A Systematic Review and Meta-Analysis Examining the Effect of Stress at Encoding on Line-up Performance

Milton Anthony Gering

A dissertation submitted as part of the requirements for the award of the degree of Master of Arts in Psychological Research

ACSENT Laboratory

Department of Psychology

University of Cape Town

2021

Supervisors:

Colin G. Tredoux

Alicia Nortje

COMPULSORY DECLARATION

This work has not been previously submitted in whole, or in part, for the award of any degree. I know the meaning of plagiarism and declare that all the work in the document is my own work and that each significant contribution to, and quotation in, this dissertation from the work, or works, of other people has been attributed, and has been cited and referenced.

Signed by candidate — Signed

01/03/2021 — date

Acknowledgements

Many people have helped me along this research journey, both directly and indirectly, and so there are too many to acknowledge individually. Some effort will be made, nonetheless.

Foremost, a thanks to Colin Tredoux who made this project appear simple at the onset, a necessary deception without which the work may not have begun. And who then provided nudges to keep raising the quality of the work.

To the eyewitness group at UCT, a big thank you for listening to all the iterations of this project and asking the important questions at the conceptual stage and providing support throughout. A great thanks as well the academics outside of UCT who took the time to let me bounce ideas off of them and provided resources to take me further.

To my family and friends, who both supported and distracted me throughout this process, my deepest gratitude as this would have been a far more arduous ordeal without you all. Particularly those of you who gave some part of this project an edit.

Finally, acknowledgement must be made to all the academics in the field who provided the foundations and materials for this work. It is only possible to go so far because of the trails already laid.

Table of Contents

Abstract

Although much research has been conducted on the effect of stress on eyewitness memory, the answer to this question remains unclear. Whereas a previous meta-analysis (Deffenbacher et al., 2004) concluded that stress negatively affects eyewitness identification ability, recent studies have shown a lack of consensus. As most crimes are stressful events and eyewitness evidence is influential in courts; clarity on the effect of stress is important to legal systems around the world. It is difficult to summarise extant research as many studies use differing methods making the source of disagreement unclear. Added to that, many studies report insufficient detail needed to judge the rigour of research designs, and thus the effects of stress. The present systematic review attempts to synthesise the literature and presents an analysis using recent meta-analytic techniques that allow for the influence of moderator variables to be quantified. It shows that the effect of stress at encoding on line-up decisions is not clear, with studies reporting both positive and negative effects, and examines reasons for differences in effects found between studies. A finding of note is that sequential or simultaneous line-up presentation has a moderating effect of stress on line-up performance. Additionally, a multilevel model shows that using continuous, rather than dichotomous, measures of stress may clarify the stress-performance relationship. Recommendations for further research are made in the hope that new studies can answer the important question of whether witnesses who experience high levels of stress at encoding are likely to make better or worse line-up decisions.

List of Figures

List of Tables

Literature Review

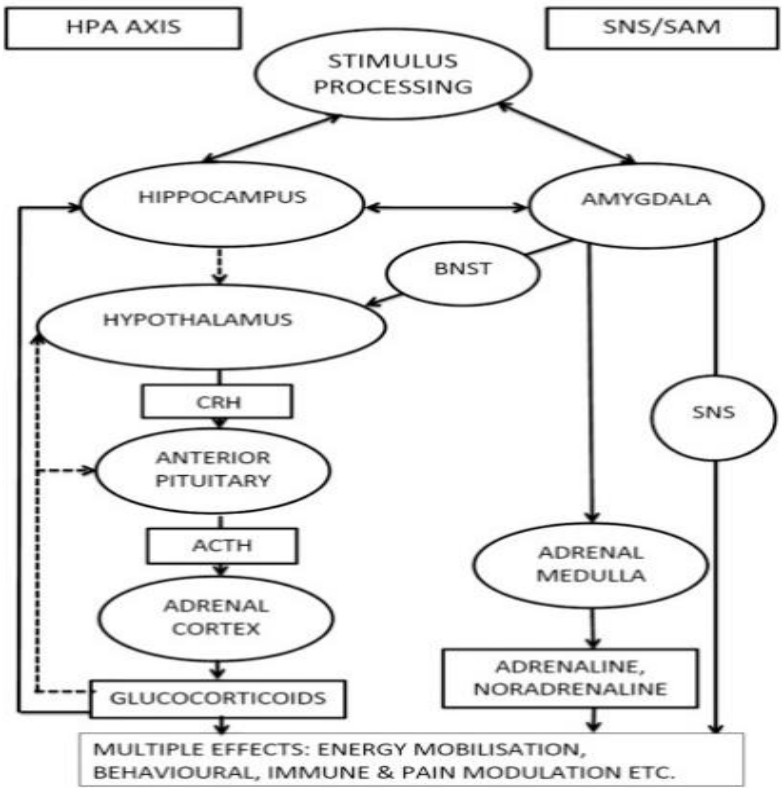
Witnessing a crime is typically a stressful experience. As witnesses are often required to recall the details of the crime and identify the perpetrator, it is important to understand how crime-related stress affects witness memory. Studies of eyewitness memory have examined the effects of stress at encoding on subsequent recall and recognition, as well as the effect of stress experienced during recall itself (Deffenbacher et al., 2004). Findings have shown that stress induced at the recall or recognition phase impairs memory, yet the effect of stress at encoding is less clear (Christianson, 1992; Het et al., 2005). The lack of clarity is reflected in the beliefs of both experts familiar with the research and lay people (Marr et al., 2020). This may be related to methodological differences between experiments (from within both applied and basic research paradigms) but may also be the result of a lack of consistency within the stress and eyewitness literature specifically (Christianson, 1992; Deffenbacher et al., 2004; Sauerland et al., 2016). As stress is not thought to have a simple linear relationship, different levels of stress might vary performance (Yerkes & Dodson, 1908). Differences in the effect of stress at encoding have been found between subsequent recall and recognition tasks (Het et al., 2005). A meta-analysis looking at various stress and memory experiments outside of the eyewitness literature found differences in method to moderate many of the differences found between studies (Shields et al., 2017). Potential confounds in method include the delay between learning and recognition, individual differences in stress reactivity, as well as the type of stimulus used (Marr et al., 2020). Pairing different stressors and event types has resulted in inconsistent results with no clear theoretical explanation (Sandi, 2013). In order to establish the link between stress and subsequent line-up decision accuracy in a mock crime paradigm as typically used by eyewitness researchers, these methodological differences must be systematically explored.

Stress and Memory

Witnessing a crime is a potentially dangerous experience and thus stressful, particularly when the witness is also the victim (Deffenbacher et al., 2004). During episodic experiences of stress, known as acute stress, a bodily response triggers the release of hormones which prepare the body for action (Joels et al., 2011). This can occur through the activation of two systems: the fast-acting sympathetic-adrenal-medullary (SAM) axis, or the slower-acting hypothalamic-pituitary-adrenal (HPA) axis shown in Figure 1. An activated SAM system increases activity of the sympathetic nervous system, which in turn releases adrenaline that stimulates the release of norepinephrine in the brain (Joels et al., 2011). The norepinephrine first activates the amygdala, which interacts with the hippocampus, caudate nucleus, and frontal lobes (Packard et al., 1994; Wolf et al., 2016). These areas are also activated through the HPA response as they contain glucocorticoid receptors (Arnsten, 2009; Gray et al., 2017; Lupien et al., 2007). HPA axis activation triggers secretion of glucocorticoids, primarily cortisol in humans. Several factors influence the effect of acute stress on memory, namely; the amount of stress experienced, the subsequent activation of the SAM or HPA axis, the memory processes being used and the time since onset of stress (Shields et al., 2016). While the SAM and HPA axes both activate the amygdala and hypothalamus, the SAM activation appears to strengthen consolidation of memory while HPA axis activation weakens it (Roozendaal et al., 2009). As a result of these contrasting effects, mild or moderate stress during encoding has been shown to improve memory, yet intense stress impairs memory (Drexler & Wolf, 2017).

Figure 1

The Stress Response



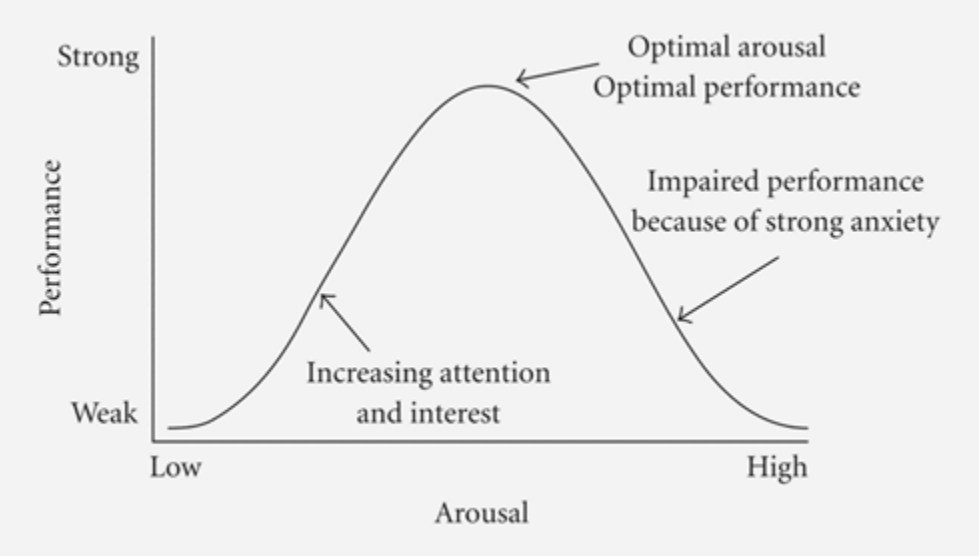
Note. From "The neurobiology of stress" by Murison, (2016). Hypothalamic–pituitary–adrenal (HPA). Sympathomedullary pathway (SAM). Bed nucleus of the stria terminalis (BNST). Corticotropin-releasing hormone (CRH). Sympathetic nervous system (SNS). Adrenocorticotropic hormone (ACTH).

A unidimensional representation of the body's response to stress has been graphically represented as an inverted U-shape, known as the Yerkes-Dodson law in which the positive effects of stress on encoding reach a peak and subsequently begin to drop after the level of stress becomes too high (Yerkes & Dodson, 1908; Drexler & Wolf, 2017). This curve (Figure 2) may not be symmetrical, having a steep drop off as stress increases which results in catastrophic forgetting when information is not adequately encoded (Sandi, 2013). This

catastrophic cusp is not shown in Figure 2 but would be on the right side of the curve, dropping more steeply than the left side rises. This non-linear relationship between stress and memory is attributable to the different effects produced by activation of either the SAM or HPA axis (Shields et al., 2016). If asymmetrical, it would suggest that the impairment from the HPA axis produces a greater effect than the benefits of the SAM axis activation.

Figure 2

The Yerkes-Dodson Curve



Note. From "The temporal dynamics model of emotional memory processing: A synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law," by Diamond et al. (2007).

This representation of the relationship between arousal and performance has been critiqued as being too simplistic (Hanoch & Vitouch, 2004). The basis for this conclusion stems from the varied nature of stressors, which extend past those tested by Yerkes and Dodson (1908) when developing this law, the outcomes measured by the authors and the

participants of the study. As the original experiments were conducted with mice and focused on habit formation, the dual critique is that studies on mice are not sufficient to understand human behaviour and that the behaviour measured by Yerkes and Dodson is not the only behaviour to which their law is applied (Hanoch & Vitouch, 2004). Reanalysis of their results by Bäumler and Lienert (1993) found that the effect varied based on whether performance was measured as hits or errors, with only hits producing the U-shape. Additionally, critics argue that arousal is not so simple a construct that it can be represented on a unidimensional axis. Arousal as a construct has been used to discuss both physiological changes, such as increased heart rate, as well as changes in mental state such as anxiety and alertness (Hanoch & Vitouch, 2004). These states are not congruent despite language used and do not always covary. As such, different experiments using similar terms and measures may be investigating different constructs. This is a potential flaw in the Yerkes-Dodson law which has not been adequately addressed.

Nonetheless, the Yerkes-Dodson law has also been built upon by subsequent researchers. Easterbrook's (1959) cue-utilisation theory proposes that at high levels of stress focus is narrowed, and only a smaller set of information can be attended. This results in a poorer subset of information being encoded. At medium levels, the whole set of information can be attended, which allows for richer encoding. This explains the drop off in performance at high arousal, while the orienting response to stress explains the initial increase in performance from low to moderate arousal (Easterbrook, 1959). It has also been argued that cue-utilisation only works in this manner when the task and stressor are related. As emotion serves an evolutionary function by aiding certain behaviours, there is an internal logic that emotion is not a general performance enhancer (Hanoch & Vitouch, 2004). However, studies in the field of neuropsychology which test memory using stimuli such as word lists often consider the stress effect generally. Stress is often induced by some lab-based stressor which

is unrelated to the task (Shields et al., 2016). Yet, a seemingly pervasive flaw in the stress literature is that stress is often dichotomised (Deffenbacher et al., 2004; Shields et al., 2016; Shields et al., 2017). As the theoretical assumptions treat stress as a continuous measure, experiments testing this assumption with a stress vs control paradigm are missing the necessary resolution to test the underlying theory. These experiments only look at two levels of stress; too few to study a non-linear relationship. To see such a relationship, a minimum of 3 levels of the independent variable must be considered. The concepts above will now be considered in relation to both the Yerkes-Dodson law and cue utilisation theory and how these apply specifically to the stress and eyewitness literature.

Attention and Memory

Given that one must notice and attend to an event for encoding to occur, attentional processes play a role in memory. One cannot always attend to everything in the environment as our attentional resources are limited (Easterbrook, 1959).There have been cases of inattentional blindness where people fail to notice a critical event for which they are present, as their attention was elsewhere (Pickel, 2015). One explanation is that people who focus on goal-related tasks do not attend to unrelated stimuli, resulting in them missing the critical event. Another is that attention acts like a beam of light - information on which the beam is centred is better encoded. Levine and Edelstein (2009) consider goal-related behaviour to attract the focus of this attentional beam, while other information falls on the periphery. It is expected that threatening stimuli will draw greater attentional resources because not doing so can endanger the agent's life (Hanoch & Vitouch, 2004). Plieger et al. (2016) found that stress facilitates selective attention at encoding and subsequent recall performance for cognitively undemanding tasks but impairs selective attention on more demanding tasks. Moreover, if multiple stimuli are present, which would make a task more cognitively

demanding, some stimuli will be prioritised. This results in poor memory for peripheral or less salient stimuli (Pickel, 2015). As criminal events often involve several elements, such as the perpetrator's actions, details of their appearance, as well as a face to memorise, some aspects of a crime might be better remembered than others. If stress narrows a witness's focus, they may remember some aspects of an event well while forgetting other details (Plieger et al., 2016). This supports the cue-utilisation theory and is congruent with the idea of emotion being beneficial to specific, related behaviours (Hanoch & Vitouch, 2004).

Emotional events, such as fear inducing crimes, tend to be conspicuous and thus attract attention. This improves the encoding of the salient event at the expense of peripheral information (Thorley et al., 2015; Drexler & Wolf, 2017). Events that are moderately emotionally arousing will be remembered well and improve later recall because they attract attention (De Quervain et al., 2009). The focussing effects of stress are one of the functions of the SAM axis activation which tend to be faster than HPA axis activation. However, high levels of emotional arousal may impair encoding and reduce memory of details, even for well-attended stimuli. (Drexler & Wolf, 2017). High levels of cortisol, promoted by HPA activation, are a likely mechanism behind this impairment. However, as both axes are triggered together, and interact with each other, it is not clear that the contrasting effects of stress are attributable to the different neural pathways (Godoy et al., 2018). Although there is some evidence of attentional benefits of SAM activation and negative effects on consolidation of memories resulting from high levels of cortisol resulting from HPA activation, the stress response is complex with many potential moderators across neurological, cognitive and behavioural responses (Godoy et al., 2018).

Experiences of stress or intense emotion also use cognitive resources which could otherwise be used on one's primary task. This intense form of arousal can impair memory by increasing cognitive load (Levine & Edelstein, 2009). In addition to this, emotion and stress

fix one's attention on the arousing stimulus and hence, should impair memory of neutral or peripheral information (Hoscheidt et al., 2014). Considering that crimes often involve a central, salient event as well as more neutral information such as contextual information about the crime scene, both neutral and emotional memories must be recalled by witnesses. Studies in the stress and eyewitness literature vary according to the stimuli used during encoding as well as the relationship between the stressor and the encoding task. Such methodological differences may further explain differences currently seen in the literature (Sauerland et al., 2016; Marr et al., 2020).

An example of this is the weapon focus effect, where the presence of a weapon draws the victim's attention, impairing their ability to remember other details or recognize the perpetrator's face (Steblay, 1992). This effect is thought to be in part because the threat posed by the weapon increases its salience, making it the central focal point for the participant. While studies have found a similar effect when an unusual item replaces the weapon, the effect is smaller (Hope & Wright, 2007). This suggests that the threat posed by a weapon, rather than only its novelty, leads people to focus on the weapon. Some studies of stress and eyewitness memory have incorporated weapons to better represent violent crimes, including them in both the low and high stress groups so as not to confound the experiment (Cutler et al., 1987; Lindberg et al., 2001). This may increase the stress experienced during the experiment for both groups. Of interest in comparing studies is whether the presence of a weapon affects high and low stress groups differently, either narrowing the attentional focus of stressed participants further, or reducing this difference by adding a narrowing effect to participants experiencing lower levels of stress. Additionally, as an extra stimulus is present, the task of remembering may be more difficult, reducing performance (Hanoch & Vitouch, 2004). The role of attention is also relevant to the stressor used in the experiment, which may pull focus away from the encoding task. As stress induction is not always related to the

learning material in eyewitness experiments, the effect of stress on event salience must also be considered (Sauerland et al., 2016).

Encoding Stimuli and Task Relevance of Induced Stress

The stimulus at recognition after a real crime can be a photograph, video or live line-up, but the encoding stimulus is always a live event (Fitzgerald et al., 2018). Yet, this is not always the case in experimental procedures, which often use photographic or video materials at encoding. While such procedures allow for convenience and control, they do so at the cost of ecological validity (Sauerland et al., 2016). Recent stress and eyewitness studies have attempted to remedy this by using a live 'mock crime'. Yet, these studies found no effect of stress at encoding on witness memory despite a significant difference in both self-report and physiological indicators of stress (Gering & Tredoux, 2018; Sauerland et al., 2016; Krix et al., 2015). This may be because these studies used an unrelated lab stressor administered shortly before the live event. The effect of the unrelated stressor on attention and encoding, may not accurately represent that experienced in a real crime (Christianson, 1992). To achieve a greater degree of ecological validity in studies of stress and eyewitness memory both the live event and the stressor should represent the character of the experience by including key elements of a real crime. While it is thought that this is relevant and will affect the outcome, it should be noted that this has not been empirically shown.

