.

reported at Time 1 ($M = 2.05$, 95% CI [1.85, 2.26]) than at Time 2 ($M = 2.86$, 95% CI [2.61, 3.12]) or at Time 3 ($M = 3.08$, 95% CI [2.83, 3.34], $F(2, 262) = 39.28$, $p < .001$, $\eta_p^2 = .23$). The comparison between Time 2 and Time 3 was not significant ($p = .213$). Critically, the interaction of Stress \times Time was significant, $F(2, 262) = 29.16$, $p < .001$, $\eta_p^2 = .18$. At the beginning of the study at Time 1, similar and lower levels of stress were reported in both the high-stress ($M = 2.17$, 95% CI [1.87, 2.46]) and the low-stress conditions ($M = 1.94$, 95% CI [1.65, 2.23], $t(131) = 1.08$, $p = .141$, $d = .19$). However, immediately after the presentation phase at Time 2, higher levels of stress were reported in the high-stress ($M = 3.86$, 95% CI [3.51, 4.22]) than the low-stress condition ($M = 1.87$, 95% CI [1.51, 2.22]), $t(131) = 7.89$, $p < .001$, $d = 1.36$. Similarly, following the test phase at Time 3, self-reported levels of stress remained significantly higher in the high-stress ($M = 3.38$, 95% CI [3.02, 3.74]) than the low-stress condition ($M = 2.79$, 95% CI [2.44, 3.15], $t(131) = 2.31$, $p = .012$, $d = .40$). Together, these findings confirm that the stress manipulation in Experiment 2 was effective.

In the first critical analysis, as predicted, discrimination accuracy (as measured by the signal detection measure d') was significantly lower in the high-stress ($M = .68$, 95% CI [.53, .82]) than the low-stress condition ($M = .95$, 95% CI [.83, 1.08]), $t(131) = 2.91$, $p = .002$, $d = .51$, and overall discrimination accuracy was higher in Experiment 2 than Experiment 1.⁸ The data for d' , hit rate, false-alarm rate, and response bias (as measured by the signal-detection measure of criterion, c) are presented in Table 1. As can be seen in Table 1, and consistent with the results of Experiment 1, high stress affected the false-alarm rates more than the hit rates. Considering the criterion data, unlike the results of Experiment 1, participants in the low-stress and high-stress conditions were similarly likely to respond “new,” $t(131) = 1.23$, $p = .222$, $d = .274$.

The primary purpose of Experiment 2 was to assess if the effect of stress on the CA relationship reported in Experiment 1 resulted as well with higher overall discrimination accuracy. We examined metamemory judgments for identifications by assessing confidence-specific accuracy for “old” responses, with accuracy defined at each level of confidence as Number of Hits_{*c*}/(Number of Hits_{*c*} + Number of False Alarms_{*c*}), where c indicates that the hits and false alarms were made with a specific level of confidence. As in the analyses of Experiment 1, we did not conduct one overall 2 (stress condition) \times 4 (confidence: 1 and 2, 3, 4, 5) ANOVA on the proportion-correct data because there were too few participants who had nonmissing data (i.e., a calculable proportion correct with nonzero values) across all confidence levels. In addition, there were few observations at the two lowest levels of confidence, so we collapsed across Levels 1 and 2 for all analyses to increase the stability of proportion-correct estimates.

To explore whether the CA relationship differed as a function of the stress condition, we compared proportion-correct data for “old” responses at each of the four levels of confidence. These data are presented in the right panel of Figure 2. Independent-groups t tests were performed on the proportion correct in the high-stress versus the low-stress condition at each level of confidence. A Bonferroni correction of $\alpha = .013$ was used. The critical finding was that at Confidence Level 5, the proportion correct did not differ between the high-stress ($M = .83$, 95% CI [.77, .90]) and the low-stress conditions ($M = .89$, 95% CI [.85, .93]), $t(119) = 1.47$, $p = .144$,

$d = .27$. In addition, at each of the other three levels of confidence, the proportion correct also did not differ between the high-stress and the low-stress conditions.⁹ These findings suggest that even when recognition accuracy was relatively high and significantly higher in the low-stress than the high-stress conditions, nonetheless, participants in the high- and low-stress conditions were similarly able to metacognitively calibrate subjective ratings of confidence, even at the highest level of confidence.

