Stress, stress-induced cortisol responses, and eyewitness identification performance

Behavioral Sciences and the Law
Behav. Sci. Law **34**: 580–594 (2016)
Published online 15 July 2016 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/bsl.2249

Stress, stress-induced cortisol responses, and eyewitness identification performance

Melanie Sauerland*, Linsey H.C. Raymaekers[†], Henry Otgaar[†,‡], Amina Memon[¶], Thijs T. Waltjen[†], Maud Nivo[†], Chiel Slegers[†], Nick J. Broers[§] and Tom Smeets[†]

In the eyewitness identification literature, stress and arousal at the time of encoding are considered to adversely influence identification performance. This assumption is in contrast with findings from the neurobiology field of learning and memory, showing that stress and stress hormones are critically involved in forming enduring memories. This discrepancy may be related to methodological differences between the two fields of research, such as the tendency for immediate testing or the use of very short (1–2 hours) retention intervals in eyewitness research, while neurobiology studies insert at least 24 hours. Other differences refer to the extent to which stress-responsive systems (i.e., the hypothalamic–pituitary–adrenal axis) are stimulated effectively under laboratory conditions. The aim of the current study was to conduct an experiment that accounts for the contemporary state of knowledge in both fields. In all, 123 participants witnessed a live staged theft while being exposed to a laboratory stressor that reliably elicits autonomic and glucocorticoid stress responses or while performing a control task. Salivary cortisol levels were measured to control for the effectiveness of the stress induction. One week later, participants attempted to identify the thief from target-present and target-absent line-ups. According to regression and receiver operating characteristic analyses, stress did not have robust detrimental effects on identification performance.

Expert witnesses are often asked to evaluate the impact of high levels of stress during a criminal event on eyewitnesses' identification performance. Such questions are driven by the concern that stress during encoding might exert negative effects on witnesses' memory, a view shared by many experts in the psychology and law domain (Kassin, Ellsworth, & Smith, 1989). In a 1989 survey among expert witnesses, Kassin et al. found that the expert panel generally agreed with the statement that high levels of stress impaired eyewitness testimony. None of the experts was in favor of the idea that stress could have beneficial effects on memory. Twelve years later, this picture changed. In a new edition of the survey (Kassin, Tubb, Hosch, & Memon, 2001), there was a drop in the consensus on this issue, with fewer experts rating the available evidence as reliable (1989, 71%; 2001, 60%), and fewer experts stating that they would be willing to testify about the issue in court (65% vs. 50%).

* Correspondence to: Melanie Sauerland, Maastricht University, Faculty of Psychology and Neuroscience, P.O. Box 616, 6200 MD Maastricht, the Netherlands. E-mail: melanie.sauerland@maastrichtuniversity.nl
[†]Department of Clinical Psychological Science, Maastricht University, the Netherlands
[‡]City University London, UK
[¶]Royal Holloway University of London, UK
[§]Department of Methodology and Statistics, Maastricht University, the Netherlands

The lack of consensus on the effects of stress on identification ability is in line with the divergent conclusions originating from two different strands of research, namely the eyewitness field and the neurobiology domain. In the eyewitness field, a meta-analysis revealed that stress during encoding has negative effects on eyewitness identification performance (Deffenbacher, Bornstein, Penrod, & McGorty, 2004), with strong effects for target-present line-ups but negligent effects for target-absent line-ups. Furthermore, the effect was more pronounced under conditions of higher ecological validity (eyewitness identification vs. face recognition tasks; staged crime vs. other means of inducing stress). Note, however, that only six of the 22 eyewitness identification studies included in the meta-analysis displayed high ecological validity (i.e., employed a staged crime).

In a stressful situation, physiological stress responses result in supplementary energy becoming available for the individual to act in response to the stressor. This includes the fight-or-flight response that is associated with epinephrine and norepinephrine release on the one hand, and the activation of the hypothalamic–pituitary–adrenal (HPA) axis, which goes along with cortisol secretion on the other side. Additionally, physiological stress responses facilitate the storage of memories related to the stressful situation by acting on brain structures involved in regulating human memory performance (e.g., the amygdala and hippocampal formation; see Phelps, 2004; Schwabe et al., 2012; Wolf, 2009).

Contrary to eyewitness identification laboratory studies, the empirical literature from the field of neurobiology indicates that memory performance for information encoded under stress tends to be superior to memory performance for neutral information (see LaBar & Cabeza, 2006 for a review; Roozendaal & McGaugh, 2011). This effect is somewhat weaker for recognition compared with recall tasks (Het, Ramlow, & Wolf, 2005), yet the effect on recognition performance is especially increased when a proper delay of at least 24 hours is inserted between encoding and recognition (Schwarze, Bingel, & Sommer, 2012; experiment 1).

The seemingly divergent findings of the two fields could be related to differences in methodology. First, previous eyewitness identification research mostly induced relatively small levels of stress (e.g., by using video or staged crimes), while neurobiological stress studies have generally reverted to pharmacological (e.g., cortisone administration) or standardized laboratory stress tests (e.g., public speaking, mental arithmetic, exposure to painful stimuli). Second, most of the neurobiological stress studies carefully separate the effects of stress on memory consolidation versus those on retrieval performance by introducing a lengthy (i.e., 24 hours or more) retention interval. Eyewitness identification studies, however, are frequently executed within a single session of roughly 1 or 2 hours so the impairing effects of stress on memory retrieval may offset the known beneficial effects of stress and stress hormones on memory formation (Schwabe et al., 2012; Wolf, 2009). Indeed, less than half (i.e., 10) of the 22 studies included in the meta-analysis (Deffenbacher et al., 2004) used an interval of at least 24 hours in one or more conditions. One notable exception is a recent study that used a 2-week retention interval (Rush et al., 2013) and found no effect of stress on identification performance for children in target-present or target-absent line-ups although there was one exception. We will return to this study later in this paper. The absence of sufficiently long retention intervals is a problem for the interpretation of the majority of single-session eyewitness identification studies examining the effects of stress at encoding. It is unclear whether the observed stress-memory effects can be attributed