Many stress studies in neuropsychology use laboratory stressors that are unrelated to the memory task despite real world stressors typically being task related (Hanoch & Vitouch, 2004). This is certainly the case for eyewitnesses, as the perpetrator of the crime is usually both the stressor (posing danger to the witness) and the target one must encode (Christianson, 1992). Thus, it may not just be the neurobiological effects of stress on memory that affect encoding during a stressful encounter for a witness; the relevance of the stressor may have an

effect (Sandi, 2013). A meta-analysis by Shields et al. (2017) found that when the stressor and the task were related, memory was less likely to be impaired. Eyewitness studies using the Maastricht Acute Stress Test (MAST) have induced acute stress prior to the encoding event (Sauerland et al., 2016; Krix et al., 2015). While the duration of the MAST and the inclusion of physical, cognitive and socio-evaluative stress result in a strong HPA axis response (Smeets et al., 2012), the MAST unpairs the stress induction from the memory task (Gering & Tredoux, 2018). This threatens the ecological validity of eyewitness procedures using the MAST or similar stressors, as the encoding event in a crime invokes the stress response. In contrast these lab-based stressors are the source of the stress response which happens separately from the mock crime. The usual procedure is to first induce the stress and then present the encoding situation (Smeets et al., 2012). While this ensures that participants are experiencing the physiological stress response during the encoding task, they are not occurring together or related (Gering & Tredoux, 2018). As the SAM axis stress response has adapted to focus the agent's attention on the threat, the severance of stress and encoding task poses a potential problem (Hanoch & Vitouch, 2004).

A recent study attempted to solve this problem by inducing stress with the Cold Pressor Test (CPT), where stress is induced physical using cold water during a screen-based face encoding task (Davis et al., 2019). While in this case the stressor is occurring at the same time as encoding, solving one of the problems, the tasks remain unrelated. The cold water poses a physical threat which elicits a stress response, but this response is not coming from the memory task and so the events are not conceptually related. This weakens any association resulting from the simultaneous timing and may result in competition for cognitive resources (Troyer & Craik, 2000). The immersion of one's hand in ice water during the CPT causes participants discomfort and draws their attention away from the primary encoding task. This likely reduces encoding efficiency and subsequent recall for the stress group (Troyer & Craik,

2000). Ideally, the stressor should draw the participants' attention to the eyewitness event as would be the case in a crime as the SAM axis response should orientate participants to the threat, which in a crime is normally the perpetrator. A field study by Morgan et al. (2004) successfully linked the encoding event and the stressor using a high stress interrogation where the interrogator needed to be identified at a later stage. However, such studies give up aspects of control found in experiments and cannot be easily replicated. This control is important not only for testing the effectiveness of the stress manipulation but also for controlling estimator variables, such as length of exposure to the stressor and target, known to affect face encoding.

A number of studies in the literature have induced stress incidentally through the experiment while maintaining various degrees of experimental control (Maass, & Köhnken, 1989; Valentine & Mesout, 2008; Johnson et al., 2019). The trade-off between closely representing a real-world scenario and maintaining control over relevant variables in an experiment is a fine one. It is not yet clear which elements of a real crime are the most important to preserve and this has resulted in the use of a range of methods within the field of stress and eyewitness memory (Christiansen, 1992; Deffenbacher et al., 2004). There is a similar trade-off in the choice of method for presenting the encoding event. Many experiments have used photographic or video materials as encoding stimuli, whereas others have used live, often unexpected, events (Deffenbacher et al., 2004). Computer screens are typically used to present photographs and occupy the whole field of vision. In contrast live events are embedded in more complex settings and thus attended to differently. It is therefore important to consider how attention may affect the encoding of different stimuli. The threat to experimental validity when using photographic materials centres on participants' knowledge that they must attend to the screen (Sauerland et al., 2016). In contrast, a live event may

provide a more holistic and higher-dimensional view of the face, whereas a screen or photograph provides only a two-dimensional representation of a face. On the other hand, a live event is harder to control and contains more variables which affect the stress reaction. Participants experiencing live events may not always have the same stress reaction or focus on the same aspect of the crime.

Differences in Stress Response

External factors such as length of exposure to a stressor and its intensity can affect the stress response (Levine & Edelstein, 2009). As studies vary these factors, methodological differences between studies may explain inconsistencies in the literature in the effects of stress on performance (Sauerland et al., 2016). Different studies have used different methods of stress induction (e.g., the CPT, the MAST and mock interrogations among others). The differences between the methods used may also be responsible for some of the lack of consistency of results (Smeets et al., 2012; Sauerland et al., 2016). Besides these external differences in stressor, some internal factors can affect the stress response. These factors, including individual differences in baseline levels of arousal or sensitivity to stimulation, may also affect the impact of stress on memory. This includes affective states such as depression, which presents with increased levels of baseline cortisol (Kirschbaum et al., 1996). As some people are more resilient to increasing task demands, which include both task difficulty and stress intensity, the point at which stress impairs memory may vary between people (Plieger et al., 2016). As all these factors can affect where one is on the Yerkes-Dodson stress curve, with increased intensity and exposure pushing one further to the right on the x axis and individual differences changing the baseline or the rate of increase in arousal when exposed to a stressor, it is important to control for as many of them as possible (Sandi, 2013). While potential individual differences in the stress response will remain a challenge in this

literature, it is possible to control for many external factors. In doing so it should become clear why findings in the literature have not been consistent (Deffenbacher et al., 2004; Morgan et al., 2004; Sauerland et al., 2016).

A 2004 meta-analysis looking at high levels of stress during the encoding of a crime found that stress reduces witness memory of details (Deffenbacher et al., 2004). The authors suggested that witnessing a crime produces intense rather than moderate stress and their findings support the notion that poor encoding due to high levels of arousal impairs the ability of witnesses to recall events. A field study by Morgan et al. (2004) found that a high intensity stressor increased the false alarm rate and reduced the successful identification rate. In contrast, laboratory research has not shown a consistent effect of stress on witness identification (Sauerland et al., 2016). This may be because Morgan et al. (2004) used military recruits as participants, they were able to induce greater stress than a typical lab study and it may thus be that their finding of dramatic impairment results from such high levels of stress. If this level of stress induced is markedly different from those of lab studies, their participants, particularly in the stress group, may fall far further to the right on the Yerkes-Dodson curve. Rather than the groups being on opposite sides of the curves crest, it is likely that both groups were to the right of the crest with the control group near the peak and the experimental group far down the slope. Although they categorise participants as stressed or not stressed, as do most studies, both groups may have been more stressed than the average participant in a lab. This could make the extreme result and difference in findings from other studies, attributable to the Yerkes-Dodson law as both groups may fall further to the right on the x axis than participants in other studies. This could result in catastrophic failure by the more stress group; resulting in extreme differences between the groups (Yerkes & Dodson, 1908; Nixon, 1982). Furthermore, the procedure they used, which included survival skills training, a wilderness evasion exercise and finally an intensive interrogation, not only

induced high degrees of acute stress but likely resulted in participants experiencing chronic stress, which impairs memory (Morgan et al., 2004). The difference in results between studies by Morgan et al. (2004) and Sauerland et al. (2016) may be better explained by differences between chronic and acute stress rather than intensity of the stressor (Sandi, 2013). However, as there are many methodological differences between these studies and others in the literature, it is difficult to make direct comparisons (Deffenbacher et al., 2004; Sauerland et al., 2016).

Given that witnesses to crimes must recall the event after an unknown delay subsequent to its occurrence, it becomes necessary to consider how acute stress affects memory over time. A meta-analysis from the neuropsychology literature found that stress impaired recall on words and pictures unless the delay between encoding and recall was very short (Shields et al., 2017). As witnesses must typically wait several days before recalling the particulars of the crime event, a very short delay is unlikely. Thus, the meta-analysis by Shields et al. (2017), would suggest an impairing effect of stress in support of a similar conclusion reached in the meta-analysis by Deffenbacher et al. (2004). Findings in an empirical study in the eyewitness literature by Fitzgerald et al. (2012) also support this conclusion. Using children as participants, Fitzgerald et al. (2012) found no significant difference between high and low stress groups on a face recognition task after a shorter delay period of a month. On the other hand, they did find some differences after a longer delay of a year (Fitzgerald et al., 2012). After a month there were no differences between the groups, yet at a year's follow up there was a significantly better performance by the low stress group on target absent (TA) line-up decisions. A limitation in the current stress and eyewitness literature is that only a few studies have included a delay of more than a few hours, making the previously mentioned study an exception (Sauerland et al., 2016). Similarly, few of the studies in the literature on delay effects have included TA line-ups (Fitzgerald et al., 2012).

Furthermore, the 'short delay' reported in the study by Fitzgerald et al. (2012) is greater than the 'long delay' reported in the meta-analysis by Shields et al. (2017) where even the long delays were days and not weeks. Therefore, the question of the moderating effect of delay on stress-performance relations cannot be answered definitively. The ambiguity in both descriptions of time delay and differences in this delay length have produced an unclear quantification of the effect of post encoding time delay on memory differences between stressful and non-stressful encoding.

Differences in the effect of stress might depend on the type of task, with studies finding that stress affects subsequent recall and recognition tasks differently (Het et al., 2005). Het et al. (2005) found stress to impair recall but not recognition, in contrast to Deffenbacher et al. (2004) who found stress to impair both recall and recognition. As witnesses to a crime are typically required to both recall events (relying on explicit memory systems) and recognise perpetrators (which may be an explicit or an implicit memory task depending on several factors discussed below), this distinction is important (Dew & Cabeza, 2011; Vakil et al., 2018). The literature suggests that explicit memory is more likely to be impaired by HPA axis activation than implicit memory, impairing recall but not recognition (Drexler & Wolf, 2017). One important piece of evidence is that humans with hippocampal damage are still able to recognise unfamiliar faces, i.e. faces that have been seen before but only shortly or not often, as in a crime (Bird, 2018). Some studies in the neuropsychology literature have found high stress to improve implicit memory tasks while impairing those requiring explicit memory (Het et al., 2005; Sandi, 2013). While face recognition is often considered an implicit memory task, as are most recognition tasks (Dew & Cabeza, 2011; Sandi, 2013), line-up identifications may rely on explicit memory processes (Yonelinas, 1998; Wixted, 2007). Recognition is currently best described by a dual process model which uses a faster familiarity check, which is implicit, and a slower recollection channel which is

explicit (Wixted, 2007). This model in not a case of either or, but rather suggests that various degrees of familiarity can occur which support the more explicit recollections (Dew & Cabeza, 2011). Of course, in the case of high familiarity (a feeling rather than the result of multiple exposures) when viewing a line-up, an additional process of recollection may not be necessary as this familiarity facilitates an automatic high confidence decision. Even so, an initial sense of familiarity when seeing a face may feed into a more explicit comparison of features against an existing memory or other present faces (Wixted, 2007).

This can be seen in simultaneous line-ups, in which multiple faces are presented to the witness and the witness is asked to choose the perpetrator among them, or state that the subject is not present (Clark et al., 2015). While participants at times report a face to 'pop-out' from the line-up, used to describe an effortless identification decision based on familiarity, in other instances participants make featural comparisons between faces, an explicit form of memory work (McQuiston-Surrett et al., 2006). This shows that in cases where participants feel a sense of familiarity, witnesses can rely on implicit processes to make a near automatic decision. Yet, having the option of comparing faces when they are simultaneously presented allows for the option of a diagnostic feature comparison (Clark et al., 2015). As such, the identification process need not be implicit nor explicit but can be a combination of both (Wixted, 2007).

Because of these task specific properties, findings in previous analyses that did not use line-up identification tasks may not have tested the same cognitive processes used by witnesses in line-up tasks. Evidence from neuropsychology supports this as face recognition ability is spared after brain damage that impacts other memory processes (Bird, 2018). Similarly, it is possible that stress may negatively impact some brain areas, but that this may have differing effects on face recognition compared to other types of memory. Witnessing a crime is a complex situation which involves different processes to memorising word lists or

series of faces and so differences in how witnesses attend to stimuli may affect later memory (Kanwisher and Yovel, 2006). It is worth considering that different encoding or retrieval materials might produce different results (Fitzgerald et al., 2018). One type of stimulus might be better in general, i.e. is unaffected by stress, if this effect applies equally across stress conditions. This potential outcome would suggest that differences in encoding stimuli may not be important when comparing studies. However, the answer to this question is unclear and best answered by a meta-analysis which can consider these differences as moderator variables.

Line-up administration

Differences in method at recognition and between encoding and recognition should also be considered as potential moderator variables. As with encoding stimuli, it is worth considering whether differences in line-up presentation might affect high and low stress groups differently. While the medium of line-up used might be relevant, so might the type of line-up presentation. Past studies on eyewitness recognition have used both simultaneous line-ups where the suspect and foils are presented together, e.g., in a 2x3 photographic array, as well as sequential line-ups where participants view only one potential suspect or foil at a time while the experimenter cycles through the line-up options (McQuiston-Surrett et al., 2006). While sequential line-ups were previously favoured as producing fewer foil identifications (Steblay et al., 2001), it is not clear if this is always the case (McQuiston-Surrett et al., 2006). Analysis using Receiver Operating Characteristics (ROC), rather than ratio-based measures, have shown no clear difference between sequential and simultaneous line-ups (Grolund et al., 2014). Furthermore, some studies have shown a trade-off, with fewer decisions being made overall when sequential line-ups are used, showing an increase in

conservative judgement making (Gronlund et al., 2015). This finding suggests that more correct decisions are made with simultaneous line-ups and that signal detection measures of discriminability, the proportion of target to foil identifications, are higher for simultaneous line-ups (Wixted & Mickes, 2014).

Supporting this, alternate ROC analysis has shown that simultaneous and not sequential line-ups produce better discriminability (Grolund et al., 2014, Seale-Carlisle et al., 2019). While different analyses have shown different results, resulting in uncertainty over which presentation method, if either, is superior, there remain differences in strategies available to participants using each line-up (Clark et al., 2015). Simultaneous line-ups allow participants to directly compare faces in a relative judgement. This allows for an explicit comparison of remembered diagnostic features, which the participant can use to either eliminate foils or identify the target as the best match relative to the other faces (Clark et al., 2015). Not being able to do this in a sequential line-up might account for differences between line-up presentation methods and explain a finding showing sequential line-ups perform better when the suspect is not present, as participants are less likely to pick someone out when faces are presented sequentially (McQuiston-Surrett et al., 2006). In contrast, using a sequential line-up forces greater reliance on absolute judgement, where the presented face can only be matched to memory (Wixted & Mickes, 2014). Not being able to make a judgement based on a face being the closest of the group and rather having to judge each face only on its closeness to memory, has been thought to reduce false alarms without affecting successful choices (Wixted & Mickes, 2014). Because of these differences in line-up presentation method, it is worth considering whether high or low stress encoding conditions produce different effects for different line-ups. This could be due to choosing rates, as some studies have shown that stressed participants are more likely to make an identification, despite being no more, or even less, accurate (Sauerland et al., 2016; Valentine & Mesout,

2008).

The delay between encoding and line-up tasks may also be relevant. That memory quality deteriorates over time has been shown in both basic and applied research starting as far back as 1880 with Ebbinghaus' forgetting curve, which has been replicated and shown in other contexts (Atkinson & Shiffrin, 1968; Hardt et al., 2013; Murre & Dros, 2015). However, the added influence that stress at the time of encoding has on decay over time is less clear. A study by Fitzgerald et al. (2012), found that after a delay of a month there was no difference in recognition performance between anxious and non-anxious children on either target present (TP) or absent line-ups. In contrast, at a follow up a year later, the group that was more anxious during encoding performed significantly better on TA line-ups and showed a slight, albeit non-significant negative difference on TP line-ups (Fitzgerald et al., 2012). This is one of very few studies to use such a long delay, as participant attrition often prevents this. This study was able to retain 75% of its participants at follow up which, while admirable, is still a large enough loss to substantially reduce the statistical power of the study. Although few studies use such long delays or test participants at multiple times, the delay used in eyewitness studies typically varies, from virtually no delay between encoding and recognition, to a delay of weeks or months (Maass, & Köhnken, 1989; Fitzgerald et al., 2012; Sauerland et al., 2016). As such, the interaction of such a delay with stressful encoding conditions has not been thoroughly studied and might be better understood through a meta-analysis.

Outcome Variables

Another variable at the recognition phase worth considering is whether the line-up includes the perpetrator or not. The current recommendation in the literature is to include both a TP and TA line-up in one's experimental design (Wells et al., 2020). This can be

useful in comparing participant choosing rates and strategies. While a successful decision in a TP line-up is choosing the perpetrator from among the foils; in a TA line-up the correct decision is to reject the line-up as not having the perpetrator in it. Both TA and TP of line-ups permit foil identifications as well as 'don't know' decisions from uncertain participants. This allows for some comparison and as such some studies have collapsed results across these types of line-ups (e.g. Steblay et al., 2001). In contrast, the different line-ups and outcome variables allow for a more nuanced analysis as some variables may affect TA line-ups differently from TP ones (McQuiston-Surrett et al., 2006). As one of the critiques of the Yerkes-Dodson law is that the relationship holds for hits but not for incorrect outcomes, line-up responses allow for this to be empirically tested (Hanoch & Vitouch, 2004). Correct responses differ between TP and TA line-ups, with TP line-ups requiring an identification (while avoiding foils) and TA line-ups requiring a rejection. Having different stimuli at the recognition stage with differing correct responses provides a broader set of outcomes against which one can test theory. Nonetheless, it should be noted that while the current trend is to include both types of line-ups and to record hits, false alarms, line-ups rejections and *don't know* responses, studies differ in method and reporting (e.g. Maass, & Köhnken, 1989; Sauerland et al., 2016). As such, not all the studies in the literature can help to unpack these questions.