We next compared the effect size for the proportion-correct data (measured by Cohen's d) for the stress variable at each level of confidence, with the effect size for d' data provided for comparison. These data are presented in the right panel of Figure 3. Based on this comparison of effect sizes, and consistent with the results of Experiment 1, it is clear that the magnitude of the effect of stress was generally larger for discrimination than for confidence-specific accuracy. In fact, the effect size for stress on the proportion correct was relatively small at each of the four levels of confidence (Level 5: $d = .27$, Level 4: $d = .25$, Level 3: $d = .18$; Levels 1 and 2: $d = .11$). Most important, and consistent with the proportion-correct data in Figure 2, the small Cohen's d at Confidence Level 5 suggests that even when stress had a medium-size effect on memory discrimination accuracy, participants were somewhat aware of this and adjusted their confidence judgments accordingly.

The CA relationship for “new” responses was also examined in the high- and low-stress conditions. These data are presented in the right panel of Figure 4. Using a Bonferroni correction of $\alpha = .013$, the results from independent t tests indicated that at each level of confidence, the proportion correct for items judged to be “new” did not differ between the high- and low-stress conditions.¹⁰ This result is consistent with the results of Experiment 1 and suggests that the effect of stress on the CA relationship is greater for “old” than for “new” faces, a result that was also reported by Nguyen et al. (2017).

In analyzing the results of Experiment 2, we noticed that 15 participants in the high-stress condition provided Time 2 self-report levels of stress that were a 1 or 2 on the scale of 1 (*not at all stressed*) to 7 (*very high level of stress*), and 2 participants in the low-stress condition provided Time 2 self-report levels of stress that were a 6 or 7 on this scale. In light of reports of individual differences in reactivity to stressful stimuli by Buchanan and Tranel (2008) and others, we anticipated that we might have some participants who had idiosyncratic reactions to stressful stimuli. To address this concern, we included the following exclusion criterion in our Open Science Framework preregistration: “In the low-stress condition, responded to the Time 2 self-report stress question with a rating of 6 or 7; in the high-stress condition, responded to the Time 2 self-report stress question with a rating of 1 or 2.” We reanalyzed the data with these 17 participants excluded to test (a) whether the effect of stress on recognition accuracy was stronger and (b) whether the proportion-correct results were sim-

⁸ The mean d_a values computed at the group level were similar but slightly lower than the mean d' values for both the high- ($d' = .68$, $d_a = .61$) and the low-stress conditions ($d' = .95$, $d_a = .85$).

⁹ For all four of these parametric tests, we replicated the pattern of results with nonparametric Mann-Whitney U tests ($ps > .14$).

¹⁰ This pattern of nonsignificant results was replicated with nonparametric Mann-Whitney U tests.

ilar when participants' self-reported stress was consistent with their assigned condition compared with the results reported previously.