DOI: 10.1002/bsl

to stress affecting the memory formation phase or retrieval processes. Third, eyewitness identification studies seldom employ objective, physiological measures of stress to verify whether the stress induction procedure succeeded in eliciting bodily stress responses. Only seven studies included in the meta-analysis report physiological indicators of stress, such as heart rate, pulse, skin conductance or blood pressure [Bothwell, Brigham, & Pigott, 1987; Brigham et al., 1983; Hosch & Bothwell, 1990; Peters, 1988, 1991 (experiment 2), 1997 (experiment 1 and 2)]. Other studies relied on self-report measures of parents' or other observers' ratings (e.g., by nurses or strangers). Among the more recent studies, only two included measurements of heart rate (Valentine & Mesout, 2009) and cortisol levels (Rush et al., 2013) to corroborate the stress manipulation. The exclusive reliance on self-reports or other reports as a manipulation check is problematic because subjective measures of stress rarely correspond to objective, physiological responses that modulate memory performance (e.g., Hellhammer & Schubert, 2012).

Across all studies on stress and identification performance that we know of, only four combined both a meaningful interval between encoding and retrieval and physiological measures of stress [Peters, 1988, 1991 (experiment 2), 1997, (experiment 2); Rush et al., 2013]. It is important to note, however, that none of these used a more realistic event (e.g., staged crime paradigm). Furthermore, all but Peters (1988) relied on child or adolescent samples. The outcomes of these studies are mixed. Two were in line with the previously mentioned meta-analysis [i.e., negative effect of stress on target-present, but not target-absent line-up performance; Peters, 1988, 1991 (experiment 2)]. Peters (1997, experiment 2), however, provided support for our hypothesis that a meaningful retention interval is crucial when studying the effect of stress on memory: No effect was found after 6 months of retention, while negative effects materialized for both target-present and target-absent line-ups after a 15-minute interval. Finally, in an experiment with two targets, Rush et al. (2013) found no effect of stress on identification performance in target-present or target-absent line-ups for one target, but for a different target, stress had a positive effect on target-absent line-up performance, leading to significantly more correct rejections, fewer false identifications, and fewer don't know responses. There was no effect for target-present line-up performance for this target. Clearly, these results demonstrate that stress does not necessarily have a detrimental effect on eyewitness identification performance – but can even have positive effects.

The aim of the current study was to fill a gap in the literature by studying the effect of stress at encoding on eyewitness identification performance in a study high in ecological validity addressing the aforementioned concerns with previous research. Ecological validity was strived for by exposing all participants to a live staged crime and by inserting a 1-week interval between the event and the administration of the line-up. Insertion of such an interval comes much closer to practice in real cases (Behrman & Richards, 2005) than the minute-long intervals frequently employed. Moreover, the procedure is essential from a theoretical point of view as it enables the separation of possible stress effects on memory formation and retrieval.

To increase the degree of stress in the mock crime and thereby elicit strong and robust levels of stress at the time of memory encoding, participants in the high-stress condition were subjected to a highly stressful experience, namely the Maastricht Acute Stress Test (MAST; Smeets et al., 2012) during the staged crime. The MAST is a powerful, standardized procedure to induce stress in the laboratory that has been shown to

DOI: 10.1002/bsl

elicit robust subjective, autonomic and glucocorticoid stress responses. Participants in the low-stress condition received a control MAST. Measuring participants' salivary cortisol stress level and examining high versus low cortisol responders furthermore allowed us to assess individual differences in the responsiveness to the stress manipulation in stimulating the HPA axis (Meyer et al., 2013; Smeets, Dziobek, & Wolf, 2009). In line with neurobiological research on the effects of stress on memory, we predicted that stress would have a positive effect on eyewitness identification performance. That is, we expected participants in the high-stress condition to outperform those in the low-stress condition.

METHOD

Participants

In all, 127 participants (21 men; age range 18–63 years, M_{age} = 22.2, SD = 4.9) took part in return for course credit or a €10 voucher. Four participants (all women) did not return to the laboratory for the identification task, leaving 123 participants for the analyses. Participants comprised students (82.1%), people who gave no indication of their profession (10.6%), people who were employed (4.1%) and some who were doing an apprenticeship (3.3%). The students mostly majored in psychology (40.6%), medicine (25.7%), or health sciences (18.8%). The study was approved by the ethical committee of the faculty.

Line-ups and Line-up Construction

Target-absent and target-present line-ups were constructed for a male and a female target who were both 22 years old. Fifty percent of the participants viewed a line-up with the male target, 50% with the female target. Line-ups were composed of six photographs (shoulders up), numbered 1–6, which were arranged in two rows of three pictures (a simultaneous line-up). All foils and the replacement (i.e., innocent suspect in target-absent line-ups) fitted the general descriptions of the referring target, as determined in pilot work with N = 30 participants using the Doob and Kirshenbaum (1973) procedure. Effective sizes for target-present and target-absent line-ups, determined as Tredoux's E values, were high with a range of 4.6 to 5.5 (Tredoux, 1998, 1999).

Stress Induction versus No-stress Control Manipulation

The MAST (Smeets et al., 2012) is a concise procedure to reliably elicit robust subjective, autonomic and glucocorticoid stress responses. It comprises a 5-minute preparation phase in which the task is explained and a 10-minute acute stress phase that includes repeated exposure to cold pressor stress and mental arithmetic challenges. Specifically, participants have to immerse their hand into ice water (4 °C; Plexiglas box with an electrical cooler and a circulation pump; Julabo Labortechnik, Seelbach, Germany) during five trials that last between 60 and 90 seconds. Alternating with the hand immersion trials, participants are engaged in mental arithmetic challenges during which they have to count backwards as fast and accurately as possible in steps of 17 starting at 2043 for 45, 60 or 90 seconds. Whenever they count

DOI: 10.1002/bsl

too slowly or made a mistake, they receive negative feedback (i.e., to count faster and/or recommence at 2043). To increase task unpredictability and uncontrollability, participants are told that the order and duration of the hand immersion and mental arithmetic trials were randomly chosen by the computer and that they were videotaped so that their facial expressions could be analysed afterwards (Smeets et al., 2012).