As the previous meta-analysis on stress and eye-witness memory only considered hits and false alarms separately, some combination of the two responses, using a Signal Detection Theory (SDT) measure of sensitivity such as d' may also provide results of interest. The use of SDT has recently become more popular in the eyewitness literature for its use in determining how different identification procedures affect participants' decision making (Wixted & Mickes, 2014). Using SDT one can view hits and false alarms together to establish both willingness to make a decision and ability to discriminate between targets and

lures (known innocents in a line-up) (Wixted & Mickes, 2014). As such a high ratio of hits to false alarms even if absolute hits are low; shows good discriminability. It must be noted that this measure of empirical discriminability is a noisy indicator of decision strategy and as participants typically only make one decision, it is not often used (McQuiston-Surrett et al., 2006). However, d' has been used in several experiments on line-ups by Meissner et al. (2005) to show that signal detection theory can be beneficial to eyewitness experiments using mock crime paradigms as well using a repeated measures design. More recently, the literature using ROC analysis has incorporated these measures, along with confidence ratings to better understand eyewitness decision making (Wixted & Mickes, 2014). While this measure was not used in the previous meta-analysis of stress and eyewitness memory, one can consider each study in a meta-analysis as a single effect of a repeat measure. In doing so, and calculating a difference in sensitivity between groups, we can see whether participants in stress conditions choose differently from low stress participants in line-up tasks. This is particularly relevant when considering the process and strategies used by simulated witnesses in these tasks which relies on both recollection and familiarity (Wixted, 2007). Previous studies have shown that the shape of a ROC curve depends on which processes are used, with a more linear curve showing reduced reliance on implicit familiarity and a curvilinear ROC resulting from explicit recollection strategies (Wixted & Mickes, 2014). As stress affects implicit and explicit memory differently (Sandi, 2013), SDT methods may help ascertain if stressed witnesses rely more heavily on implicit processes. Despite being imperfect measures, as the scope of this thesis is limited to line-ups this additional measure will be considered so as to provide additional depth of analysis.

In summary, recall and recognition performance may be affected by neurobiological and psychological states that influence attention and processing at encoding. Acute stress likely influences these processes by orientating one towards the stressor through SAM axis

activation, prioritising a response to the stress. While this initial SAM axis response facilitates encoding processes, cortisol response from the HPA axis activation impairs memory consolidation (Roozendaal et al., 2009). As such, stress will affect encoding both through attentional and memory processes. This effect may be beneficial, if the stress induced is mild or moderate but may also impair overall memory according to both the Yerkes-Dodson law and the cue-utilisation theory (Hanoch & Vitouch, 2004). Methodological differences in stress induction between studies in eyewitness studies and neuropsychology might account for the range of reported effects between these fields (Shields et al., 2017). Within the field of eyewitness research, differences in methodology such as crime event and delay between encoding and recall may have produced inconsistencies in this literature (Sauerland et al., 2016; Marr et al., 2020). Moreover, differences in methods which can be recorded as moderator variables may produce interactions with stress, allowing for more nuanced consideration of the stress effect. The effects and interactions of stress may also affect TP and TA outcomes differently and may have differing effects depending on the specific outcome variables measured. These differences are best considered in the context of a systematic review and meta-analysis where many potential moderator effects can be identified and quantified.

Rationale and Research Aims

As witness testimony has far-reaching consequences, knowledge that can help establish its accuracy is important. This is relevant both for supporting reliable and discrediting unreliable evidence provided by witnesses. As the number of studies has grown and methods evolved since the last meta-analysis on stress and eyewitness memory; it is important to review the existing studies. This study aimed to systematically review the literature on stress during encoding of eyewitness events and to conduct a meta-analysis to

quantify the effect of stress as well as several moderator variables on subsequent line-up decisions. If there are variables that have significant moderating effects on the stress-performance relationship, they could be a potential source of unmeasured confounds in the literature. The following hypotheses will be tested:

Hypothesis 1: On average across studies, participants experiencing stress during encoding tasks will perform worse (scoring fewer hits and more false alarms) than those in low stress, control conditions on-line up recognition tasks, as found in the previous meta-analysis by Deffenbacher et al., 2004.

Hypothesis 2: This effect will also be seen in the signal detection measure with d' being lower for stressed participants.

Hypothesis 3: Differences in the direction and magnitude of this effect between studies will be attributable to methodological differences between studies.

Hypothesis 4: Viewing stress as a continuous variable with a non-linear relationship to performance will produce a different result than using stress as a dichotomous predictor or performance.

Method

Design and Procedure

Ethics for this study was granted by the University of Cape Town (UCT), Psychology Department internal ethics committee (see appendix A). A systematic review of the literature and meta-analysis of existing data will be used to assess the effect of stress at encoding on subsequent line-up performance. The protocol for this systematic review was guided by the

outline in the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA-P) statement (Moher et al., 2015; Shamseer et al., 2015; see appendix B).

Study Sample

Data Sources and Data Search Strategy. Articles for this review were found using the Wiley online library and Web of Science using the Boolean phrases in Table 1. Google Scholar's 'cited by' function was also used to search articles citing the previous meta-analysis. The articles used in the previous meta-analysis were also considered for inclusion. Theses that were not peer reviewed, but published by universities online and found during the search process were considered for inclusion. All combinations from each category in Table 1were used as search terms for each online database. The search phrases were selected by two reviewers (The author and a PhD student in the lab working on a cognate topic) in accordance with the aim of the study. Additionally, the reference list of the previous mete-analysis by Deffenbacher et al., (2004), as well as the list of articles which cite this paper, were used as a source of articles. Studies only included in book chapters were not included in this analysis due to unavailability. As the country has been in various levels of lockdown, the UCT library has been closed and access to books further hampered. Moreover, it is no longer common to only publish such studies in books, with the number of journals greatly increasing.

Table 1

Boolean phrases for literature search

Category 1: Stress	Category 2: eyewitness	Category 3: memory	Category 4: line-up

stress	eyewitness	encoding	line-up
acute stress	crime	memory	parade
Induced stress	witness	recognition	identity parade
arousal			lineup
			identification parade
			line up

Note. All phrases, plurals and variants of the base word within each category (column) were searched using "LOR" and those between categories (across rows) were searched using "AND".

A PhD student as well as the author of this review, both conducted the searches for an initial screening of articles. All articles were recorded using Mendeley referencing software to keep the database and to exclude any duplications. This reference list was uploaded onto Rayyan, which allows reviewers to make notes about inclusion or exclusion of articles (Ouzzani et al., 2016). Using this software, each reviewer was able to review each article and decide to include or exclude it, whilst giving reason for exclusion. From the initial list of search phrases, line-up and line up were found to be equivalent terms. Parade, identification parade and identity parade are equivalent. "Line up" produced significantly more hits than other phrases. Of these equivalent words, only "Line up" was used when conducting a second search of terms flagged as key words in articles.

The reviewers met via video link software once the set of articles had been reduced so that any differences in inclusion could be discussed and settled. Once the set of articles were agreed upon, a coding sheet was developed by the author and supervisor to capture all the relevant variables of interest. This coding sheet was created in Microsoft Excel and used by the author and the independent coder to record the data extracted from the set of articles. An

Intraclass Correlation Coefficient (ICC) for continuous variables was calculated, using the *irr* package in R to quantify the level of agreement between the coders for each outcome variable. Two of the variables initially had low correlations (0.65 and 0.75) which on closer examination was due to a decimal point error. Re-analysis showed ICC > 0.99 for all outcome variables. The small differences were for values taken off of graphs for the Brown (2003) and Read et al. (1992) studies which differed by no more than 0.02.

Eligibility Criteria. Articles were assessed according to their eligibility, ensuring that only high-quality studies were included. Eligible studies had to include an encoding event, a stress induction at or prior to encoding with some subsequent manipulation check and a well administered line-up recognition test. It was decided that a variety of encoding events would be allowed; including live and video events as this is a potential moderator variable of interest. Similarly, stress induction events that were allowed for inclusion included both incidental stress induced through the encoding task (e.g., walking through the London Dungeon; Valentine & Mesout, 2008), as well as stress induced through previously validated tasks such as the MAST or Trier Social Stress Test (TSST; Smeets et al., 2012; Kirschbaum et al., 1993). As there is debate in the literature over the effectiveness of various stressors both within and outside of eyewitness studies, this was considered as a variable of interest in potentially explaining the differences found in the field (Sauerland et al., 2016, Smeets et al., 2012).

However, these studies did need to be manipulating acute stress, as this response differs from chronic stress and is the most likely type of stress response experienced during crimes. As such, the study by Morgan et al. (2004) was not included in the final analysis as, while the details are not fully reported, their participants underwent an extended period of several days during which they were exposed to physical and psychological stressors prior to a mock prisoner of war event. During the 48 hours prior to the encoding event, participants

were also deprived of regular sleep and food. This extended period of stress prior to the encoding event made to rattle experienced soldiers, makes this study too different from the rest of the sample. Although not retained in the final analysis because of these differences in method, data from this study was initially analysed along with the rest of the studies. First including and then removing the study allows for a sensitivity analysis to show that excluding this study does not significantly affect the results.

Two studies were included where there was no control group but where stress groups were created post-hoc based on a stress response to the paradigm (Valentine & Mesout, 2008; Fitzgerald et al., 2012). In these studies, participants all experienced the same event and then were divided using a median split of their stress response. The clear reporting of manipulation checks in these studies showed differences in stress experienced between their post-hoc grouping. As individual differences in stress response occur and likely affect outcomes, this sort of split is methodologically valid; although a median split may not be as useful as dividing participants into terciles or quartiles which would allow analysis of a non-linear stress-performance relationship. Furthermore, these groups were not confounded by other characteristics which divided them.

In contrast a study by Buckhout et al. (1974) created their groups based on line-up performance and then showed differences in stress. By doing this, rather than dividing the groups by stress and then showing differences in line-up performance, a perfect difference between groups results. This is because the grouping by line-up performance on the TP line-up creates a successful hit group and a foil identification group with no hits. Average stress was shown for groups after this split, showing the perfect hit group to have a significantly lower stress score. The results of taking the average stress into account after the fact being that, despite variations in stress within the groups, there is no variation in outcome as all 'high stress' participants produced only false alarms while all 'low stressed' participants only

hit (Buckhout et al., 1974). This is at odds with all the other studies in the literature and produces a logical problem of looking at the effect of stress on decision as a causal relationship, as decision is established before the stress grouping. While a difference in stress was found between the post-hoc groups, it is confounded by another distinct difference between the groups, namely performance on the outcome measures for the line-up task. In this event, comparing these groups as stress groups creates the appearance that a small difference in stress caused a stark difference in performance. This study was ultimately excluded from the final analysis but results including this study are shown as well.

Study two by Goodman et al. (1991), was also excluded as it contained only a stress group whose data was reused in study three when compared to a control group. As such, including study three captured the effect of this group in an experimental comparison (Goodman et al., 1991). These last two cases are mentioned specifically as they were included in the previous meta-analysis by Deffenbacher et al. (2004). Other studies in which stress was not experimentally manipulated were also excluded, such as a study by Clifford and Hollin (1981) where violence was used as a proxy for stress in the video materials but no manipulation check regarding stress, arousal or anxiety was conducted.

Four studies included a second person or bystander who was included in a separate line-up. This was not considered a confound as the experimenters distinguished these people from the primary target, by specifying their role. The data for the additional person was not coded or analysed as this study was more concerned with the perpetrator or target and not with questions regarding peripheral or central details. One of these studies by Read et al. (1992) also included an alcohol intoxication condition alongside a placebo condition. Arousal was manipulated within each of these groups. As such, the control group was further divided into a high and low arousal group whose data was reported separately and used in this analysis.

The type of recognition task used in studies was also considered as a criterion for inclusion or exclusion, considering only line-up tasks as suitable for inclusion, as opposed to facial recognition tasks such as face matching or old/new face recognition paradigms. This was decided on for several reasons. Firstly, these tasks differ significantly in process as line-ups require only one or two decisions (when both TP and TA line-ups are used) while other recognition tasks often require participants to make many decisions (often as many as 80), rendering the experience quite different to that likely after experiencing a crime. (Deffenbacher, 2004). Secondly, the encoding tasks used for these recognition paradigms are different. Where for a line-up, a participant typically views a single salient event in some context, other recognition tasks show a series of static faces, often from the neck or jaw/chin up only. In the eyewitness task, the visual stimuli are complex events which include critical information along with neutral information. In contrast, the stimuli for face matching tasks show only the critical information needed at recognition. These two differences in encoding and recognition task were deemed enough to consider these bodies of literature separately, with face matching being basic research and line-up tasks being applied research. Finally, the meta-analysis by Deffenbacher (2004) concluded that a moderate difference existed between these types of paradigms. As there is a difference both in encoding task and in outcome it was decided that face recognition tasks not using line-ups could be excluded from this analysis. These studies were not collected and coded for empirical comparison as it was not practical given the time constraints of this thesis.

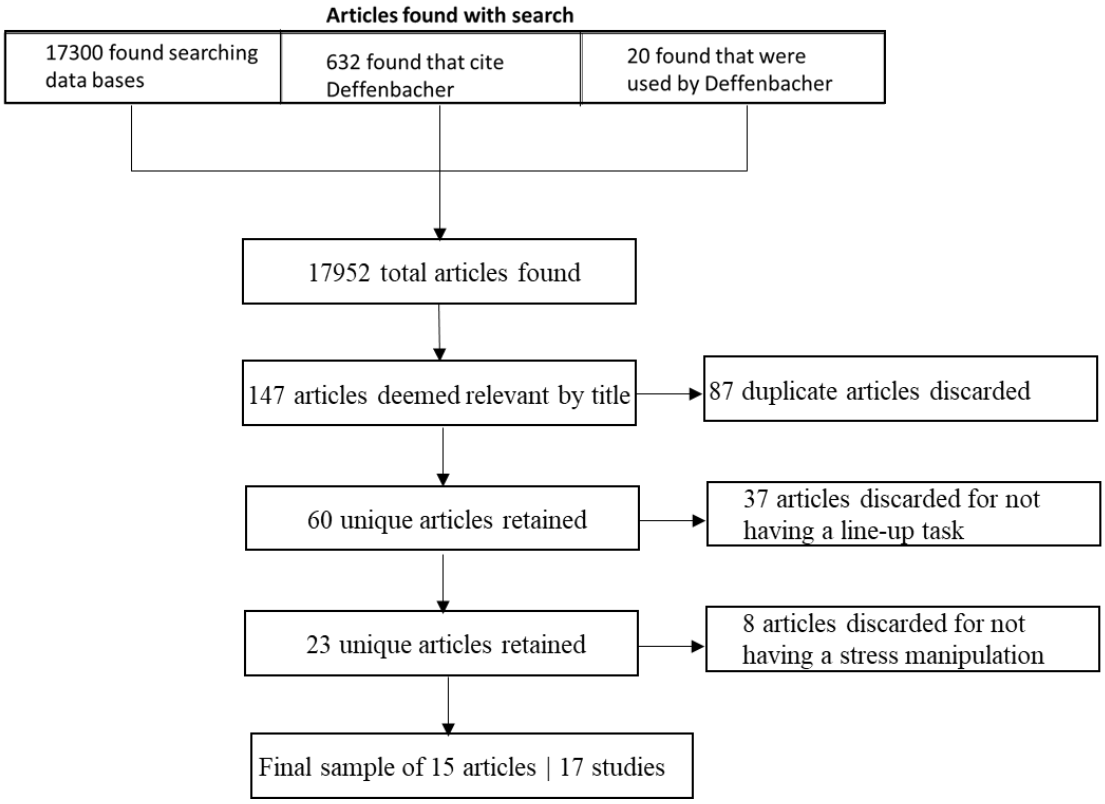
Other variables deemed eligible and included in the coding sheet were the use of simultaneous or sequential line-up presentation; the presence or absence of a weapon; the delay between encoding and retrieval; the participants' age as well as the medium of presentation of the line-up (live, video or photo). As these variables differ between studies and are likely to affect stress and recognition memory, they were considered as potential

moderators in the meta-analysis. While some of these variables may have a general effect (such as the presence of a weapon potentially reducing recognition accuracy), this was not considered as a confound if the variable was present for both the experimental and control group. Similarly, reporting on method such as line-up construction, counterbalancing, instruction and presentation were not factors considered in the eligibility criteria where they were the same between the groups of interest.

Final Sample. 15 articles containing 17 studies and reporting 32 effects of induced stress on line-up decisions were included in this review. All of these studies are high quality eyewitness studies with a clear stress manipulation. While they differ in other aspects, all included studies have an encoding task, an element of stress induction, a subsequent manipulation check for the induced stress, and a well administered line-up task. Figure 3 shows how the original pool of articles was narrowed down to the final set of studies included in the review and analysis. This sample included studies conducted between 1987 and 2019, with a total of 2045 participants across effects ($M_{number\ of\ participants} = 63.91$, $SD = 44.67$). The data from these studies was captured in an Excel spread sheet (Appendix C). A second spread sheet with data of some additional, ultimately excluded studies can be found in Appendix D.

Figure 3

Articles search and inclusion



Data Analysis

The primary analysis for this study was a comparison of low and high stress (or control and stress groups) for each study on a range of recognition variables discussed below. The effect size used in the meta-analysis was log odds ratio (LOR) on the recognition variables between the stress groups. A simple effect analysis was used for the main effects while a mixed effects analysis was used when including moderator variables (Viechtbauer, 2010). TP and TA line-ups were analysed separately as well as by collapsing correct responses and false alarms across TA and TP line-ups. A SDT sensitivity statistic, d' prime, was also calculated for which the effect size used in the meta-analysis was mean difference (MD) in d' between groups. Two of the studies included had 3 levels of stress. Each of the three possible comparisons was coded as a separate effect. Likewise, studies that included

additional levels (e.g., interview type before the recognition task or different time delays) and which had the stress and control groups coded separately for each of these levels were coded at the lowest level. As such, some studies were coded on multiple lines, when multiple stress vs control effects could be coded.

The potential outcome variables for TP line-ups included were successful hits, foil identifications (false alarms), line-up rejections and 'don't know' responses. While many agree that one ought to record all these outcomes from line-up recognition studies (Wells et al., 2020), not all the studies included in the analysis set used each of these options, with some authors coding responses only as 'successful hits' or 'other'. As such the number of articles used in the various analyses differed. Similarly, the TA line-ups where responses can be coded as correct rejections, foil identifications (false alarms) or 'don't know' responses were sometimes coded only as 'correct rejection' or 'other'. While the current recommendations include participants not to make a decision (Wells et al., 2020), this was not the case in all of the studies included.