As predicted, the effect of stress on discrimination accuracy (d') was even greater with these 17 subjects excluded. Discrimination accuracy was significantly lower in the high-stress ($M = .61$, 95% CI [.45, .76]) than the low-stress condition ($M = .97$, 95% CI [.84, 1.09], $t(114) = 3.66$, $p < .001$, $d = .69$, and this difference was greater than that reported in the previous analysis that included these 17 participants who had indicated idiosyncratic responses to stress ($d = .51$).¹¹ To assess whether this result affected the CA relationship, we conducted independent-groups t tests on the proportion correct in the high-stress versus the low-stress condition at each level of confidence. A Bonferroni correction of $\alpha = .013$ was used. Again, at each level of confidence, the proportion correct did not differ between the high-stress and low-stress conditions. The critical finding was that even at Confidence Level 5, the proportion correct did not differ between the high-stress ($M = .81$, 95% CI [.73, .90]) and low-stress conditions ($M = .89$, 95% CI [.85, .94]), $t(104) = 1.76$, $p = .082$, $d = .35$.¹² These findings corroborate the result of the initial analyses of Experiment 2 data and suggest that even when recognition accuracy was relatively high and significantly higher in the low-stress than the high-stress conditions, nonetheless, participants in the high- and low-stress conditions were similarly able to metacognitively calibrate subjective ratings of confidence, even at the highest level of confidence.

Finally, we assessed whether participants were generally accurate in their assessment of the effect of stress on their own memory. Our results suggest that they were. At the end of the study, all participants were asked, "How do you think being stressed affects your own memory?" The results for the high-stress participants are most telling because they had just experienced "being stressed." For the full sample of 66 high-stress participants, the frequency of responses for each response option was as follows: *significant beneficial effect* ($N = 1$), *small beneficial effect* ($N = 7$), *no effect* ($N = 2$), *small detrimental effect* ($N = 32$), and *significant detrimental effect* ($N = 24$). Consistent with the results of Experiment 1, in Experiment 2, most of the high-stress participants (85%) responded that stress had a small or significant detrimental effect on their memory.

The results of Experiment 1 were replicated in Experiment 2, in which higher overall discrimination accuracy was obtained. The results of Experiment 1 are thus not restricted to conditions with low overall recognition accuracy. Second, in Experiment 2, we modified the procedure following that of Xie and Zhang (2016) to test if the high-stress images simply drew attention away from the face stimuli. Rather than simultaneously presenting each stress image with each face, as was the procedure in Experiment 1, in Experiment 2, each stress image immediately preceded the face with which it was randomly paired. This precluded the possibility that attention to faces was reduced by time spent attending to high- or low-stress images. The fact that the results of Experiment 1 were replicated in Experiment 2 thereby excludes the alternative competing-attention account of these results. Together, the results of our two experiments suggest that the effect of stress on recognition memory (but not the CA relationship) is robust because it was replicated with a different stress procedure and using different high- and low-stress stimuli.

General Discussion

A consistent pattern of findings resulted across both experiments. First, eyewitness discrimination accuracy was impaired by elevated levels of stress; in both experiments, d' was significantly higher in the low-stress than the high-stress condition, with a medium effect size measured by Cohen's d ($d_{\text{Exp.1}} = .48$, $d_{\text{Exp.2}} = .51$). But further, participants appeared to be sensitive to this effect and were metacognitively able to adjust their subjective confidence ratings to account for what they perceived to be relatively poorer encoding conditions in the elevated-stress condition. In both experiments, the proportion correct for "old" judgments did not significantly differ between stress conditions at high levels of confidence. It is important to note, however, that even at the highest level of confidence, the proportion correct was not perfect (i.e., 100%) but approximately 80–90% (Experiment 1: $M_{\text{high stress}} = .78$, $M_{\text{low stress}} = .82$, $d = .15$; Experiment 2: $M_{\text{high stress}} = .83$, $M_{\text{low stress}} = .89$, $d = .27$). Thus, although the results of this study suggest that confidence is a better predictor of recognition-memory accuracy than is stress, confidence alone is still an imperfect predictor. This is consistent with earlier work by Stretch and Wixted (1998), in which it was reported that although participants largely adjusted their confidence criteria following a likelihood-ratio decision model, they did so with less than 100% accuracy.