The low-stress condition also comprises a 5-minute preparation phase and a 10-minute hand immersion phase, albeit in lukewarm water (25 °C), alternated with a simple counting task during which they had to repeatedly count from 1 to 25 at their own pace. The experimenter (the same one who will administer the line-up later) remains in the room to check participants' compliance with the instructions, but participants are not given any feedback on their performance and they are not videotaped. The duration and order of hand immersion and arithmetic trials parallel that of the MAST (see Smeets et al., 2012, study 3, for more details).

Salivary Cortisol Responses

Cortisol stress measures prior to and in response to the (control) MAST were obtained with synthetic Salivette (Sarstedt, Etten-Leur, the Netherlands) devices 5 minutes before ($t_{\text{pre-stress}}$) and three times after the MAST ($t_{+0\text{min}}$, $t_{+10\text{min}}$, $t_{+20\text{min}}$ with reference to the end of the stress or control procedure). Samples were stored at −20 °C until cortisol levels were determined by a commercially available luminescence immune assay kit (IBL, Hamburg, Germany). Mean intra- and inter-assay coefficients of variation are typically less than 5%, and the lower and upper detection limits were 0.015 mg/dL (0.41 nmol/L) and 4.0 mg/dL (110.4 nmol/L), respectively. Cortisol levels for two male low stress group participants could not be determined because there was insufficient saliva in the obtained samples for the analyses.

Design

Participants were randomly assigned to a 2 (stress condition: high vs. low) × 2 (target presence: present vs. absent) between-factors design. Identification accuracy (accurate vs. inaccurate) served as the dependent variable and was defined as the proportion of correct decisions across target-present and target-absent line-ups. We treated don't know answers in two different ways. In one analysis, they were coded as neither correct nor incorrect, i.e., don't know answers were treated as missing values. In an additional analysis, they were treated as rejections (see the Results section). Target gender had no effect on identification accuracy [$\chi^2(1, N = 112) = 0.23$, $p = 0.603$] and was not included in our analyses.

Procedure

Participants were instructed not to consume any foods or drinks, or to engage in physical exercise for at least 2 hours prior to session 1. After signing the informed consent form, the first saliva sample ($t_{\text{pre-stress}}$) was taken and participants were informed that the aim of the study was to examine cognitive and physical reactions to a stressful experience. They were not informed about the upcoming mock crime. After engaging in the MAST for 15 minutes, participants were informed that there would be a short "break"

DOI: 10.1002/bsl

before the task continued and the second saliva sample was taken (t_{+0min}). The information that the task would be continued was falsely provided to avoid decreases in stress level at this time (Smeets et al., 2012).

Next, the experimenter excused himself or herself, indicating that the saliva samples had to be placed into the freezer. Meanwhile, one of the two possible targets entered the room for about 1 minute (M = 52.2 seconds, SD = 20.1). Pretending that s/he was one of the previous participants, s/he took a phone from the desk and then left the room. When the experimenter re-entered the room, s/he searched for her/his phone. Following the participant's account of what had occurred in the past few minutes, participants were informed that they had witnessed a staged theft. They then completed a free recall form or a Self-Administered Interview (SAI; Gabbert, Hope, & Fisher, 2009) in another room. While writing down their testimony, the third (t_{+10min}) and fourth (t_{+20min}) saliva samples were taken 10 and 20 minutes after termination of the MAST. The SAI data were collected to investigate the possible beneficial value of retrieval support (as provided by the SAI) under varying levels of stress on immediate recall (i.e., no delay). These data are reported elsewhere (Krix et al., 2016). We hypothesized that the long time interval between completing these forms and performing the identification tasks (6–8 days) makes it unlikely that the recall task had an impact on identification performance. This was confirmed by preliminary analyses and, accordingly, interview was not included as a factor in the analyses. Specifically, the quantity of reported person and event details as well as the accuracy of these reports (total number of reported correct details divided by total number of details reported) did not predict accuracy of line-up decisions (all $p \geq .191$). This is in accordance with previous studies reporting that prior recall did not affect subsequent recognition (e.g., Howe et al., 2010; Marche, Brainerd, Lane, & Loehr, 2005).

Six to eight days after session 1 (M = 7.0, SD = 0.8), participants returned to the laboratory in which session 1 had ended and were informed that they were about to see a photographic line-up of the thief from session 1. They had not been given any information of what would be expected of them in session 2 beforehand. Participants were instructed that the target may or may not be in the line-up and they were also given the option to make a "don't know" decision. Subsequently, participants indicated their post-decision confidence on an 11-point scale ranging from 0 to 100%. No post-decision confidence ratings were obtained for don't know responses. Finally, participants were thanked and later debriefed via e-mail upon termination of data collection.