Unlike the meta-analysis by Deffenbacher et al. (2004), a combined measure of recognition accuracy across TA and TP line-ups, as well as a specially constructed signal detection theory measure, d', is used in the meta-analysis reported here. The previous review and meta-analysis of this literature did include both TP and TA line-ups, these were only analysed separately (Deffenbacher, 2004). Collapsed responses across TP and TA line-ups were also calculated, in studies where both types of line-up were used. Hits for TP line-ups and rejection decisions for TA line-ups are both correct responses and so some previous studies have collapsed these (Hosch & Bothwell, 1990; Steblay et al., 2001). Similarly, false alarms in both TP and TA line-ups can be collapsed as incorrect responses, as a false accusation is the most damaging potential outcome leading to false imprisonment (Wells et al., 2020). D' was calculated for each of the studies, both for TP and TA line-ups, by

calculating standardised scored for correct responses and false alarms then subtracting the false alarms from the hits i.e. d' = Z(H) – Z(FA). This SDT measure shows the overlap between correct identifications and false alarms, showing not just the ability to make a good identification but rather the ability to discriminate well between targets and lures (Wixted & Mickes, 2014). While this measure is not enough to establish decision making strategy, it does provide more insight into the memory strength of the participants by taking into account the difference between hits and false alarms. As this study uses secondary data, this must be done across a whole studies sample. To calculate the variance for this measure using secondary data, the Gourevitch and Galanter method was used (Suero et al., 2017).[1]

The main effect size calculated for the various outcomes was the odds ratio and then LOR (Viechtbauer, 2010). Rather than calculate *Z* scores directly from proportions, the *Metafor* package allows counts and totals for each effect to be entered as a 2x2 Table (Viechtbauer, 2010). Effect sizes variances are determined by the function *escalc* and are used in the model building to calculate weighted *Z* scores. As the *Metafor* package allows for a moderator analysis within a meta-analysis, several moderator variables were included in the coding sheet (Viechtbauer, 2010). These included: the method of stress induction; the encoding stimuli used; the type of line-up; the medium of the line-up, the delay between encoding and recognition, the participants' age as well as the presence or absence of a weapon. Table 2 (at the end of the section) shows the relevant moderator variables for the studies. These variables represent the major differences in method between the studies in the field which could explain the difference found between studies. All data was coded in a Microsoft Excel spreadsheet and read into R version 3.6.1 for analysis. The primary package used for the analysis was *Metafor* (Viechtbauer, 2010). All R code used for the analysis can

[1] This method provides an estimation of the variance as the actual variance cannot be known.

be found in Appendix E.

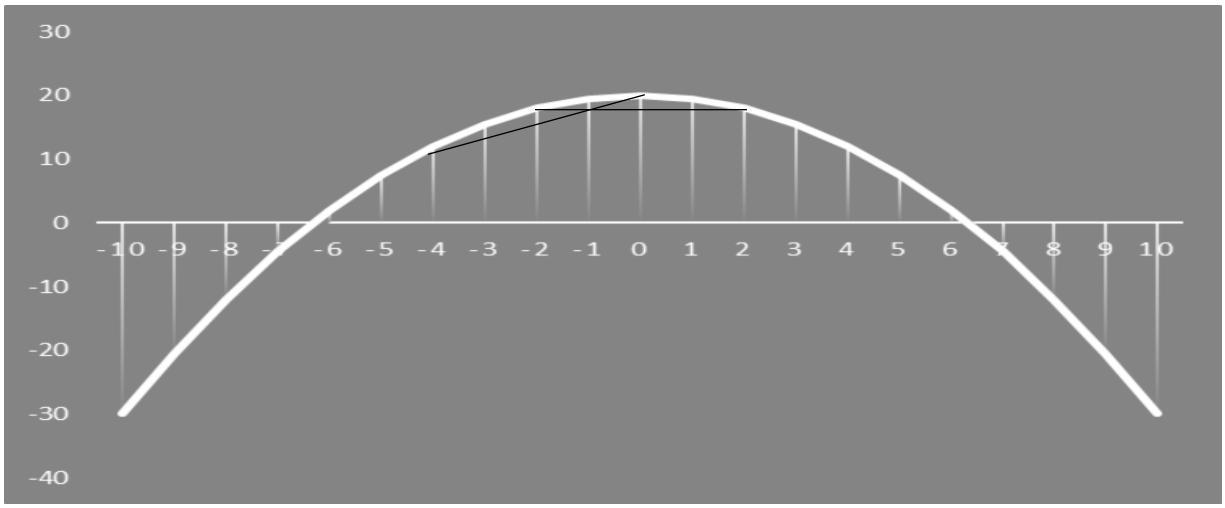
An analysis of the manipulation checks was also attempted but as manipulation checks were not consistent across studies, including various self-report scales as well as physiological measures, it was not possible to conduct an analysis with existing data. As such, a coding sheet and scale were developed in an attempt to assign stress values to each condition (see Appendix F). The data was captured in an Excel spread sheet to be used for analysis (see Appendix G). Rather than code for difference in stress between each group per effect, an individual score was assigned to each condition. This is necessary as on a non-linear curve, as the Yerkes-Dodson curve theorises the stress curve to be, a given difference on the x axis can result in a varying y axis effects depending on the position of the independent variable on the x axis. This is demonstrated in Figure 4, where one can see that a difference of one unit on the x axis produces a large positive change in y on the left of the curve, only a small change in the centre and a large negative change on the right. As one cannot assume that all stress conditions across studies fall on the same point on the x-axis, or that control groups are equally un-arousing, a given difference between groups need not correlate with outcome effects. As such, each condition within each study should be considered as its own stress effect, with control groups always being lower than experimental groups.

This scale was validated by the same PhD student, referred to on p. 30, so as to increase reliability of the scale using the coding guide in Appendix F. An initial Kappa for ordinal data was calculated value of 0.86, indicating reasonable agreement between the coders. Discrepancies between the coders were resolved during a discussion after the coding. However, the author of this study acknowledges that such a scale, created post-hoc, does have limitations in validity. This scale and subsequent analysis using the scale were conducted to show that some other form of continuous measurement of stress might be

helpful in establishing the stress performance relationship as the current experimental manipulations used in the literature, dichotomising stress into two groups, are a low-resolution tool which may be missing a more complex relationship.

Figure 4

Negative parabola with grid lines showing change on the x -axis.



Note: This is one possible version of what a non-linear stress – performance relationship could look like. Black lines show the possible differences in slope for a given change on the x axis, in this case four units, based on the starting point

This scale was used in a multi-level linear model with the same outcome variables as the meta-analysis to see whether such a continuous measure of stress could be useful in predicting stress. The analysis was conducted twice, once assuming a linear relationship and again assuming a quadratic relationship. If the relationship were to be linear, a dichotomous measure would likely be sufficient as the gradient remains constant and a difference of one

unit at any point on the x-axis will have the same difference in the outcome measure. As mentioned above, if this relationship is not linear, this analysis would highlight the need for reform in the current practice of stress measurement.

Table 2

Studies listed by year showing moderator variables of importance.

Author	Year	Study	Stressor	Encoding Scenario	Participant Age	Line-up Presentation	TP/TA	Weapon	Manipulation Check
Cutler et al.	1987	1	incidental	video	adults	sequential	Collapsed across	Yes	Self-report
Maass and Kohnken	1989	1	incidental	live	adults	simultaneous	TA only	Yes	Self-report
Hosch & Bothwell	1990	1	incidental	live	adults	simultaneous	Collapsed across	No	Physiological
Hosch & Bothwell	1990	2	incidental	live	adults	simultaneous	Collapsed across	No	Physiological
Goodman et al.	1991	1	incidental	live	children	simultaneous	TP only	No	Self-report
Goodman et al.	1991	3	incidental	live	children	simultaneous	TP only	No	Self-report
Read et al.	1992	2	incidental	live	adults	simultaneous	TP and TA	No	Self-report
Lindberg et al	2001	1	vicarious	video	children	simultaneous	TP only	Yes	Self-report
Brown	2003	1	incidental	photo	adults	sequential	TP only	No	Self-report
Hulse & Memon	2006	1	incidental	video	adults	simultaneous	TA only	No	both
Valentine and Mesout	2008	1	incidental	live	adults	simultaneous	TP and TA	No	both
Houston et al.	2012	2	incidental	video	adults	simultaneous	TP and TA	No	Self-report
Fitzgerald et al.	2012	1	incidental	live	children	simultaneous	TP and TA	No	Other rater
Rush et al.	2014	1	TSST	live	children	simultaneous	TP and TA	No	Self-report
Rush et al.	2014	1	TSST	live	teenagers	simultaneous	TP and TA	No	Self-report
Sauerland et al.	2016	1	MAST	live	adults	simultaneous	TP and TA	No	Physiological
Gering & Tredoux	2018	1	MAST	live	adults	simultaneous	TP and TA	No	both
Johnson et al.	2019	1	incidental	live	adults	simultaneous	TP and TA	No	both

Note: The Rush et al. (2014) study is given two lines despite being one study as participant age was used as an experimental group and is displayed here as a moderator.

Results

Results for both the review of the literature and the subsequent meta-analysis will be presented below. The systematic review of the literature, being the foundation for the subsequent analysis, will be presented first. The results of the search and set of final articles will be presented, as will the characteristics of the studies that have been included in the statistical analysis. The statistical analysis will first look at the studies descriptively before examining the results of the meta-analysis. As the studies included investigated the effects of stress on line-up performance, the results of the stress manipulation and several types of line-up outcomes were analysed. As the stress induction varied so widely and the data reported was often limited, the stress manipulations were considered descriptively. Stress was included as a dichotomous variable for the meta-analyses, with effects calculated between groups. The results of a set of multilevel linear models (MLM) are also presented where stress was considered first as a dichotomous and then as a continuous variable using an inter-study rating system.

Descriptive Analysis of Studies

Manipulations checks. All 17 studies included performed some type of manipulation check to show differences in levels of stress were significant between the experimental and control groups, or between different levels of arousal. Four studies included both self-report and physiological measures of stress or arousal, three used only physiological measures and the remaining 10 used only a self-report or rater-report in the case of child participants. Of the seven studies that did include a physiological measure of stress, one used only cortisol levels measured in the participant's saliva, two measured only skin conductance, three measured heart rate and one measured both heart rate and skin conductance. In addition to these differences, studies took measures at different points. Only one study by Sauerland et

al. (2016) took four readings across the experiment, at baseline; pre-stress, post stress and post-cooldown, as is the practice in the neuropsychology literature on stress. Johnson et al. (2019) took three measures, baseline, pre-stress and post-stress. Two studies measured only pre- and post-stress so as to calculate a change score, with the remaining three studies only taking a post stress measure to compare across groups. Of the 14 studies that included a self-report or rater measure, only 10 reported what that measure was. Three of these studies used the State Anxiety Inventory, one used the negative items from the Positive and Negative Affect Schedule (PANAS) and the rest used a range of self-developed Likert type scales (Spielberger et al., 1983; Watson & Clark, 1999). Differences in the reporting of these measures occurred, with four studies only reporting the significant difference between the groups; one study reporting the overall mean and use of a median split to form groups; a couple of studies calculating change scores and the rest only taking post-stress measures.

Differences in method. Several differences in method were found across the studies, which could be coded for use as potential moderator variables, shown in Table 2. Three of the studies included a weapons in their experiments. Two of the studies used sequential line-up presentation with the rest presenting their line-ups members simultaneously. One study used a video line-up with the others all using photographic line-ups. 12 of the studies used adult participants, four used children below 13 years of age, with one of these including teenagers as a comparison group. Three of the studies induced stress using lab-based stressors previously validated in the neuropsychology literature, namely the MAST and the TSST, with the other 13 studies inducing stress incidentally through the eyewitness encoding event. These encoding events also differed with one study using a photographic slide show, four studies using video materials and the remaining 11 using a live encoding event. While all the studies using a lab-based stressor had a live encoding event, those using incidental encoding varied including photographic, video and live encoding events.

The final moderator considered in this study was the delay between the encoding event and the line- up task. Half of the studies used a delay of less than 24 hour with the mean delay for these being 13.75 minutes (SD = 15.73), with a median of 12.50 minutes. The high standard deviation here indicates a bimodal distribution, even among the studies with a short delay, as well as the small sample size. Of the remaining studies, the study by Cutler et al., (1987) collapsed results across several delays, but the mean result for the remaining seven studies was 50 days. In contrast, the median for this second half of studies was only a week, with the mean being heavily skewed by one study which had a follow up condition a year after the encoding task. No standard deviation is given here, and these measures should not be considered precise as there were differences within studies that were not reported in detail. Particularly for longer studies with follow up periods of weeks or months, a variation of hours between participants is virtually guaranteed. The variation within studies notwithstanding, this data shows how widely the delay between encoding and retrieval varies between studies in the stress and eyewitness literature.

Line-up outcomes measured. Of the final sample of 17 studies (recovered from the final sample of 15 articles), eight included both a TP and a TA line-up), four included only a TP line-up, the remaining two only a TA line-up and three provided results collapsed across TP and TA line-ups. Of the two studies that only made use of a TA line-up, Hulse and Memon (2006) stated that no suitable photograph of the target could be obtained, while Maass and Köhnken (1989) simply noted the lack of a TP line-up as a limitation. Reasons for not including a TA line-up were not given by studies which did not include them. As including a TA line-up requires either a repeat measures design or a much larger sample, it does create practical difficulties. Studies set outside of the lab (e.g., Valentine & Mesout, 2008) report difficulties recruiting participants and may have found it impractical to collect TA data. Similarly, studies looking at variables other than stress and as such needing more

participants to achieve the necessary statistical power (e.g., Brown, 2003), might also have decided that the comparison and rigour gained by a TA line-up required too much additional data. The study by Read et al. (1992) only reported data from the TP line-up, just noting a non-significant result for the TA line-up.

Outcomes also differed in terms of variables measured. While all studies with TP line-ups measured hits, two did not measure false alarms, three did not include line-up rejections and six did not include *don't know* responses. Similarly, disparities exist for TA line-ups with all the studies measuring false alarms and correct rejections but two studies not measuring *don't know* responses. For those studies which collapsed results across TP and TA line-ups, one only reported correct decisions while the other two only reported false alarms. This provides a severe limitation for individual studies, as a comparison of different responses provides richer detail and greater insight into how independent variables affected participant's decision making. Regarding this review, it resulted in some analyses having large samples and others having smaller, less statistically powerful samples.

Meta-analyses

Separate meta-analyses are reported for each of the four possible line-up decisions on TP line-ups as well as the three possible decisions on TA line-ups. Thereafter, two additional analyses are reported for collapsed correct answers across TP and TA, namely hits on TP and rejections on TA, as well as collapsed false alarms across TP and TA line-ups. A final set of analyses were conducted using d' for TP and TA line-ups separately. Tables reporting proportions of respondents making each type of line-up decision, separated by high and low stress groups, are reported in Table 3 for TP line-ups and Table 4 for TA line-ups (next section). The LOR was calculated as the measure of effect size, for all outcome variables

other than d', in order to compare accuracy of line-up decisions made by participants in the low and high stress groups. For d', the MD between groups was calculated.

Table 3

Studies by year showing the proportion of each line-up choice on TP line-ups for low and high stress groups.

Author	Hits		False Alarms		Rejections		Don't know		N	
	low	high	low	high	low	high	low	high	low	high
Goodman et al.	0.77	0.55	0.00	0.11	0.33	0.44			9	9
Goodman et al.	0.35	0.47	0.60	0.53	0.05	0.00			17	17
Read et al.	0.63	0.76							16	16
Lindberg et al.	0.21	0.32							43	43
Lindberg et al.	0.16	0.36							43	43
Brown	0.30	0.62	0.40	0.21					41	40
Brown	0.53	0.70	0.27	0.13					39	42
Brown	0.51	0.62	0.40	0.21					41	40
Brown	0.35	0.70	0.24	0.13					41	42
Brown	0.51	0.30	0.40	0.40					40	40
Brown	0.35	0.53	0.24	0.27					42	42
Valentine & Mesout	0.75	0.18	0.21	0.54	0.00	0.00	0.04	0.29	28	28
Fitzgerald et al.	0.65	0.67	0.30	0.33	0.00	0.00	0.04	0.00	23	15
Fitzgerald et al.	0.21	0.27	0.74	0.73	0.00	0.00	0.05	0.00	19	11
Houston et al.	0.40	0.27	0.25	0.47	0.35	0.25			57	59
Rush et al.	0.55	0.60	0.11	0.20	0.11	0.20	0.22	0.00	9	10
Rush et al.	0.50	0.55	0.00	0.36	0.40	0.09	0.10	0.00	10	11
Rush et al.	0.91	0.40	0.00	0.30	0.09	0.20	0.00	0.10	11	10
Rush et al.	0.54	0.64	0.18	0.00	0.18	0.27	0.09	0.09	11	11
Sauerland et al.	0.53	0.53	0.00	0.07	0.40	0.33	0.70	0.07	30	30
Gering & Tredoux	0.29	0.29	0.29	0.21	0.43	0.36	0.00	0.14	14	14
Johnson et al.	0.71	0.50	0.14	0.33	0.00	0.17	0.14	0.00	36	24
Johnson et al.	0.71	0.77	0.14	0.18	0.00	0.06	0.14	0.00	36	28
Johnson et al.	0.77	0.50	0.18	0.33	0.06	0.17	0.00	0.00	28	24

TP Line-ups

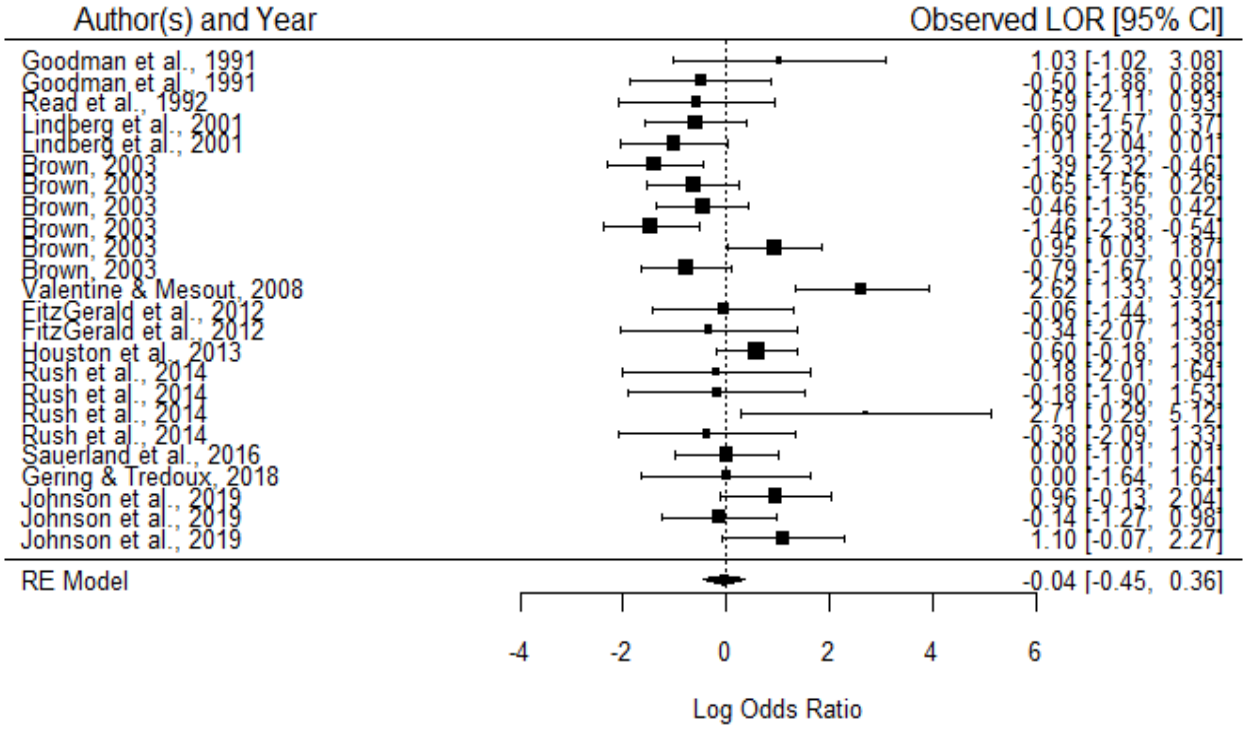
Correct identifications (hits). Figure 5 reports the effect sizes, LOR, for hits on TP line-ups per study. The analysis was based on $k = 24$ effects and $N = 1333$. The average LOR is -0.04 ($SE = 0.21$), 95% CI [-0.45, 0.36]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant ($Z = -0.21$, $p = .833$). This suggests that there was no difference in performance between low and high stress groups on hit rate for TP line-ups. Still, before one can say that there is no effect present, one must test for heterogeneity as high heterogeneity would suggest that there are different subgroups within the data. Effects for these groups can differ from the overall effect and may suggest moderator effects. The measures of heterogeneity were indeed high $I^2 = 65.91\%$, with the significance test being significant $Q(23) = 64.44$, $p < .001$. Looking at the Figure below, one can see that the effect sizes vary around 0 but that some studies did produce both positive and negative significant effects. After conducting an outlier analysis, only one study was overly influential (see Appendix H) for all outlier and influence plots). That study by Valentine and Mesout (2008) showing far worse performance by the high stress group. However, removing this study does not change the sign or significance of the main effect. The moderator analysis was run both including and excluding this study.