The results of this study clarify inconsistent findings reported in previous studies regarding the effect of high levels of stress on eyewitness identification accuracy and the CA relationship. Sauerland et al. (2016) reported similar levels of accuracy in identifying the perpetrator from a lineup in their high- and low-stress conditions using several signal-detection measures of recognition accuracy. Not surprisingly, then, they reported that the CA relationship did not differ between conditions either, a conclusion that is actually not possible without a significant main effect of stress on recognition accuracy.

More recently, Davis et al. (2019) reported in two experiments that recognition memory was more accurate (higher d' values) in a low-stress than a high-stress condition. This finding is consistent with a wealth of research reporting that stress impairs memory and cognitive processing more generally (Wolf, 2009). However, regarding the effect of stress on the CA relationship, Davis and colleagues reported somewhat inconsistent findings across their two experiments. Whereas the proportion correct did not differ between the high- and low-stress conditions for the highest confidence judgments in Experiment 1 (see their Figure 1), this difference was significant in Experiment 2 (see their Figure 2). The results of their study are thus inconclusive regarding the effect of stress on the CA relationship. More important, though, in the Davis study as in ours, the important and consistent finding is that confidence is a better predictor of recognition accuracy than is stress.

¹¹ In addition, the mean d_a values computed at the group level were lower than the mean d' values for both the high- ($d' = .81$, $d_a = .56$) and the low-stress conditions ($d' = .89$, $d_a = .87$).

¹² Again, the results of the nonparametric Mann–Whitney U tests replicated the nonsignificant patterns of results of the parametric t tests ($ps > .1$). Specifically, for Confidence Level 5, $U = 1,191.5$, $p = .210$.

The finding that stress impairs eyewitness identification accuracy is an important finding, and it is especially important to test this relationship in a study in which the induced stress is intricately linked to the encoding of the faces. A limitation of the results of both Davis et al. (2019) and Sauerland et al. (2016) in their generalizability to real-world eyewitnesses is that in both studies, a version of the cold pressor task was used to manipulate stress. This is a task in which the stress manipulation is not intricately linked to the encoding of the faces, and this procedure may have mitigated the effect of stress on the cognitive processing of the faces. This may thus limit the generalizability of their results to real-world eyewitnesses. In the present study, a unique image was paired with each face, and high- and low-stress stimuli were contiguously and temporally linked to each face with which they were paired. This was achieved using two different procedures for presenting the stimuli. The results of this study are thus an important step toward generalizing this research to a range of stress manipulations that are likely to occur with real-world eyewitnesses.

That said, one limitation of the current study is that although there was a significant difference in both experiments between the mean self-reported levels of stress at Time 2 in the high-stress and low-stress conditions (Experiment 1: $M_{\text{high stress}} = 3.66$, $M_{\text{low stress}} = 3.02$, $d = .38$; Experiment 2: $M_{\text{high stress}} = 3.86$, $M_{\text{low stress}} = 1.87$, $d = 1.36$), the stress levels reported in the high-stress condition were actually only slightly above the midpoint of the 1–7 stress scale. This suggests that although the self-reported stress levels were elevated in the high-stress condition relative to the low-stress condition, extremely high levels of stress were not achieved in our study. This should be taken into consideration in generalizing the results of this study to real-world eyewitnesses. Stronger manipulations of stress (yielding even larger effects of stress on recognition-memory performance) have been induced, for example, by Morgan et al. (2004) in research on active-duty military personnel enrolled in military survival school training and by Valentine and Mesout (2009) in a study of people in the Horror Labyrinth of the London Dungeon. And importantly, Wixted, Mickes, et al. (2016) reported that the CA relationship was reliable among real-world eyewitnesses, most of whom had been a victim of an armed robbery and thus were likely to have experienced very high stress at the time of the event.