RESULTS

Manipulation Check: Cortisol Stress Responses

For each participant, cortisol increases were computed, defined as maximum cortisol level after the MAST or control task minus baseline level. A responder rate was then calculated representing participants in the MAST group with a cortisol increase (i.e., cortisol reactivity) $\geq$ 2.5 nmol/L (i.e., high cortisol responders; see, e.g., Kirschbaum, Pirke, & Hellhammer, 1993; Smeets et al., 2012). Based on this, 17 participants were categorized as low responders and 45 as high responders. A 3 (stress: high responders vs. low responders vs. low stress) × 4 (measurement: $t_{pre\text{-}stress}$ vs. t_{+0min} vs. t_{+10min} vs. t_{+20min}) ANOVA was run to confirm the differences between high and low cortisol

DOI: 10.1002/bsl

responders. Cortisol data were log-transformed before analysis, as Shapiro–Wilk tests of normality showed typical skewness of the data. Figure 1 depicts the salivary cortisol levels over time. Cortisol concentration at baseline did not differ across groups (all $p \geq .181$ and cortisol reactivity did not differ as a function of gender [Wald $\chi^2(1, N = 121) = 2.38$, $p = .123$, $b = 0.96$].

As expected, a statistically significant measurement × group interaction emerged [$F(6, 348) = 39.55$, $p < .001$; $\eta_p^2 = 0.41$]. Follow-up tests showed: a simple main effect of measurement within high cortisol responders [$F(3, 132) = 45.06$, $p < .001$; $\eta_p^2 = 0.51$], with cortisol increases from $t_{\text{pre-stress}}$ to $t_{+0\text{min}}$ and from $t_{+0\text{min}}$ to $t_{+10\text{min}}$ (all $p < .001$), but not from $t_{+10\text{min}}$ to $t_{+20\text{min}}$ ($p > .99$); a simple main effect of measurement within low cortisol responders [$F(3, 45) = 3.21$, $p = .032$; $\eta_p^2 = 0.18$], with only a statistically

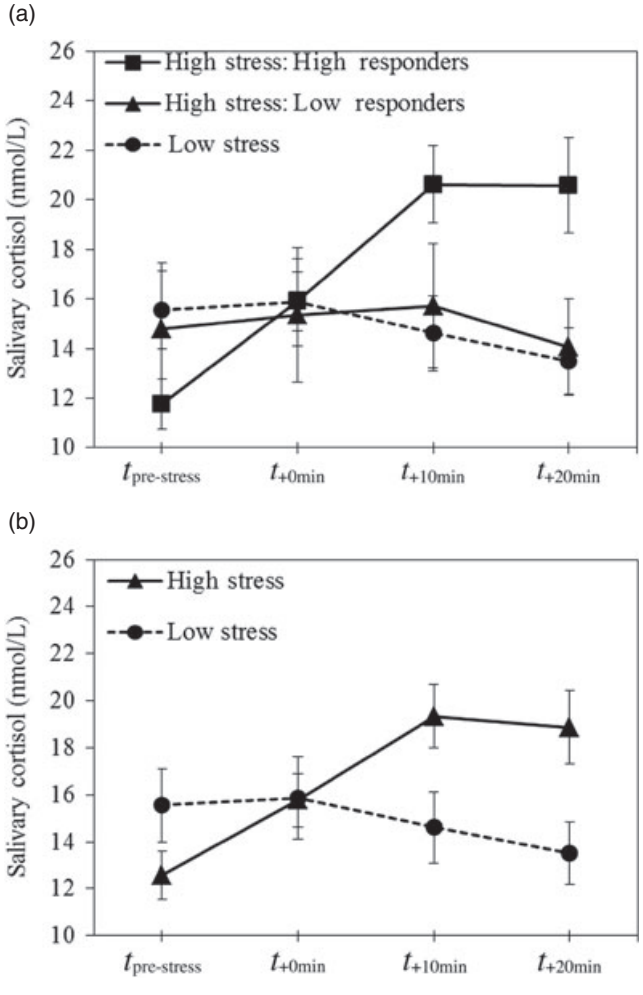

Figure 1. Salivary cortisol levels (nmol/L) over time for: (a) high responders (high-stress condition), low responders (high-stress condition) and low-stress participants; and (b) the high- and low-stress conditions. $T_{\text{pre-stress}}$, measurement before administration of the Maastricht Acute Stress Test (MAST)/control task; $t_{+0\text{min}}$, measurement upon termination of the MAST/control task; $t_{+10\text{min}}$, measurement 10 min after termination of the MAST/control task; $t_{+20\text{min}}$, measurement 20 min after termination of the MAST/control task.

DOI: 10.1002/bsl

significant decrease between t_{+10min} and t_{+20min} ($p = .005$); and for low-stress participants, cortisol decreases from t_{+0min} to t_{+10min} and from t_{+10min} to t_{+20min}, but not from $t_{pre\text{-}stress}$ to t_{+0min}. A 2 (stress: high vs. low) × 4 (measurement: $t_{pre\text{-}stress}$ vs. t_{+0min} vs. t_{+10min} vs. t_{+20min}) ANOVA yielded similar results. Analogously, the high- and low-stress groups differed significantly in cortisol reactivity [$F(1, 119) = 33.43$, $p < .001$; $\eta_p^2 = 0.22$]. The salivary cortisol levels over time can be found in Figure 1.

The results of these analyses confirm the success of our stress manipulation as well as the general finding in the literature that cortisol elevation occurs with temporary delay after a stressor has unfolded its impact.

Impact of Stress on Identification Accuracy

An overview of the identification outcomes in the high and low stress conditions can be found in Table 1. Using binary logistic regression analysis, we tested the effect of stress (high vs. low) and target presence (target-present vs. target-absent) on identification accuracy (correct vs. incorrect). Of the 123 participants, 11 gave a don't know response and were omitted from the analysis. Initially, we included main effects and the interaction in the equation. We then sequentially excluded nonsignificant two-way interactions. Here, we focus on (main or interaction) effects of stress and will not report other effects. The stress × target-presence interaction as well as the main effect of stress were statistically nonsignificant [Wald $\chi^2(1, N = 112) \leq 1.23$, p-values $\geq .267$, b-values ≤ 0.58].