This analysis was repeated with the inclusion of the Buckhout et al. (1974) as well as the Morgen et al. (2004) study. These effects did not change the significance of the effect despite both showing better performance by the low stress group, with the effect estimate changing to 0.26 ($SE = 0.23$), 95% CI [-0.18, 0.71] and ($Z = 1.17$, $p = .243$). Appendix I shows the forest plot with these studies included, showing the Buckhout et al. (1974) study as having a large effect size but very wide CI and the Morgan studies having a modest but significant negative effects of stress on hit rate. This analysis was done to show that these

effects were considered and that extremity of effect was not the reason for exclusion. The methods of these studies, one for reasons of data analysis and the other procedure, were the basis for their exclusion. As these studies differ so substantially from the other studies in this review, their inclusion confounds the over-arching homogeneity of the sample.

Figure 5

Forest plot showing effect of high vs low stress on TP line-up hits.



Note: positive LOR indicated more hits by the low stress group.

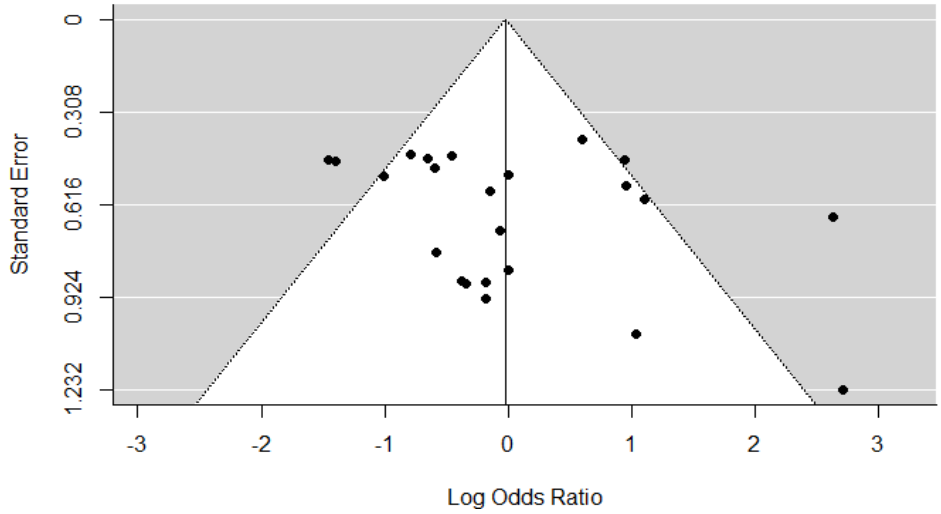
A funnel plot (Figure 6) was also analysed to check for any publication bias of studies. Most of the studies fall in the pyramid funnel shape, with two studies falling either side of the funnel. It appears that the biggest effects falling outside of the funnel report negative effects of stress on correct TP identifications, but that more reported effects show

either no effect or a small positive effect. A second funnel was created without the previously identified outlier, but this did not have a notable effect, with only that point (the higher outlier on the right) changing.

As, this analysis indicated heterogeneity between the studies, further analysis was conducted to see if moderators could account for the differences in effect seen across studies (Viechtbauer, 2010). It is worth noting that developer of the package *metafor* states that these measures are not always reliable, particularly when the sample of articles is small (Viechtbauer, 2010). As the review of the literature showed that there were many differences between the methods used in the current set of studies, such heterogeneity was expected. As such, although the effects of these analyses are reported below, a separate mixed effect analysis using the moderator variables coded was also conducted.

Figure 6

Funnel plot of LORs for hits on TP line-ups.



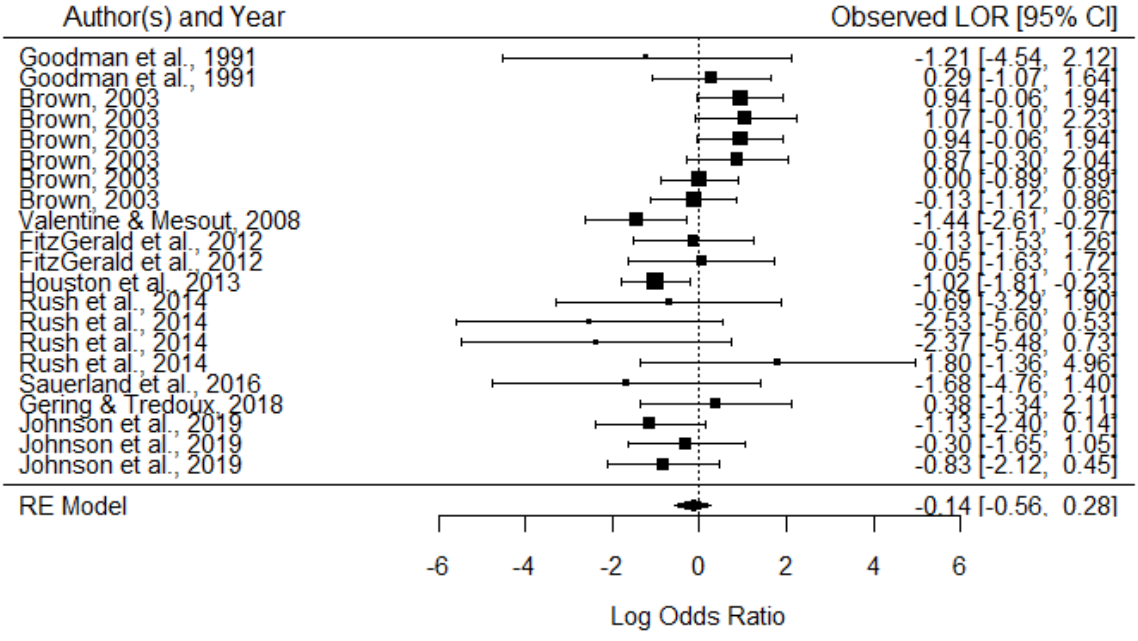
Looking at the mixed effects model with the moderators for hits on TP line-ups, a significant result was observed for the moderation effect for type of line-up presentation, being either simultaneous or sequential when all of the studies were included. With influential study the Mesout and Valentine (2008) outlier study removed, the effect was no longer significant but approaching significance when comparing successful hits on TP line-ups between low and high stress groups LOR = 1.08 (SE = 0.56), 95% CI [-0.02, 2.18], (Z = 1.91, p= .057. This effect suggests that high stress participants were more likely than low stress participants to make correct identifications on TP line-ups, when line-ups were presented sequentially. It should be noted that there were only six effects included with sequential line-ups that had separate TP line-ups, all from one study by Brown (2003) No difference in performance was observed for simultaneous line-ups. The test for residual heterogeneity was still high with I^2 = 46.75% and the significance test being significant $Q(20)$ = 37.71, p = .010 but reduced from the previous value. The other moderator variables, namely: participant age, delay between encoding and retrieval, the type of stressor used and whether the encoding was live, video or photo, were not significant. The code and results for all effects can be found in appendix E.

False alarms (foil identifications). The results for false alarms on TP line-ups were similar to those for hits. As hits and false alarms both require an identification, it is unsurprising that there is a strong correlation between the two, in this data set the correlation was r = -0.67. As such we expect a similar pattern to the hits but with an effect of the opposite sign. Unfortunately, not all studies that reported hits also reported false alarms. Figure 7 reports the effect sizes, LOR, for false alarms on TP line-ups per study. The analysis was based on k = 21 effects and N= 1129. The average LOR is -0.14 (SE = 0.21), 95% CI [-0.56, 0.28]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant ($Z = -0.65$, p = .511). This suggests that

there was no difference in performance between low and high stress groups on false alarm rate for TP line-ups. However, one must test for heterogeneity, as high heterogeneity would suggest that there are different subgroups within the data, before concluding that there is no effect. The measures of heterogeneity were not as high, I^2 = 45.79% as that for hits, but was significant $Q(20) = 37.16$, $p = .011$, indicating that a moderator analysis might be beneficial. Looking at the Figure 7, one can see that the effect sizes vary around 0 and that only one study, by Houston et al. (2013) had a CI that did not cross 0. After conducting an outlier analysis, no outliers were found. A funnel plot (Figure 8) was also checked with no studies falling far outside of the pyramid shape. As hits are normally the focus in eyewitness research, it is interesting that the false alarm values are less extreme. This suggests that there is more publication pressure on the hit data than the false alarm data.

Figure 7

Forest plot showing effect of high vs low stress on TP line-up false alarms.

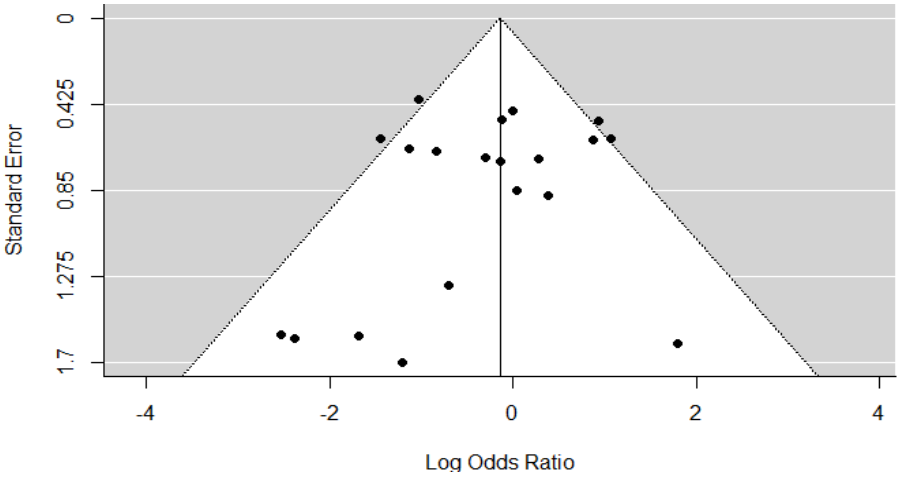


Note: positive LOR indicated more false alarms by the low stress group.

Looking at the mixed effects model with the moderators for false alarms on TP line-ups, a significant result for line-up presentation method was observed. The effect, LOR = -1.25 (SE = 0.30), 95% CI [-1.84, -0.66], (Z = -4.18, p< .001), was similar in magnitude and opposite in sign from that of the same variable on TP hits. This suggests that when line-ups are presented sequentially, high stress participants make fewer identifications of foils than low stress participants while there is no difference between groups on simultaneously presented line-ups. Again, there were only six effects included with sequential line-ups that had separate TP line-ups, all from one study by Brown (2003) and so must be interpreted cautiously as with small uneven sample sizes there is the potential of confounds. The test for residual heterogeneity was much lower with I^2 = 0.00% and the significance test no longer significant $Q(18) = 16.95$, $p = 0.551$. The other moderator variables, mentioned above, were not significant. This analysis was repeated with the Buckhout et al. (1974) study included. There was no change to the sign or significance of either the main effect or moderator effect. The complete analysis and output can be replicated using the data file in Appendix D and the R code in Appendix E.

Figure 8

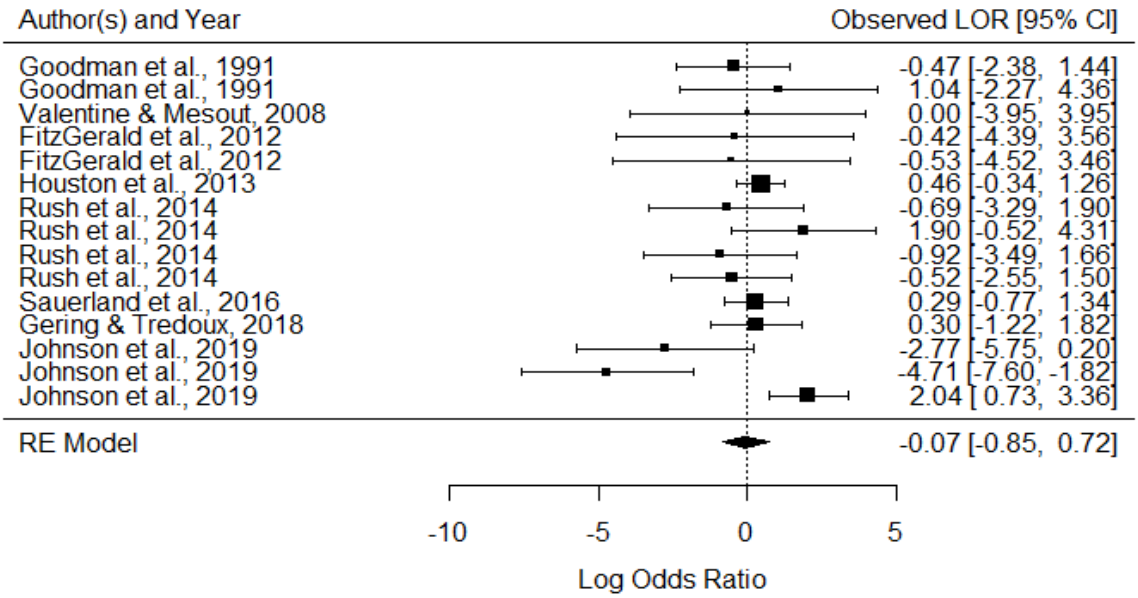
Funnel plot of LORs for False Alarms on TP line-ups



Line-up rejections. Figure 9 reports the effect sizes, LOR, for line-up rejections on TP line-ups per study. The analysis was based on k = 15 effects and N= 639. This is a notable decrease from the sample used in the previous analyses. As seen in Table 3, fewer studies reported or gave participants the option to reject the line-up if they believed the target was not present. For line-up rejections, the average LOR is -0.06 (SE = 0.40), 95% CI [-0.84, 0.72]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant ($Z = -0.31$, p = .757). This suggests that there was no difference in performance between low and high stress groups on line-up rejection rate for TP line-ups. The measures of heterogeneity were high, I^2 = 66.23% and significant $Q(14)$ = 27.43, p = .017, again suggesting that a moderator analysis would be beneficial. A funnel plot (Figure 10) was also checked with no studies falling outside of the pyramid shape.

Figure 9

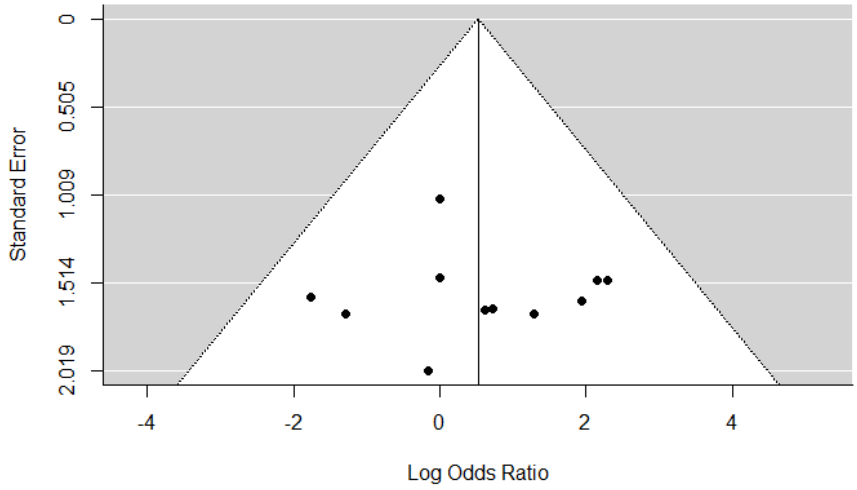
Forest plot showing effect of high vs low stress on TP line-up rejections.



Note: positive LOR indicates more rejections by the low stress group.

Figure 10

Funnel plot of LORs for line-up rejections on TP line-ups



An outlier analysis indicated that the second effect from the Johnson et al., (2019) study was an outlier. Indeed, looking at Figure 7 below, one can see that there is a particularly large effect *LOR* = -4.71, 95% CI [-7.60, -1.82] which is the only significant negative effect whose CI does not cross 0. Removing this effect, the average LOR changes to 0.26 (*SE* = 0.31), 95% CI [-0.35, 0.87] but remains non-significant ($Z = 0.83$, $p = .404$).