In addition, two observations highlight the need for future research to replicate and extend our findings using research paradigms that more closely approximate a highly stressful eyewitness event. First, we noted that Davis et al. (2019) reported inconsistent effects of stress on the proportion correct at high confidence across their two experiments (i.e., a significant difference in Experiment 2 but not in Experiment 1). Second, although we consistently observed no significant effect of stress on the proportion correct at high confidence and small effect sizes, these effect sizes fluctuated between Experiments 1 and 2 ($d_{\text{Exp.1}} = .15$, $d_{\text{Exp.2}} = .27$). Thus, it is possible that there may yet be an effect of stress on high-confidence identifications in real eyewitness identification scenarios. Despite these observations, our critical and consistent finding is that confidence level, not stress, provides relatively reliable information about whether an identification is likely to be accurate. In other words, knowing the level of *stress* at the time of an observation does not tell you very much about whether an identification is likely to be accurate, whereas knowing the level of

confidence at the time of the identification tells you a great deal about whether an identification is likely to be accurate.

There are two additional possible constraints on the ecological validity of our study that suggest directions for future research. First, static photographs were used as face stimuli in this study, and there are surely differences between the recognition memory for two-dimensional versus three-dimensional versions of faces. Nonetheless, there is no reason to think that stress would differentially affect the CA relationship for two-dimensional versus three-dimensional versions of the same faces. Second, in our study, stress was introduced at encoding. However, given that the test phase immediately followed the encoding phase, stress was likely to have also been in effect at the time of the test. In light of reports that stress has a state-dependent effect on memory (cf. Robinson & Rollings, 2010), it remains to be seen if the results reported in this study are observed when stress is manipulated at encoding but not at testing. Future research is necessary to assess if our results reflect an encoding phenomenon, a retrieval phenomenon, or both.

But why do stress and other estimator variables have a large effect on discriminability but a smaller effect on the proportion correct at a given level of confidence? Several models have been proposed to account for this finding. Semmler, Dunn, Mickes, and Wixted (2018) compared the optimality hypothesis with the likelihood-ratio model for determining the relationship between memory accuracy and confidence judgments. According to the optimality hypothesis, the CA correlation should vary directly with the optimality of the encoding conditions, with confidence judgments based on the familiarity or memory strength of the test item (Deffenbacher, 1980). But this model does not fit recent data comparing discrimination accuracy with CA data. Alternatively, a better fit to existing data is offered by the likelihood-ratio model. According to this model, people have learned, from error feedback in everyday experience, what factors contribute to strong versus weak memory signals and then adjust their confidence ratings such that a stronger memory-match signal is needed before deciding with high confidence that a test stimulus matches the observed stimulus (Mickes, Hwe, Wais, & Wixted, 2011).

Although the current study does not actually test the likelihood-ratio model, this model can account for the obtained results. The participants in our study did seem aware that elevated stress impairs memory, a finding implicit in the likelihood-ratio model. In both experiments, the large majority of participants who had experienced the high-stress condition (71% in Experiment 1, 78% in Experiment 2) correctly responded that stress had a small or significant detrimental effect on their memory. In our previous study regarding the effect of marijuana on the CA relationship (Pezdek et al., 2020), we hypothesized that it might generally be the case that people are less aware of the effect of state variables (e.g., drugs, alcohol, stress, sleep, etc.) on their memory than the effect of situational variables (cross-race faces, brief exposure time, retention interval, physical distance, etc.). The results of the current study suggest that this hypothesis is not correct. Stress and marijuana would both be considered state variables, but they revealed a different pattern of results in terms of the CA relationship.

Alternatively, the results of this study, along with the previous results reported by Pezdek et al. (2020), support the hypothesis that to metacognitively calibrate subjective ratings of confidence, people need to be aware of how a specific variable affects their

memory accuracy so that they can more cautiously allocate high-confidence judgments in relatively poorer memory conditions. People were not generally accurate in assessing how marijuana affected their memory in our previous study (Pezdek et al., 2020), but they were generally accurate in assessing how stress affected their memory in the current study.