We followed the current practice of using receiver operating characteristic (ROC) analysis as an alternative for establishing effects on identification accuracy in line-up procedures (Gronlund et al., 2012; Gronlund, Wixted, & Mickes, 2014; Mickes, Flowe, & Wixted, 2012). Although ROC analysis is mathematically related to logistic regression (Austin & Steyerberg, 2012) and would not be expected to yield a different outcome for the effect of stress on accuracy, the construction and comparison of separate ROC curves for the stressed and non-stressed conditions will provide a slightly different perspective on the possible existence of a stress effect. A ROC curve is constructed by plotting the rate of correct identifications against the rate of false identifications for separate levels of confidence (rated by participants on a scale from 0 to 100 on an 11-point Likert scale).

Figure 2 shows the ROC curves for the two stress conditions, together with a positive diagonal line that represents chance performance. Points on this line would correspond to equal rates of correct and false identifications. The different points that form the ROC curves represent pairs of correct and false identification rates at different levels of confidence. Looking at the ROC curve for the stressed group, the point located at

Table 1. Percentage of identification outcomes for the high- and low-stress conditions

Stress condition	High stress	Low stress	Total
Target-present line-ups	$n = 30$	$n = 30$	$n = 60$
Hits (correct identifications)	53.3	53.3	53.3
Foil choices	6.7	0.0	3.3
False rejections	33.3	40.0	36.7
Don't know responses	6.7	6.7	6.7
Target-absent line-ups	$n = 32$	$n = 31$	$n = 63$
Correct rejections	59.4	80.6	69.8
False alarms	25.0	12.9	19.0
Don't know responses	15.6	6.5	11.1

DOI: 10.1002/bsl

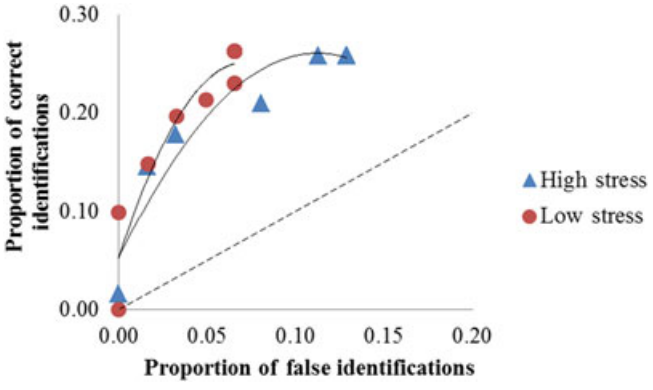

Figure 2. Receiver operating characteristic (ROC) plots for the high- and low-stress conditions. The dashed line represents chance performance.

the upper right of the curve shows the rates of correct and false identifications, determined over all confidence levels taken together. One data point further to the left, we find the rates of correct and false identifications computed over all confidence levels taken together, with the exception of the lowest level of confidence. As we move further towards the left, we ultimately end at the lower right-most point, where only responses that were rated with the highest possible confidence (100%) were included for the computation. Although, based on this, we would expect to see 11 points in the curve, in fact we observed fewer points. That is because particular combinations of correct and false identification rates proved to be identical for multiple levels of confidence.

To assess a possible effect of stress on identification performance, we calculated the area under the curve for the relevant range of false identification rates. For the high-stress group, this partial area under the curve (or pAUC) pertains to a range of false identifications from 0 to 0.13; for the low-stress group the relevant range for determining the pAUC is 0–0.07. Using a bootstrap analysis of the pROC package in R (Robin et al., 2011), we found the pAUC for the ROC curve to be .025 (95% CI [.012, .040]) for the high-stress group, and .013 (95% CI [.006, .022]) for the low-stress group. Note that the pAUC for the ROC curve of the high-stress group is higher in spite of the fact that this curve consistently lies below the ROC curve for the low-stress group. This is caused by the fact that the range of false identification rates is larger for the high-stress group than for the low-stress group. In order to be able to make a fair comparison between the two curves, we compared them for a limited range (i.e., from 0 to 0.07) or for an extended range (from 0 to 0.13, where an extrapolation of the ROC curve for the low-stress group is required). For our data, neither strategy of comparison yielded a statistically significant difference, indicating that accuracy did not differ as a function of stress level (we found the former standardized difference between the pAUC values to be -0.35, $p = .73$, and the latter standardized difference between the pAUC values to be -0.50, $p = .62$).

Additional Analyses

Other possible ways of analysing these data include: implementing three levels of stress (high responders vs. low responders vs. low stress participants); using participants' individual cortisol reactivity score as a measure of stress rather than forming two or three

DOI: 10.1002/bsl

categories; and including don't know responses as non-selections in the model rather than excluding these responses. Treating the data in those different ways using regression analyses consistently led to result patterns analogous to the ones reported earlier. Note that no additional ROC analyses were performed because no confidence ratings were obtained for don't know responses.

Power Analysis

Our data showed that level of stress, dichotomized into low vs. high, did not exert a marked effect on identification accuracy. To examine the possibility that the absence of a stress effect in our study may have been related to a lack of power, we conducted a power analysis using the pwr package in R, version 3.1.2 (2014). For a one-sided test of the difference of two independent proportions, assuming a true effect size of $h = -0.31$ (based on Deffenbacher et al., 2004), the power of our statistical test was 0.50 (0.53 when treating don't know answers as rejections). Additionally, we attempted to assess the power of the analysis using cortisol reactivity as the predictor in a logistic regression model. Because continuous scales provide richer information than a crude dichotomy, the power for this continuous predictor is likely to be higher than for a dichotomous predictor. However, such a calculation requires the estimated probability of a correct identification for the average level of the continuous predictor (i.e., for an average cortisol level), and to specify the estimated probability of a correct identification at a cortisol level one standard deviation above the mean. These two probabilities would then be used to compute a standardized sort of odds ratio, which is taken to represent the size of the effect of the continuous variable. However, we cannot make realistic assumptions about these probabilities, due to a lack of prior research on the effect of cortisol level on eyewitness identification accuracy. In addition, we would need to specify the population distribution of cortisol level. To conclude, it seems that we currently lack the necessary prior information that would allow us to make a credible statement of the power of our significance test of the effect of cortisol level on identification accuracy.