Separate moderator analyses were conducted, one including and the other excluding the outlier. Neither produced a result for any moderator variable that was approaching significance and residual heterogeneity remained high in both cases. As there were so few data points for TP line-up rejections, it is unlikely that the heterogeneity results are meaningful (Vichtbauer, 2010). The paucity of data in this analysis also makes meaningful conclusions on rejections of TP line-ups difficult to draw. This analysis was also repeated with the effects from the study by Morgen et al. (2004). This did not produce any significant effects for either the main analysis or subsequent moderator analysis and can be found in appendix I.

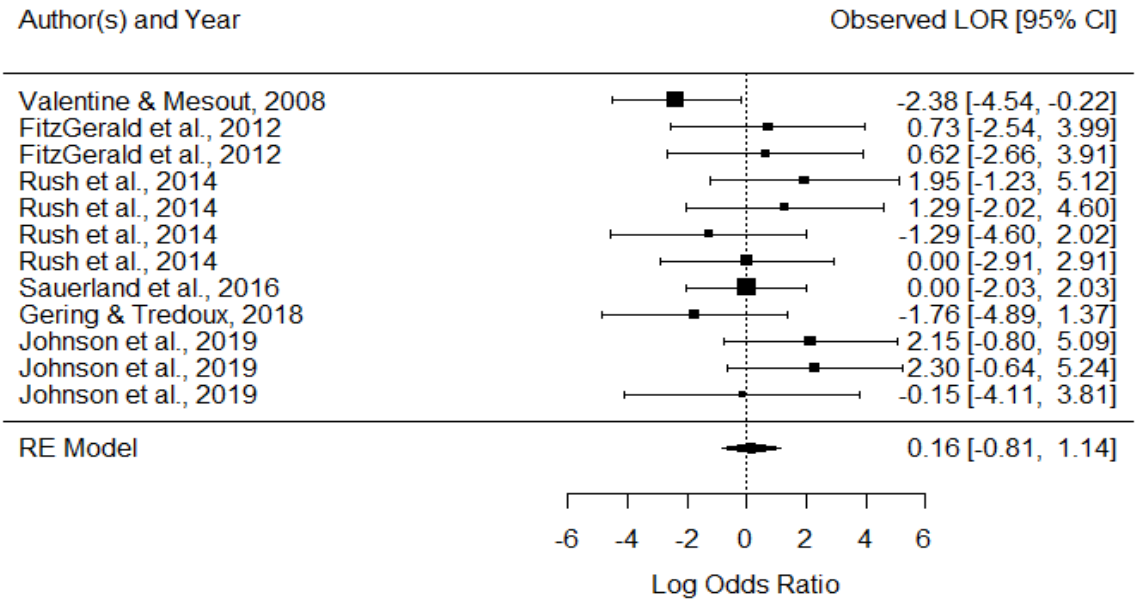
Don't know responses. Figure 11 reports the effect sizes, LOR, for don't know responses on TP line-ups per study. The analysis was based on k = 12 effects and N= 471. This is the smallest data set for TP line-ups as few studies allowed for this response option. For *don't know* responses, the average LOR is 0.16 (*SE* = 0.50), 95% CI [-0.81, 1.14]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant (Z = 0.33, p = .743). This suggests that there was no difference in performance between low and high stress groups on *don't know* response rate for TP line-ups. An outlier test showed one study, by Valentine and Mesout (2008), that was an outlier. Removing this study increased the magnitude of the effect size *LOR* = 0.52 (*SE* = 0.46), 95% CI [-0.37, 1.43] but it remained non-significantly different from 0 (Z = 1.14, p = .255). A funnel plot (Figure 12) was also analysed which showed all the studies, other than the outlier, falling within the funnel, with half falling on either side of the middle line.

The measures of heterogeneity were low, I^2 = 24.26% and non-significant $Q(11)$ = 13.17, p = .282 suggesting that a moderator analysis would not be beneficial. However, as this heterogeneity test is less accurate with smaller samples and the review indicated

generally high levels of heterogeneity between studies, a moderator analysis was conducted. The results showed that the variable recording whether studies used only a TP, or both a TP and TA line-up was significant *LOR* = -0.95 (*SE* = 0.33), 95% CI [-1.60, -0.30] and significantly different from 0 (*Z* = -2.84, *p* = .005). Yet, as only one of these studies, the outlier study by Valentine and Mesout (2008), recording *don't know* responses used only a TP line-up, this result should be interpreted cautiously. With the outlier removed, the effect is no longer significant and so indicates it is likely an anomaly.

Figure 11

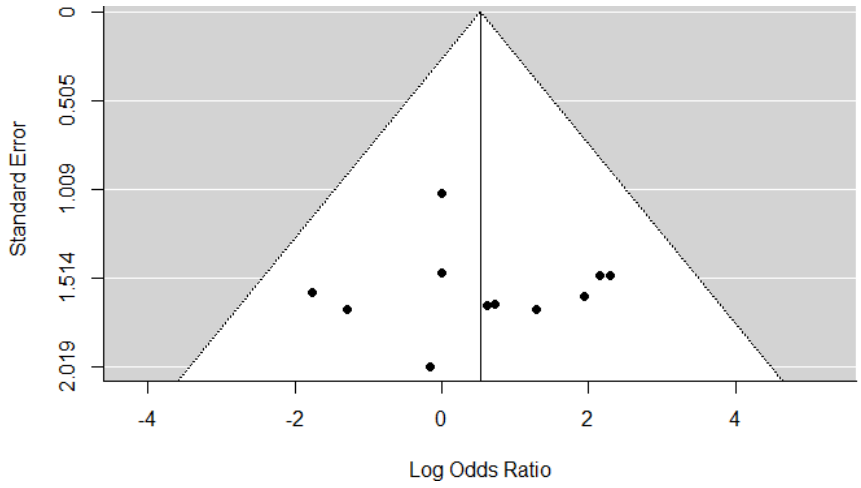
Forest plot showing effect of high vs low stress on TP line-up 'don't know' responses.



Note: positive LOR indicates more 'don't know' responses by the low stress group.

Figure 12

Funnel plot for don't know responses on TP line-ups



TA Line-ups

Correct Rejections. There are fewer studies in the literature that report separate TA line-up data compared to those that report TP line-up data. Figure 13 shows the effect sizes, LOR, for correct rejections of TA line-ups per study. The analysis was based on $k = 15$ effects and $N = 688$. The average LOR is 0.12 ($SE = 0.35$), 95% CI [-0.57, 0.81]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant ($Z = 0.34$, $p = .731$). This suggests that there is no difference between high and low stress groups on correct rejections of TA line-ups. Looking at the tests of heterogeneity one can see substantial heterogeneity $I^2 = 77.79\%$, with significant heterogeneity between studies $Q(14) = 46.68$, $p < .001$ suggesting that a moderator analysis may show some effects. However, no significant moderator effects or subsequent changes in heterogeneity were found. An outlier analysis indicated that there were no influential cases. Looking at a funnel plot for this data (Figure 14), one can see an

equal distribution of effects across the mid-line, with four studies within the pyramid falling either side of the line and two effects outside of the pyramid on either side.

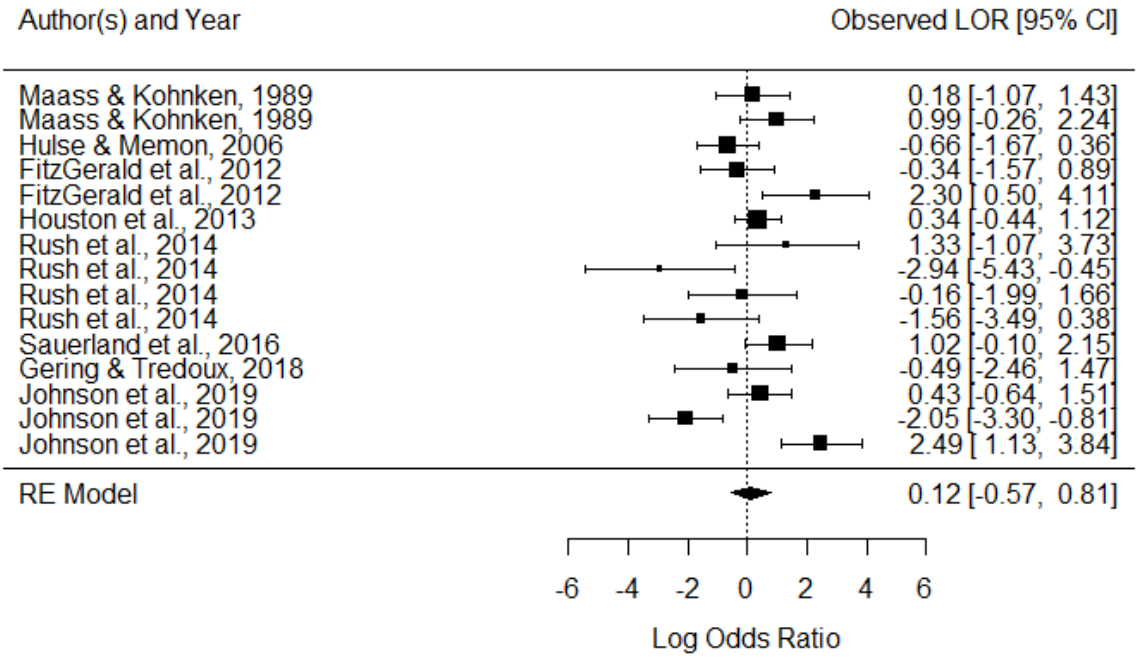
Table 4

Studies by year showing the proportion of each line-up choice on TA line-ups for low and high stress groups.

Author	Rejection		False Alarm		Don't know		N	
	low	high	low	high	low	high	low	high
Maass & Kohnken	0.36	0.32	0.64	0.68			22	22
Maass & Kohnken	0.67	0.43	0.33	0.57			21	21
Hulse & Memon	0.61	0.74	0.39	0.26			35	35
FitzGerald et al.	0.50	0.58	0.44	0.38	0.06	0.04	18	24
FitzGerald et al.	0.86	0.38	0.13	0.56	0.00	0.06	14	16
Houston et al.	0.36	0.29	0.64	0.71	0.00	0.00	58	59
Rush et al	0.36	0.13	0.36	0.63	0.27	0.25	11	8
Rush et al	0.11	0.70	0.44	0.10	0.44	0.20	9	10
Rush et al	0.56	0.60	0.44	0.10	0.00	0.30	9	10
Rush et al	0.33	0.70	0.67	0.30	0.00	0.00	9	10
Sauerland et al.	0.80	0.59	0.13	0.25	0.07	0.17	31	32
Gering & Tredoux	0.14	0.21	0.64	0.71	0.21	0.07	14	14
Johnson et al.	0.44	0.33	0.44	0.67	0.13	0.00	36	24
Johnson et al.	0.44	0.86	0.44	0.14	0.13	0.00	36	28
Johnson et al.	0.86	0.33	0.14	0.67	0.00	0.00	28	24

Figure 13

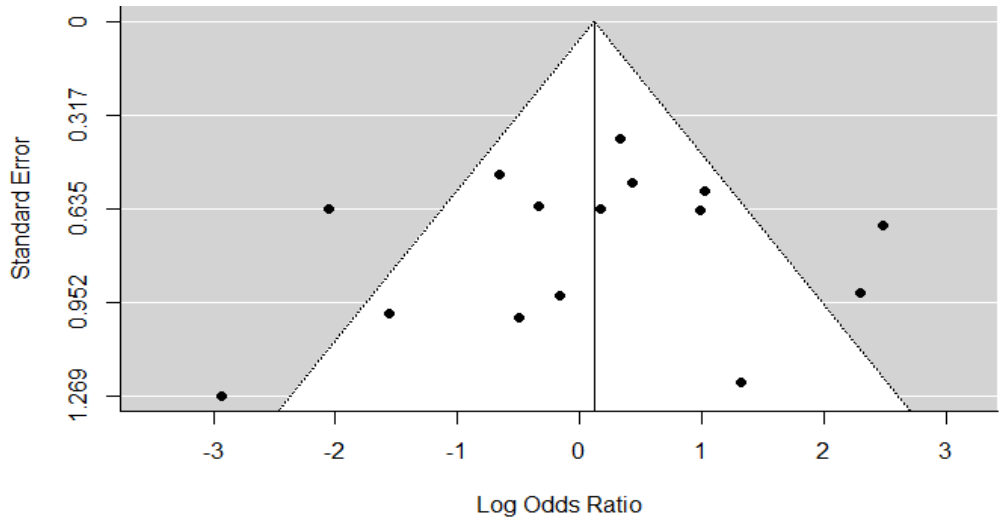
Forest plot showing effect of high vs low stress on correct rejections for TA line-ups.



Note: positive LOR indicates more correct rejections by the low stress group.

Figure 14

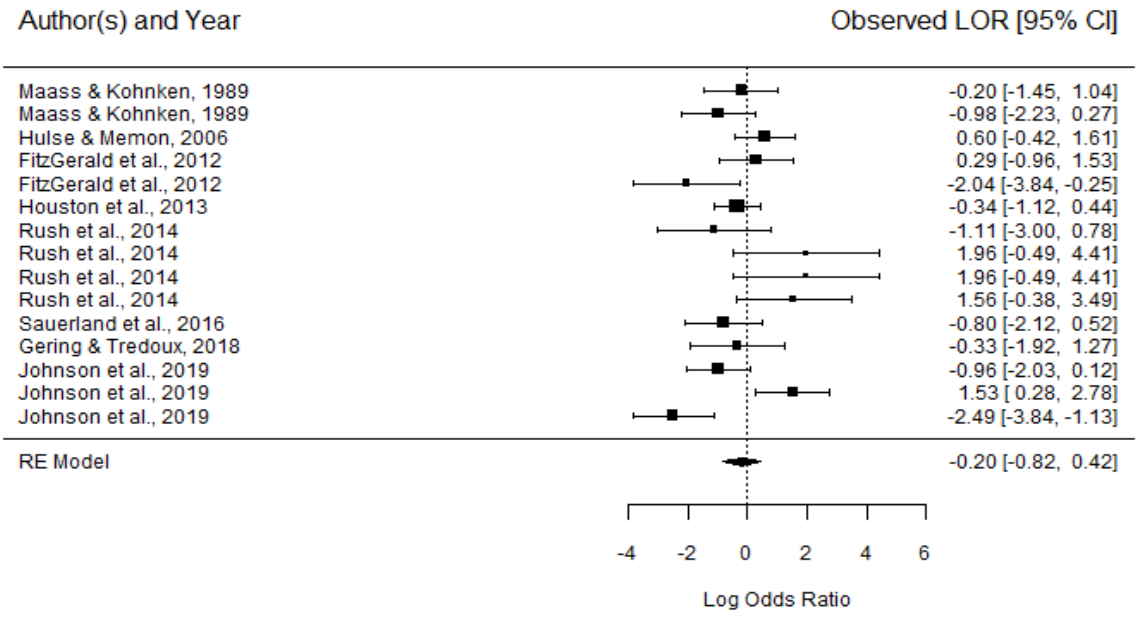
Funnel plot showing LORs of correct rejections on TA line-ups.



False alarms. Figure 15 shows the effect sizes, LOR, for false alarms of TA line-ups per study. The analysis was based on k = 15 effects and N= 688. The average LOR is -0.20 (*SE* = 0.32), 95% CI [-0.82, 0.42]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant (Z = -0.63, p = .528). This suggests that there is no difference between high and low stress groups on false alarm rates for TA line-ups. Looking at the tests of heterogeneity one can see substantial heterogeneity I^2 = 67.54%, with significant heterogeneity between studies $Q(14)$ = 39.55, p < .001 suggesting that a moderator analysis may show some effects. However, the moderator analysis showed no significant effects or reduction in heterogeneity. Looking at the diagnostic tests, there were no outliers. A funnel plot (Figure 16) shows an even distribution of effects with three points falling outside the pyramid. The even distribution suggests no publication bias.

Figure 15

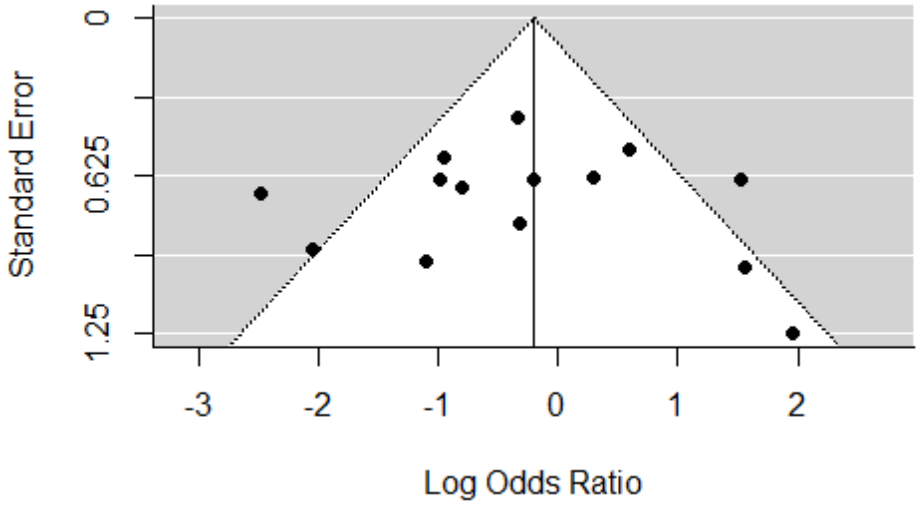
Forest plot showing effect of high vs low stress for false alarms on TA line-ups.



Note: positive LOR indicates more false alarms by the low stress group.

Figure 16

Funnel plot showing LORs of correct rejections on TA line-ups.