These results indicate that elevated stress impairs eyewitness memory accuracy as measured by discriminability (d' and d_a). However, this is not the measure that is most informative to judges and jurors, who want to know that if an eyewitness identifies a given suspect, what is the probability that that suspect is actually the perpetrator? In the eyewitness memory literature, this is referred to as the proportion correct by “choosers.” Based on this measure, the results of both experiments reported here suggest that highly confident eyewitnesses—those who are most likely to be brought to court to testify—are likely to be similarly accurate under low and elevated levels of stress. It is important to note, however, that in this study, even at the highest level of confidence, the proportion correct was not perfect but approximately 80–90%. We conclude that, yes, stress impairs eyewitness identification accuracy overall. However, confidence is a much better predictor of recognition accuracy than is stress, even though confidence is an imperfect predictor.

References

- Bergmann, H. C., Rijpkema, M., Fernández, G., & Kessels, R. P. (2012). The effects of valence and arousal on associative working memory and long-term memory. *PLoS ONE*, 7, e52616. <http://dx.doi.org/10.1371/journal.pone.0052616>
- Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 379–390. <http://dx.doi.org/10.1037/0278-7393.18.2.379>
- Buchanan, T. W., & Tranel, D. (2008). Stress and emotional memory retrieval: Effects of sex and cortisol response. *Neurobiology of Learning and Memory*, 89, 134–141. <http://dx.doi.org/10.1016/j.nlm.2007.07.003>
- Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory & Cognition*, 6, 82–92. <http://dx.doi.org/10.1016/j.jarmac.2016.04.001>
- Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, 12, 41–55. <http://dx.doi.org/10.1007/BF01064273>
- Davis, S. D., Peterson, D. J., Wissman, K. T., & Slater, W. A. (2019). Physiological stress and face recognition: Differential effects of stress on accuracy and the confidence–accuracy relationship. *Journal of Applied Research in Memory & Cognition*, 8, 367–375. <http://dx.doi.org/10.1016/j.jarmac.2019.05.006>
- Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, 4, 243–260. <http://dx.doi.org/10.1007/BF01040617>
- Deffenbacher, K. A., Bornstein, B. H., McGorty, E. K., & Penrod, S. D. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied*, 14, 139–150. <http://dx.doi.org/10.1037/1076-898X.14.2.139>
- Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior*, 28, 687–706. <http://dx.doi.org/10.1007/s10979-004-0565-x>
- Erk, S., Kiefer, M., Grothe, J., Wunderlich, A. P., Spitzer, M., & Walter, H. (2003). Emotional context modulates subsequent memory effect. *NeuroImage*, 18, 439–447. [http://dx.doi.org/10.1016/S1053-8119\(02\)00015-0](http://dx.doi.org/10.1016/S1053-8119(02)00015-0)
- 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. <http://dx.doi.org/10.3758/BF03193146>
- Gaetano, J. M., Lancaster, S., & Tindle, R. (2015). *Signal detection theory calculator 1.0* [Excel workbook]. Retrieved from https://www.researchgate.net/publication/284714960_Signal_detection_theory_calculator
- Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1304–1316. <http://dx.doi.org/10.1037/0278-7393.22.5.1304>
- Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). *Behavior Research Methods*, 49, 457–470. <http://dx.doi.org/10.3758/s13428-016-0715-3>
- Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual* (Technical Report A-8). Gainesville: University of Florida.
- Lindsay, R. C. L., Semmler, C., Weber, N., Brewer, N., & Lindsay, M. R. (2008). How variations in distance affect eyewitness reports and identification accuracy. *Law and Human Behavior*, 32, 526–535. <http://dx.doi.org/10.1007/s10979-008-9128-x>
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49, 433–442. <http://dx.doi.org/10.3758/s13428-016-0727-z>
- Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other race faces: A dual-process approach. *Applied Cognitive Psychology*, 19, 545–567. <http://dx.doi.org/10.1002/acp.1097>
- Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory & Cognition*, 4, 93–102. <http://dx.doi.org/10.1016/j.jarmac.2015.01.003>
- Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, 140, 239–257. <http://dx.doi.org/10.1037/a0023007>
- Morgan, C. A., III, Hazlett, G., Doran, A., Garrett, S., Hoyt, G., Thomas, P., . . . Southwick, S. M. (2004). Accuracy of eyewitness memory for persons encountered during exposure to highly intense stress. *International Journal of Law and Psychiatry*, 27, 265–279. <http://dx.doi.org/10.1016/j.ijlp.2004.03.004>
- National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: National Academies Press.
- Nguyen, T. B., Pezdek, K., & Wixted, J. T. (2017). Evidence for a confidence–accuracy relationship in memory for same- and cross-race faces. *The Quarterly Journal of Experimental Psychology*, 70, 2518–2534. <http://dx.doi.org/10.1080/17470218.2016.1246578>
- Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence–accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, 19, 55–71. <http://dx.doi.org/10.1037/a0031602>
- Pezdek, K., Abed, E., & Reisberg, D. (2020). Marijuana affects the accuracy of eyewitness memory and the confidence–accuracy relationship too. *Journal of Applied Research in Memory & Cognition*, 9, 60–67. <http://dx.doi.org/10.1016/j.jarmac.2019.11.005>