Exploratory Confidence–Accuracy Characteristic Analysis[1]

Although it is not the focus of this paper, the current data can also inform us about the confidence–accuracy relationship across different stress conditions. Figure 3 shows the confidence–accuracy characteristic (CAC) curve (cf. Mickes, 2015). CAC analysis focuses only on suspect identifications. Innocent suspects selections were established by dividing the number of foil selections by the number of line-up members (Palmer, Brewer, Weber, & Nagesh, 2013; Sauerland, Stockmar, Sporer, & Broers, 2013). Confidence categories were collapsed as low (30–60%), medium (70–80%), and high (90–100%). Suspect identifications with confidence $< 30\%$ did not occur. The CAC plot indicates that, for the high-confidence category, accuracy in both stress conditions was almost perfect. In other words, a high-confidence suspect identification was as likely to be correct when the encoding situation was stressful as when the encoding situation was not stressful. For identifications made with a low or medium level of

[1] We would like to thank Laura Mickes for suggesting this analysis

DOI: 10.1002/bsl

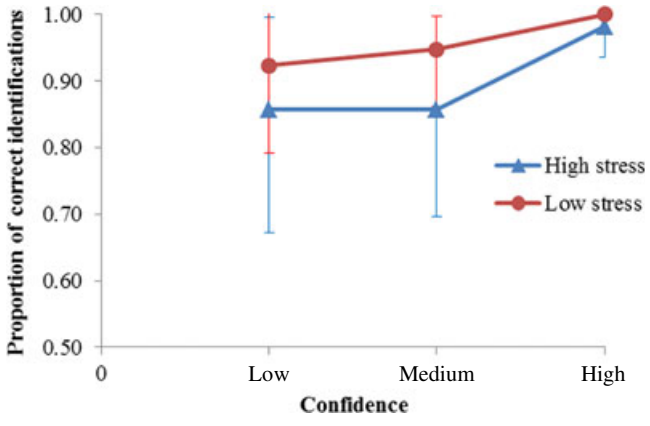

Figure 3. Confidence–accuracy characteristic (CAC) analysis for the high- and low-stress conditions. Low confidence refers to 30–60% confidence (suspect selections with confidence < 30% did not occur), medium confidence refers to 70–80% confidence, and high confidence to 90–100%. The bars represent standard error bars.

confidence, there was a tendency for accuracy to be lower for the high-stress condition than for the low-stress condition. These results should be interpreted with caution, because of the small number of data points that went into this analysis (32 guilty suspect identifications +12/6 foil innocent suspect identifications).

DISCUSSION

The aim of the current study was to examine the impact of stress at encoding on recognition performance. It is the first one to combine four methodological characteristics that are essential when studying this issue. First, we used a staged crime and inserted a 1-week retention interval between encoding and retrieval to increase ecological validity, which has been shown to be an important moderator of the stress–accuracy relationship (Deffenbacher et al., 2004). Second, the insertion of this 1-week interval ensured that the current design examined the effect of stress on memory formation rather than on retrieval. Third, the use of a powerful and standardized procedure for stress induction in the laboratory (MAST; Smeets et al., 2012) allowed us to exert relatively high levels of stress during a mock crime in a laboratory experiment. Fourth, by measuring participants' salivary cortisol stress level and examining high- versus low-cortisol responders, we assessed individual differences in the responsiveness to the stress manipulation in stimulating the HPA axis. Finally, we increased ecological validity by using two targets as stimulus persons rather than one.

Stress was expected to have a positive effect on eyewitness identification performance. Our results, however, showed that stress had no impact on identification performance in target-present or target-absent line-ups using both regression and ROC analyses. Although to some extent in conflict with the literature in the field of neurobiology of learning and memory (but see Payne et al., 2007; Schwabe &Wolf, 2010; Smeets, Otgaar, Candel, & Wolf, 2008), these findings are in line with some earlier studies that combined both a meaningful retention interval between encoding and retrieval and measurement of objective indicators of stress (Peters, 1997, experiment 2;

DOI: 10.1002/bsl

Rush et al., 2013). Two other studies, however, found a negative effect of stress (Peters, 1988, 1991, experiment 2) on identification performance. Methodological differences, and hence heterogeneity as to the obtained levels of stress, might underlie these conflicting findings. Specifically, when comparing the four previous studies with each other, we found that the stress-inducing events were quite diverse, including a fake fire alarm (Peters, 1997, experiment 2), inoculation at a medical clinic (Peters, 1988), having one's head rubbed by a stranger until attempting to avoid the rubbing (Peters, 1991, experiment 2), and performing the Trier Social Stress Test (Rush et al., 2013). None of these studies, however, used a (more realistic) staged crime paradigm (cf. Deffenbacher et al., 2004). Additionally, most (with the exception of Peters, 1988) relied on child or adolescent samples, in which stress might have different effects than in adults (Quas, Rush, Yim, & Nikolayev, 2013). In line with these findings, we also did not find stress to have a meaningful impact on the confidence–accuracy relationship. Specifically, the CAC analysis tentatively indicated that high-confidence suspect identifications are equally likely to be correct when the encoding situation was stressful and when it was not stressful. For identifications made with a low or medium level of confidence, there was a tendency for accuracy to be lower for the high-stress condition than for the low-stress condition. Whether or not these differences are meaningful needs to be established in future research.