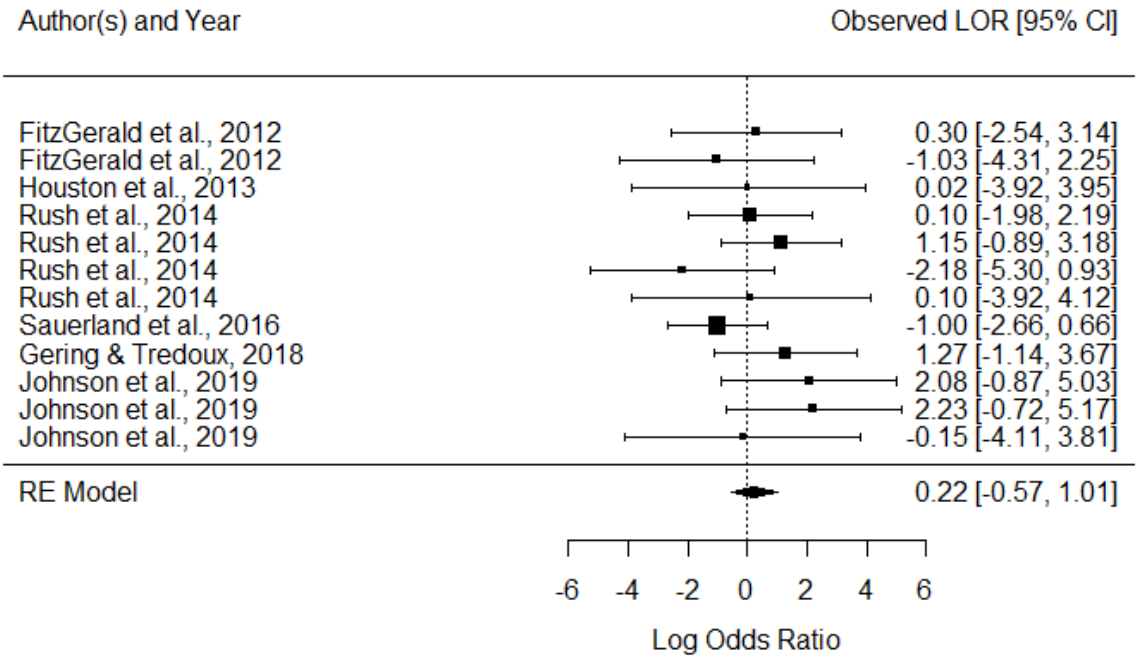
.

Don't know responses. Figure 17 shows the effect sizes, LOR, for don't know responses on TA line-ups per study. This analysis was based on k = 12 effects and N= 532 as not all of the studies with TA line-ups reported *don't know* responses. The average LOR is 0.22 (SE = 0.40), 95% CI [-0.05, 1.01]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant (Z = 0.56, p = .578). This suggests that there is no difference between high and low stress groups on don't know responses for TA line-ups. Looking at the tests of heterogeneity one can see little heterogeneity I^2 = 6.06%, that was not significant $Q(11)$ = 9.81, p = .547 suggesting that a moderator analysis would be unlikely to show effects. The diagnostic tests of influence did indicate that the study by Sauerland et al. (2016) was an influential case. As this study did not have an extreme effect size, this is likely due to the small number of *don't know* responses in general. Looking at the data in Table 4, one can see that many studies offering a *don't know* option did not have participants making this choice. In fact, the influential study had the most

don't know responses, seven, across the conditions with other studies that had high proportions of *don't know* responses being studies with much smaller samples. As removing this case does not affect the sign, magnitude, or significance of the analysis, it was not removed. A funnel plot (Figure 18) was also examined which showed all the points falling within the pyramid and distributed evenly around the midline.

Figure 17

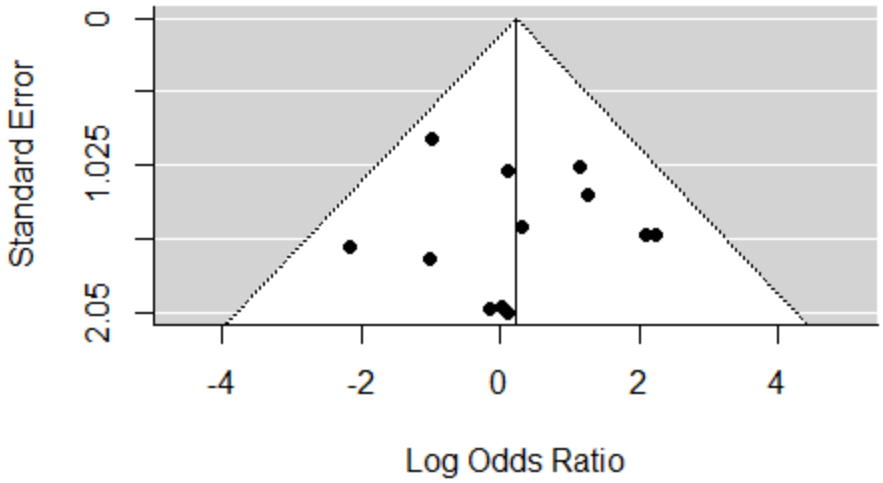
Forest plot showing effect of high vs low stress for don't know responses on TA line-ups.



Note: positive LOR indicates more *don't know responses* by the low stress group.

Figure 18

Funnel plot showing LORs for don't know responses on TA line-ups.



Collapsed TP and TA

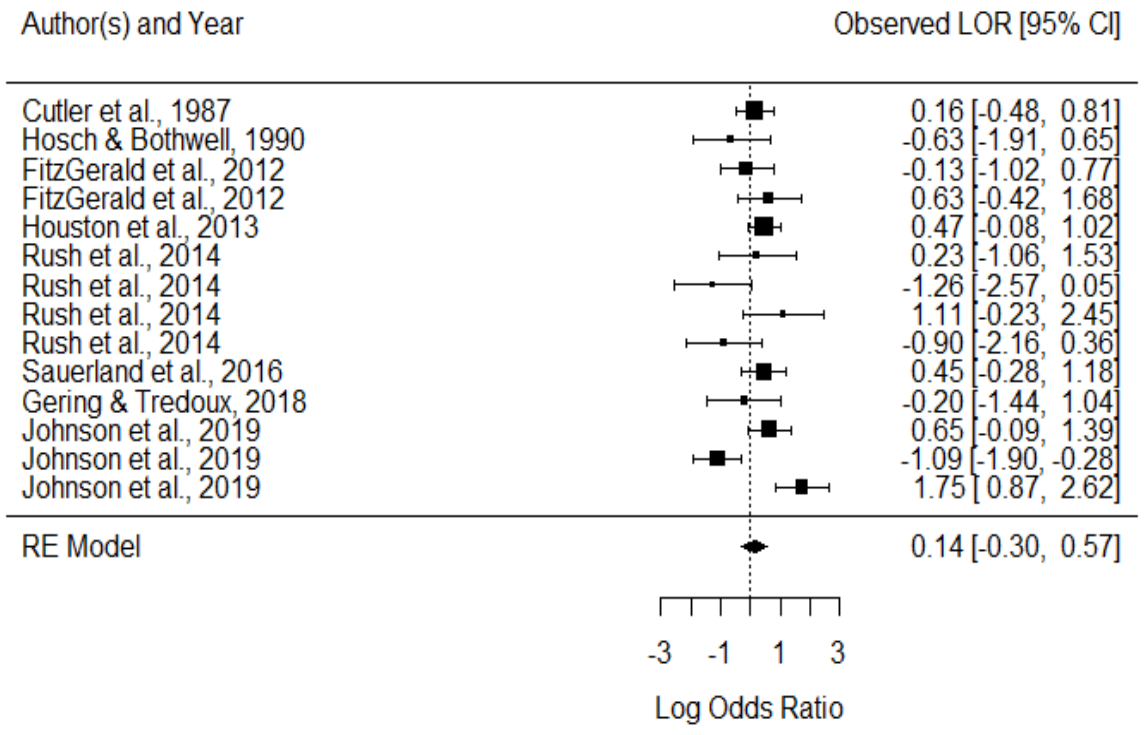
As some previous studies have collapsed correct and incorrect responses across TP and TA line-ups, including some of the studies included in this review ,this was also done here. Correct responses are hits for TP line-ups and rejections for TA line-ups. For incorrect responses, false alarms are considered for both TP and TA line-ups.

Collapsed Correct Responses. Figure 19 shows the effect sizes, LOR, for correct responses collapsed across TP and TA line-ups per study. As with the analysis of correct rejections, this analysis was based on $k = 14$ effects and $N = 1265$. As not all studies reported both TP and TA line-up data, the k value is low. In contrast, the N value is high as there is more data per study included. The average LOR is 0.13 ($SE = 0.22$), 95% CI [-0.30, 0.57]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant ($Z = 0.62$, $p = .535$). This suggests that there is no difference between high and low stress groups for correct responses collapsed across TP and TA line-ups. Looking at the tests of heterogeneity one can see high heterogeneity $I^2 =$

67.57%, that is significant $Q(13) = 37.23$, $p < .001$ suggesting that a moderator might show additional effects. However, no significant moderators were observed. Influence plots revealed no outliers or influential cases. A funnel plot was also analysed showing an even distribution with few points falling outside of the pyramid's area (Figure 20). As this is a composite variable, not always reported, it may be less affected by publication bias. Yet, as it is comprised of other measures subject to publication bias, it is still worth checking. The funnel plot shows a range of effects including several close to zero, indicating that publication bias is likely not an issue for this measure in this sample.

Figure 19

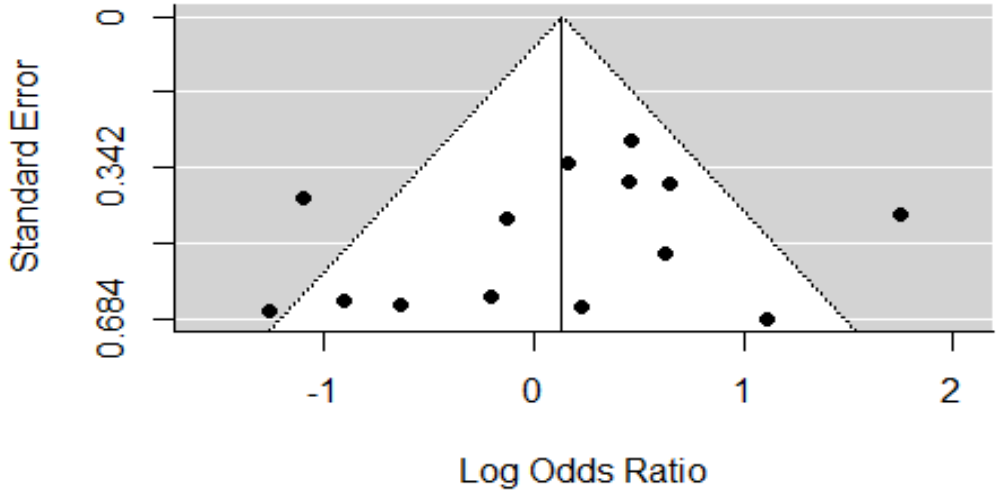
Forest plot showing effect of high vs low stress for correct responses collapsed across TP and TA line-ups.



Note: positive LOR indicates more correct responses by the low stress group.

Figure 20

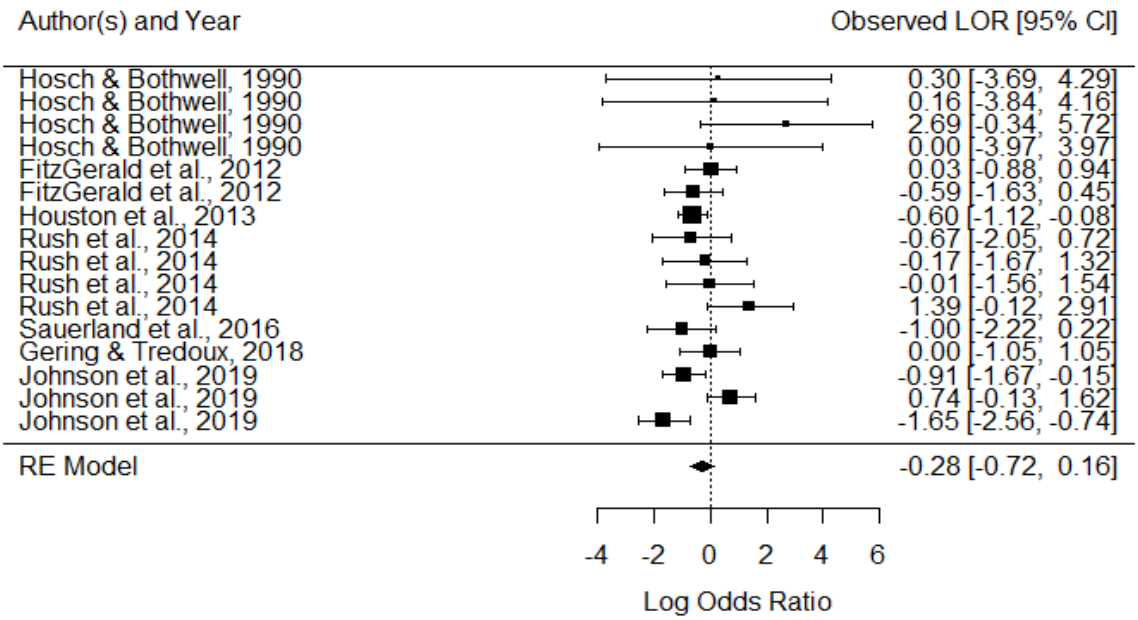
Funnel plot showing LORs for correct responses collapsed across TP and TA line-ups.



Collapsed False Alarms. Figure 21 shows the effect sizes, LOR, for collapsed false alarms across TP and TA line-ups per study. As with the analysis of correct rejections, this analysis was based on k = 16 effects and N= 1183. The average LOR is -0.28 (SE = 0.22), 95% CI [-0.72, 0.16]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant (Z = -1.26, p = .209). This suggests that there is no difference between high and low stress groups on false alarm rates for TA line-ups. Looking at the tests of heterogeneity one can see high heterogeneity I^2 = 67.57%, that is significant $Q(15) = 37.23$, p = .018 suggesting that a moderator might show additional effects. Despite this, as with the correct responses, no significant moderator variables were seen. Showing the same pattern as above, there were also no points of concern in any tests of influence or the funnel plot (Figure 22).

Figure 21

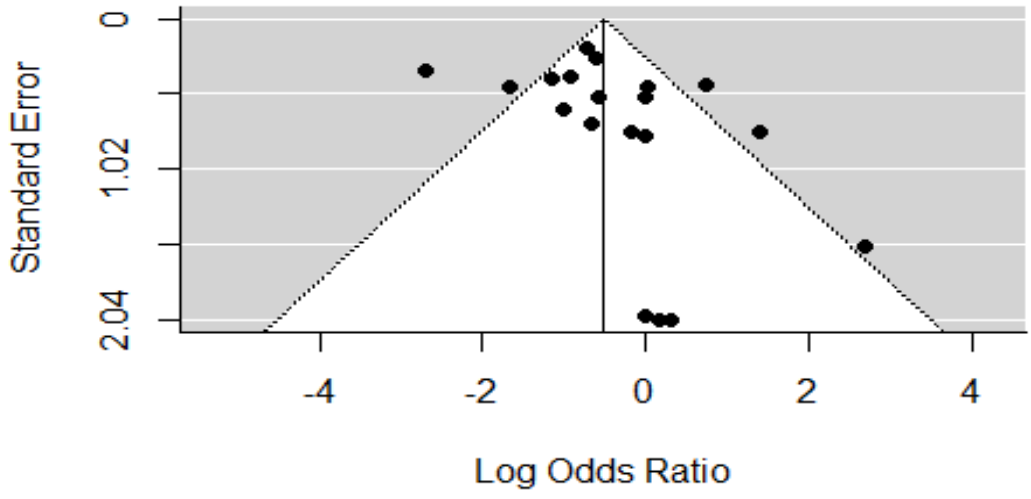
Forest plot showing effect of high vs low stress for false alarms collapsed across TP and TA line-ups.



Note: positive LOR indicates more false alarms by the low stress group.

Figure 22

Funnel plot showing LORs for false alarm collapsed across TP and TA line-ups.



This analysis was conducted again with the three effects from the study by Morgan et al. (2004). This analysis based on k = 19 effects and N= 1942 did produce a significant main effect, in the same direction as the trend without the additional effects, with an average LOR of -0.52 (SE = 0.24), 95% CI [-1.00, -0.03], (Z = -2.08, p = .037) indicating more false alarms across TP and TA line-ups by high stress group participants. However, as 759 of the 1942 participants (39%) come from the study by Morgen et al., (2004) study which is so different from the rest of the sample, conclusions should be drawn with caution. Other than the difference in stress induction and severity of the stressor previously mentioned this study did not provide participants with a *don't know* option which can increase false alarms. As such, this effect will be discussed but should not be considered a strong and defensible claim.

Signal Detection Sensitivity

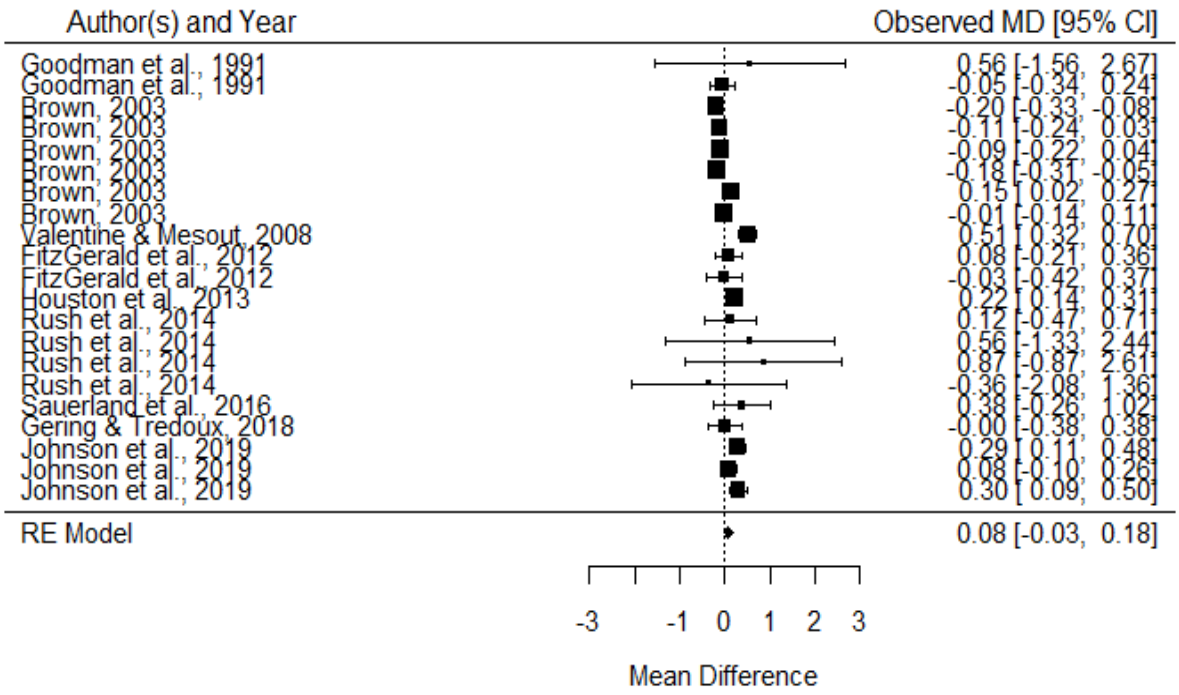
The following analysis of TP and TA line-ups using the signal detection measure d' uses MD rather than LOG as d' is a continuous, rather than binary, outcome. This measure is calculated as the difference between standardised correct decisions and false alarms, with a large positive d' value indicating high sensitivity and negative values indicating poor sensitivity. Unlike the analyses using only one outcome at a time, these difference scores allow us to see how good participants are at distinguishing true positives from foils. For this analysis, the unit of measure is groups compared across conditions rather than individual participants in a study.