- Robinson, S. J., & Rollings, L. J. (2010). The effect of mood-context on visual recognition and recall memory. *Journal of General Psychology*, 138, 66–79. <http://dx.doi.org/10.1080/00221309.2010.534405>
- Sauerland, M., Raymaekers, L. H., Otgaar, H., Memon, A., Waltjen, T. T., Nivo, M., . . . Smeets, T. (2016). Stress, stress-induced cortisol responses, and eyewitness identification performance. *Behavioral Sciences & the Law*, 34, 580–594. <http://dx.doi.org/10.1002/bsl.2249>
- Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied*, 24, 400–415. <http://dx.doi.org/10.1037/xap0000157>
- Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118, 315–327. <http://dx.doi.org/10.1037/0033-2909.118.3.315>
- Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1397–1410. <http://dx.doi.org/10.1037/0278-7393.24.6.1397>
- Suk, H. J. (2006). *Color and emotion—A study on the affective judgment across media and in relation to visual stimuli* (Doctoral dissertation). University of Mannheim, Mannheim, Germany. Retrieved from <https://madoc.bib.uni-mannheim.de/1336/>
- Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology*, 23, 151–161. <http://dx.doi.org/10.1002/acp.1463>
- Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology*, 88, 490–499. <http://dx.doi.org/10.1037/0021-9010.88.3.490>
- Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, 66, 688–696. <http://dx.doi.org/10.1037/0021-9010.66.6.688>
- Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, 7, 45–75. <http://dx.doi.org/10.1111/j.1529-1006.2006.00027.x>
- Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory & Cognition*, 4, 329–334. <http://dx.doi.org/10.1016/j.jarmac.2015.08.007>
- Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, 70, 515–526. <http://dx.doi.org/10.1037/a0039510>
- Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police line-ups. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 304–309. <http://dx.doi.org/10.1073/pnas.1516814112>
- Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence-accuracy relationship. *Journal of Applied Research in Memory & Cognition*, 5, 192–203. <http://dx.doi.org/10.1016/j.jarmac.2016.04.006>
- Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18, 10–65. <http://dx.doi.org/10.1177/1529100616686966>
- Wolf, O. T. (2009). Stress and memory in humans: Twelve years of progress? *Brain Research*, 1293, 142–154. <http://dx.doi.org/10.1016/j.brainres.2009.04.013>
- Xie, W., & Zhang, W. (2016). The influence of emotion on face processing. *Cognition and Emotion*, 30, 245–257. <http://dx.doi.org/10.1080/02699931.2014.994477>

Received January 2, 2020

Revision received May 16, 2020

Accepted May 19, 2020 ■