When thinking about possible explanations for our findings, one interpretation of our data is that stress has differential effects on face recognition versus recall. Specifically, research in neurobiology has shown more pronounced positive effects of elevated levels of the stress hormone cortisol on recall than on recognition (Het et al., 2005). While analyses of the recall data collected in the current study also showed no effect of stress (Krix et al., 2016), this result cannot speak to the effect of stress on recall, because the recall and encoding phases followed closely on each other. Another explanatory account refers to the level of stress that can be induced in the laboratory. While the MAST is a potent stress manipulation that is known to evoke reliable and strong stress reactions in the laboratory (Smeets et al., 2012), and the cortisol data confirmed reliable differences in physiologically experienced stress between the two groups, this may still not be as high as the extreme stress levels reached when witnessing violent crimes. Detrimental effects of stress on memory might be expected at the high and extreme end of the stress scale, as demonstrated in a field study of military personnel. Morgan et al. (2004) had officers participating in a mock prisoner of war camp identify their interviewers under high- and low-stress interrogations. In this setting, stress had adverse effects on hits and false alarms, while the number of false rejections decreased for one line-up presentation mode (eight-person simultaneous photographic line-ups). No effect of stress on false rejections materialized for other presentation modes (15-person simultaneous live line-ups; 16-person sequential photographic line-ups) or for correct rejections. Unfortunately, however, this study employed only a short retention interval, thus confounding the effects of stress on memory formation and retrieval. The study also failed to control for familiarity and exposure to the interviewers (outside of the context of the experiment) sufficiently. Field studies inducing very high stress levels (while avoiding the ethical constraints that apply in this field) and implementing a longer retention interval, would be extremely informative in this matter.

A limitation of the study is that we neither collected information about female participants' menstrual cycle nor excluded women using hormonal contraception, although these can influence cortisol reactivity (Kirschbaum et al., 1999). As a

DOI: 10.1002/bsl

consequence, some female participants' cortisol reactivity may have been reduced. Importantly, however, the cortisol levels in the stress group were significantly elevated by the stress task, while cortisol levels in the low-stress group constantly decreased, as expected with the circadian rhythm. Hence, in both the high- and the low-stress groups, the course of the cortisol levels was as expected. Consequently, not collecting information about the menstrual phase seems to be unproblematic for our results.

We would like to conclude with a note that from a theoretical perspective, methodological rigor is essential when researching the impact of stress on eyewitness performance. The combination of a sufficient retention interval between encoding and retrieval and the use of a powerful stress induction are indispensable to this end. In line with our rationale, a comparison of studies following those guidelines and those that do not showed that the outcome of stress studies varies as a function of methodological differences in the operationalization of stress. Note that the discussed measures not only enhance the ecological validity of the research, but are ultimately a prerequisite for ensuring the overall validity of the findings and conclusions. The current paper provides a good model and starting point to a new line of research studying the impact of stress on eyewitness (identification) testimony. It seems probable that such research will challenge the assumption that there is a simple relationship between stress and eyewitness memory (Deffenbacher et al., 2004) and replace this with research that comes closer to what happens in the real world.

REFERENCES

Austin, P. C., & Steyerberg, E. W. (2012). Interpreting the concordance statistic of a logistic regression model: Relation to the variance and odds ratio of a continuous explanatory variable. *BMC Medical Research Methodology*, *12*, 82. doi:10.1186/1471-2288-12-82.

Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, *29*, 279–301. doi:10.1007/s10979-005-3617-y.

Bothwell, R. K., Brigham, J. C., & Pigott, M. A. (1987). An exploratory study of personality differences in eyewitness memory. *Journal of Social Behavior and Personality*, *2*, 335–43.

Brigham, J. C., Maass, A., Martinez, D., & Wittenberger, G. (1983). The effect of arousal on facial recognition. *Basic and Applied Social Psychology*, *4*, 279–93. doi:10.1207/s15324834basp0403_6.

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. doi:10.1007/s10979-004-0565-x.

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups — partial remembering. *Journal of Police Science and Administration*, *18*, 287–93.

Gabbert, F., Hope, L., & Fisher, R. P. (2009). Protecting eyewitness evidence: Examining the efficacy of a self–administered interview tool. *Law and Human Behavior*, *33*, 298–307. doi:10.1007/s10979-008-9146-8.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221–8. doi:10.1016/j.jarmac.2012.09.003.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*, 3–10. doi:10.1177/0963721413498891.

Hellhammer, J., & Schubert, M. (2012). The physiological response to Trier Social Stress Test relates to subjective measures of stress during but not before or after the test. *Psychoneuroendocrinology*, *37*, 119–24. doi:10.1016/j.psyneuen.2011.05.012.

Het, S., Ramlow, G., & Wolf, O. T. (2005). A meta-analytic review of the effects of acute cortisol administration on human memory. *Psychoneuroendocrinology*, *30*, 771–84. doi:10.1016/j.psyneuen.2005.03.005.

Hosch, H. M., & Bothwell, R. K. (1990). Arousal, description and identification accuracy of victims and bystanders. *Journal of Social Behavior and Personality*, *5*, 481–8.

Howe, M., Candel, I., Otgaar, H., Malone, C., & Wimmer, M. C. (2010). Valence and the development of immediate and long-term false memory illusions. *Memory*, *18*, 58–75. doi:10.1080/09658210903476514.

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The "General Acceptance" of psychological research on eyewitness testimony: A survey of the experts. *The American Psychologist*, *44*, 1089–98. doi:10.1037//0003-066X.44.8.1089.

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "General Acceptance" of eyewitness testimony research: A new survey of the experts. *The American Psychologist*, *56*, 405–16. doi:10.1037//0003-066X.56.5.405.

Kirschbaum, C., Kudielka, B. M., Gaab, J., Schommer, N. C., & Hellhammer, D. H. (1999). Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary–adrenal axis. *Psychosomatic Medicine*, *61*, 154–62.

Kirschbaum, C., Pirke, K. -M., & Hellhammer, D. H. (1993). The ‚Trier Social Stress Test': A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*, 76–81. doi:10.1159/000119004.

Krix, A. C., Sauerland, M., Raymaekers, L. H. C., Memon, A., Quaedflieg, C. W. E. M., & Smeets, T. (2016). Eyewitness evidence obtained with the Self-Administered Interview© is unaffected by stress. *Applied Cognitive Psychology*, *30*, 103–12. doi:10.1002/acp.3173.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. Nature Reviews. Neuroscience, *7*, 54–64. doi:10.1038/nrn1825

Marche, T. A., Brainerd, C. J., Lane, D. G., & Loehr, J. D. (2005). Item method directed forgetting diminishes false memory. *Memory*, *13*, 749–58. doi:10.1080/09658210444000377.

Meyer, T., Smeets, T., Giesbrecht, T., Quaedflieg, C. W. E. M., & Merckelbach, H. (2013). Acute stress differentially affects spatial configuration learning in high and low cortisol responding healthy adults. *European Journal of Psychotraumatology*, *4*, 19854. doi:10.3402/ejpt.v4i0.19854.

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 and Cognition*, *4*, 93–102. doi:10.1016/j.jarmac.2015.01.003.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied*, *18*, 361–76. doi:10.1037/a0030609.

Morgan, C. A., 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*, 265–79. doi:10.1016/j.ijlp.2004.03.004.

Palmer, M., 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. doi:10.1037/a0031602.

Payne, J. D., Jackson, E. D., Siobhan Hoscheidt, L. R., Jacobs, W. J., & Nadel, L. (2007). Stress administered prior to encoding impairs neutral but enhances emotional long-term episodic memories. *Learning & Memory*, *14*, 861–8. doi:10.1101/lm.743507.

Peters, D. P. (1988). Eyewitness memory in a natural setting. In Gruneberg, M. M., Morris, P. E., & Sykes, R. N. (Eds.), *Practical aspects of memory: Current research and issues* (pp. 89–94), Memory in everyday life *1*. Chichester, UK: Wiley.

Peters, D. P. (1991). The influence of stress and arousal on the child witness. In Doris, J. (Ed.), *The suggestibility of children's recollections* (pp. 60–76). Washington, DC: American Psychological Association.

Peters, D. P. (1997). Stress, arousal, and children's eyewitness memory. In Stein, N. L., Ornstein, P. A., Tversky, B., & Brainerd, C. J. (Eds.), *Memory for everyday and emotional events*. Mahwah, NJ: Erlbaum.

Phelps, E. A. (2004). Human emotion and memory: Interactions of the amygdala and hippocampal complex. *Current Opinion in Neurobiology*, *14*, 198–202. doi:10.1016/j.conb.2004.03.015.

Quas, J. A., Rush, E. B., Yim, I. S., & Nikolayev, M. (2013). Effects of stress on memory in children and adolescents: Testing causal connections. *Memory*, *22*, 616–32. doi:10.1080/09658211.2013.809766.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Biochemistry*, *12*, 77. doi:10.1186/1471-2105-12-77.

Roozendaal, B., & McGaugh, J. L. (2011). Memory modulation. *Behavioral Neuroscience*, *125*, 797–824. doi:10.1037/a0026187.

Rush, E. B., Quas, J. A., Yim, I. S., Nikolayev, M., Clark, S. E., & Larson, R. P. (2013). Stress, interviewer support, and children's eyewitness identification accuracy. *Child Development*. doi:10.1111/cdev.12177.

Sauerland, M., Stockmar, A. K., Sporer, S. L., & Broers, N. J. (2013). The reliability of identification evidence with multiple lineups. *The European Journal of Psychology Applied to Legal Context*, *5*, 49–71. Retrieved from http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1889-18612013000100003

Schwade, L., & Wolf, O. T. (2010). Learning under stress impairs memory formation. Neurobiology of Learning and Memory, *93*, 183–188. doi:10.1016/j.nlm.2009.09.009

Schwabe, L., Joëls, M., Roozendaal, B., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. Neuroscience and Biobehavioral Reviews, *36*, 1740–1749. doi:10.1016/j.neubiorev.2011.07.002

Schwarze, U., Bingel, U., & Sommer, T. (2012). Event-related nociceptive arousal enhances memory consolidation for neutral scenes. *The Journal of Neuroscience*, *32*, 1481–7. doi:10.1523/JNEUROSCI.4497-11.2012.

Smeets, T., Cornelisse, S., Quaedflieg, C. W. E. M., Meyer, T., Jelicic, M., & Merckelbach, H. (2012). Introducing the Maastricht Acute Stress Test (MAST): A quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses. *Psychoneuroendocrinology*, *37*, 1998–2008. doi:10.1016/j.psyneuen.2012.04.012.

Smeets, T., Dziobek, I., & Wolf, O. T. (2009). Social cognition under stress: Differential effects of stress-induced cortisol elevations in healthy young men and women. *Hormones and Behavior*, *55*, 507–13. doi:10.1016/j.yhbeh.2009.01.011.

Smeets, T., Otgaar, H., Candel, I., & Wolf, O. T. (2008). True or false? Memory is differentially affected by stress-induced cortisol elevations and sympathetic activity at consolidation and retrieval. *Psychoneuroendocrinology*, *33*, 1378–86. doi:10.1016/j.psyneuen.2008.07.009.

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior*, *22*, 217–37. doi:10.1023/A:1025746220886.

Tredoux, C. G. (1999). Statistical considerations when determining measures of lineup size and lineup bias. *Applied Cognitive Psychology*, *13*, 9–26. doi:10.1002/(SICI)1099-0720(199911)13:1+%3CS9::AID-ACP634%3E3.0.CO;2-1.

Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology*, *23*, 151–61. doi:10.1002/acp.1463.

Wolf, O. T. (2009). Stress and memory in humans: Twelve years of progress? *Brain Research*, *1293*, 142–54. doi:10.1016/j.brainres.2009.04.013.