TP line-ups. Figure 23 shows the effect sizes, MD, for d' on TP line-ups per study. This analysis was based on k = 21 effects and N= 1129. The average MD is 0.08(SE = 0.05), 95% CI [-0.03, 0.18]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant (Z = 1.44, p = .150). This suggests that there is no difference between high and low stress groups on d' for TP line-ups.

Looking at the tests of heterogeneity one can see high heterogeneity I^2 = 78.81%, that is significant $Q(19) = 92.86$, $p < .001$ suggesting that a moderator might show additional effects. The funnel plot (Figure 24) shows most of the points falling within the cone. Only points with very low estimated variability fall just outside of the desirable region. The moderator analysis indicated that none of the variables produced a significant moderator effect. A look at the influence plots indicated that the study by Valentine & Mesout (2008) is overly influential. Despite this, removing it resulted in not significant change MD is 0.04 (*SE* = 0.05), 95% CI [-0.05, 0.13], $Z = 0.90$, $p = .377$).

Figure 23

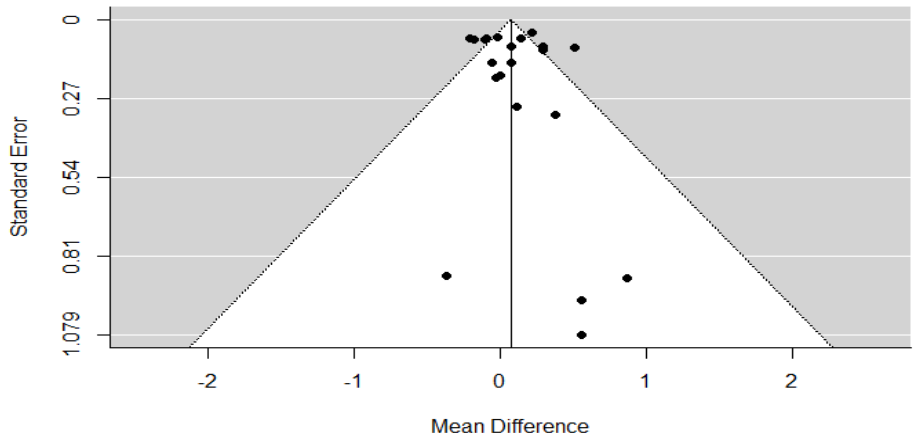
Forest plot showing MD of d' between high and low stress groups on TP line-ups.



Note: positive MD indicates greater signal sensitivity by the low stress group.

Figure 24

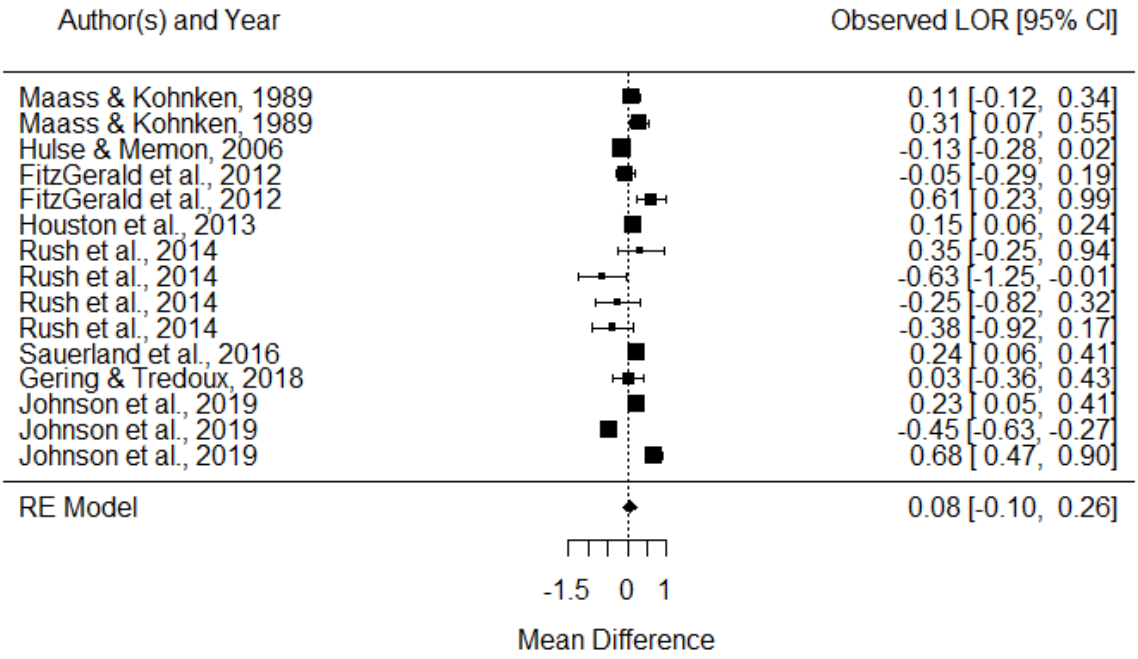
Funnel plot showing effect of high vs low stress for d' values on TP line-ups.



TA line-ups. Figure 25 shows the effect sizes, MD, for d' on TP line-ups per study. This analysis was based on $k = 15$ effects and $N = 688$. The average MD is 0.08 ($SE = 0.09$), 95% CI [-0.10, 0.26]. A test using normal theory to check if this effect size differs from 0 (representing outcomes that are equally likely) was not significant ($Z = 0.54$, $p = .583$). This suggests that there is no difference between high and low stress groups on d' for TA line-ups. Looking at the tests of heterogeneity one can see high heterogeneity $I^2 = 88.99\%$, that is significant $Q(14) = 100.65$, $p < .001$ suggesting that a moderator might show additional effects. However, as with the other TA analyses there were no significant moderator effects. The influence plots showed no outliers, influential cases, or high residuals. The funnel plot for this analysis (Figure 26) does show several points falling outside of the ideal pyramid area. As they fall on both sides of the pyramid suggesting that any bias applies to both positive and negative results. As this was not a variable reported in the studies, but rather one calculated for this analysis it should not be concerning as the authors of individual studies would not have considered this variable.

Figure 25

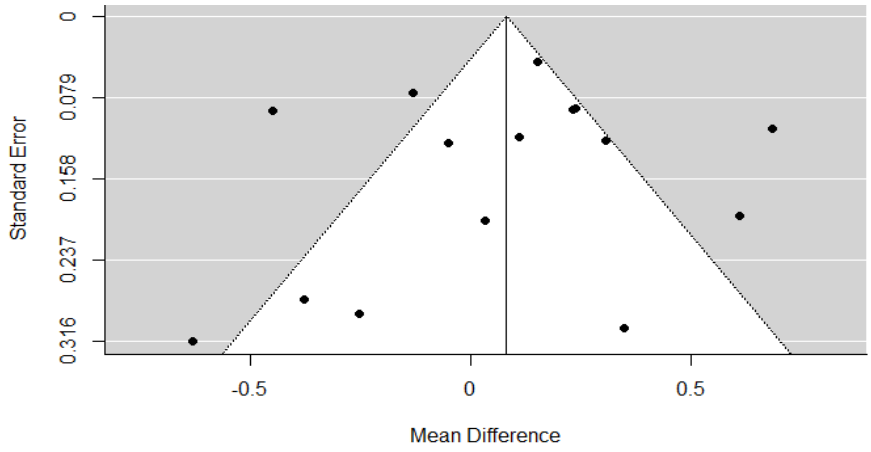
Forest plot showing MD of d' between high and low stress groups on TA line-ups.



Note: positive MD indicates greater signal sensitivity by the low stress group.

Figure 26

Funnel plot showing effect of high vs low stress for d' values on TA line-ups.



Multilevel Linear Modelling

As much of the heterogeneity between studies in the meta-analyses remained unaccounted for, additional analyses were undertaken using MLM. This method allows one to run regressions where the effects of predictor variables can be clustered by some categories which appear in the data. In this case, the data was clustered by author as a proxy for study, as all of the articles were compiled by different authors and few articles contained multiple studies. As this analysis was conducted to account for residual heterogeneity between studies, previously discussed variables were not considered. The main variable of interest was induced stress as a continuous rather than dichotomous measure. A scale, discussed in the methods section, was developed so as to allow for quantitative comparison of stress induction across studies. Each experimental and control group per study was given a rating on a Likert type scale ranging from 1-9. This scale was used as a predictor of the outcome variables used in the meta-analysis, assuming first linearity and then a non-linear relationship between stress and performance.

TP hits. The model for correct identifications on TP line-ups, using 43 observations, was successful only when assuming a quadratic function. This matches the shape of the Yerkes-Dodson curve where stress will at first improve performance and then, at higher levels, hinder performance (Yerkes & Dodson, 1908). The coefficient estimate of the model was -0.62 *SE* (0.23), $p = .010$ and *pseudo-*$R^2 = 0.15$, suggesting that 15% of the variance is explained by the continuous stress measure. The MLM also showed that the Inter Class Correlation was 0.29, suggesting that differences between the studies account for 29% of the variation in hit rate. This echoes the findings from the meta-analyses which showed high heterogeneity between the studies, even when moderator variables were considered.

TP false alarms. The model for false identifications, using 38 observations, was

similarly only successful when assuming a quadratic relationship between the continuous stress measure and the outcome variable. The coefficient estimate of the model was 0.40 *SE* (0.19), $p = .042$ and *pseudo-*$R^2 = 0.11$, suggesting that 11% of the variance is explained by the continuous stress measure. The MLM also showed that the Inter Class Correlation was 0.16, suggesting that differences between the studies account for 16% of the variation in hit rate. The shape of this parabolic curve is positive, unlike the TP hits, showing the opposite pattern – that stress first reduces and then increases the number of false alarms. As false alarms are a mark of poor performance, the pattern here again matches the theory proposed by Yerkes and Dodson (1908) with performance first increasing with stress and then decreasing as stress is further increased.

The rest of the models using TP variables were all non-significant, as were those using TA line-up variables, variables collapsed across TP and TA line-ups as well as the signal detection measures. The samples for these analyses were all smaller than those for TP hits and false alarms with the ICC values varying widely from 0.13 to 0.93. It should be noted that the MLM does not account for sample size through weighted averages. This analysis is presented primarily to illustrate alternate ways of considering stress data as a continuous variable.

Discussion

This meta-analysis included 17 studies investigating the effect of stress on line-up tasks, more than any previous review. The studies included in the final sample had over 2000 participants, allowing for a robust aggregate over many strong effects. From the 17 studies included in the present analysis one can see that there are many differences in the methods used by different studies. This, despite only including high quality studies which performed a manipulation check on the stress induction and included a line-up task. Several differences in

both method and result were found within this body of literature. While most of these differences, whose effect was investigated in moderator analyses, did not produce significant differences in the results, they highlight the rigour needed in planning new studies to ensure that studies are comparable. This review highlights some of the reasons for lack of consensus in this body of literature and in discussing these will attempt to provide clarity and make recommendations so that future research improves on existing work.

Stress Measurement. Of greatest concern is the varying standard of stress induction and manipulation checks used. Despite acknowledging that stress is experienced both subjectively and physiologically and that both these elements can affect psychological processes, studies in this literature have not employed a standardised battery of measures. Some excluded studies did not include manipulation checks and some included studies used either self-report or physiological checks rather than both. The lack of standard measures means that even when these measures are well used, they are often difficult to compare directly. This is true for self-report measures using different scales, or only one item rather than a multi-item scale, as well as in physiological measures where studies vary between cortisol, heart rate, blood pressure and skin conductance measures but rarely use more than one. Only four of the 16 studies used both self-report and physiological measures concurrently. As stress affects the body through at least two channels, namely the SAM and HPA axes of s activation and people are found to be affected differently both in reaction to the stressor and in the performance effects of stress, typically by perceiving the stressful event and its consequences differently, a more detailed set of measures would allow for greater understanding of how stress affects outcomes of interest. As research methods have advanced, methods such as MLM allow more nuanced manipulation checks than before.

Manipulation checks are crucial, but a simple check of differences between groups is not necessarily sufficient. This reduction of continuous measures into dichotomous variables

runs against the best practice methods in other bodies of literature and is an area which should be improved (Senn, 2005). As early models of physiological stress and performance showed non-linear and continuous relationships it stands to reason that studies should attempt to use such models (Yerkes & Dodson, 1908; Easterbrook, 1959). Only two studies in this review used more than two levels of stress, one by Hosch and Bothwell (1990) and the other by Johnson et al. (2019). Even these studies using three points do not treat stress as a continuous variable, making the result difficult to test against the Yerkes-Dodson law. Beyond this, a more rigorous standard of reporting is required, to match the levels seen in all other aspects of quantitative psychological research. Several studies, albeit primarily older ones, only reported that manipulation checks took place without reporting the means or measures used. Others reported means without some measure of variance, using non-standard measures for which no population statistics can be found. In the literature on neuropsychology, stress is measured at various points during an experiment to show that levels changed and later returned to normal after some manipulation. In contrast the eyewitness literature has often settled for single point measures. Such measurement can be sufficient in showing group differences, it is a low-resolution measure as levels of stress and anxiety differ greatly between people and repeated measures showing changes within individuals are a better solution.

Stress Induction. Within this set of studies, stress was manipulated in a variety of ways including live arousing events, videos and lab-based stressors. This variety is a positive as, poor measurement notwithstanding, it allows for the empirical comparison of methods. As questions remain over the importance of ecological validity of stress manipulations, it is critical to test whether studies in this literature that either do not use a live event or which unpair the stressor and encoding remain valid. As discussed in the literature review, emotion is expected to provide benefits only for task related behaviour. The results of this analysis

showed no moderation effect for method of stress induction which align with the findings in the neuropsychology literature. Those findings show that increase in physiological measures of stress affect brain regions related to memory whether the tasks are related or not. Most focus on the effect of cortisol on hippocampal regions related to memory and show that these areas are impaired by cortisol (Shield et al., 2016). While it would be tempting to conclude that this holds true for eyewitness memory, the inconsistencies in measurement and reporting do not allow for such strong conclusions to be made. As attention is affected by stressors, with a narrowing of focus on the source, a strong theoretical case can still be made that stressor and encoding tasks should be related (Plieger et al., 2016). Having a task that induces stress during the crime would still appear to be the most ecologically valid solution as stress is induced by the target and is induced concurrently with encoding. As the physiological stress reaction changes over time, with different neurotransmitters and hormones being released, the timing would appear to be as important as the task-relatedness of the stressor (Joels et al., 2011). Yet, this was not seen in the results of this study with tasks occurring before or during encoding having similarly varied effects.

Outcomes Measured. As with stress induction, the outcome variables measured in this set of studies varied greatly. For individual studies, including only a TP or a TA line-up reduces the richness of the analysis. Being able to see how your participants behave when the target is not present can provide insight into their choosing rates. Furthermore, informing the participants that targets may or may not be present can affect participant decision making. Building on that, not providing participants with the options of rejecting the line-up or responding that they *don't know* whether the participant is present is also important. These options are recognised as important as they stop low confidence participants from guessing (Wells et al., 2020). For this analysis, it means that the findings related to the latter options are less robust as there were fewer data points.

Main Statistical Findings. The main random effects meta-analysis showed no overt effect of stress on eyewitness identifications for either TP or TA line-ups, on any of the outcome measures. Thus, we must reject our two initial hypotheses that stress at encoding would adversely affect line-up performance of eyewitnesses. This finding was in contrast to that of Deffenbacher et al. (2004), upon whose finding our initial hypotheses were based. The current finding was not a result of no studies showing an effect but rather that studies showed a range of positive and negative results, as well as some null results, which balanced out, resulting in no significant main effect of stress. As such we cannot simply say that there is no effect of stress at encoding on later line-up decisions but rather that this effect varies across studies as the result of some other variables. Furthermore, the results of the main analysis showed that there was significant heterogeneity between the studies on all the outcome measures. This was likely due to the range of methods used in the studies, as well as the outcomes measured and was confirmed by statistical tests. Various moderator variables were analysed to see if the source of this variation could be attributed to some of the common differences in method between the studies in the current sample.

Moderators. There were many differences in method worth noting and which were coded for analysis. These differences reflect recognised methods in the literature and are useful variables that can only be adequately measured between studies within the stress and eyewitness literature. The moderator variables tested in the meta-analysis included the type of line-up, the medium of the line-up, the delay between encoding and recognition, the participants' age as well as the presence or absence of a weapon. Of these effects, only the type of line-up (either simultaneous or sequential presentation) had a significant effect. It should be noted that all the effects calculated using sequential line-ups came from one study by Brown (2003), albeit with 244 participants and as such further evidence is needed in support. This effect was seen in hits and false alarms for TP line-ups providing some support

for Hypothesis 3, that moderator variables would explain and quantify some of the varied findings between experiments.

Early studies explained differences in the results seen between simultaneous and sequential line-ups as the result of relative or absolute judgements (Wells, 1984). In the eyewitness literature, a relative line-up decision is one where the closest match to the perpetrator out of the various faces seen is chosen, whereas an absolute judgement is one where a face is chosen as the closest match only when it passes some threshold for similarity with the chooser's memory. Relative decisions are easiest to make in simultaneous line-ups as all the faces may be compared unlike those in sequential line-ups (Wells, 1984). However, a different reasoning based on signal detection theory was used by to explain the differences seen between these line-ups in the meta-analysis reported by McQuiston-Surrett et al. (2006). Considering the results of Meissner et al., (2005), which used signal detection theory in several experiments where line-up type was manipulated, they surmise that sequential line-ups make participants more conservative rather than more accurate. In their final experiment, instructions were given that made all participants more conservative. In doing so, the difference between the line-up types was negated (Meissner et al., 2005). Such experiments should be repeated with stress as an additional experimental manipulation to better understand the interaction between stress at encoding and line-up presentation on recognition performance.

While studies in the eyewitness literature have compared line-up presentation methods and studies in the neuropsychology literature have compared stress induction methods, no comparison of sequential and simultaneous line-up presentation has been made within the stress and eyewitness literature (McQuiston-Surrett et al., 2006, Smeets et al., 2012). As this literature is already combining two distinct bodies of research, it is difficult for individual studies to make comparisons of all the possible combinations of variables. As the

outcome measures used are often single choices of individual participants, this field already struggles with smaller effect sizes when compared to other facial recognition research which uses repeated measures in old/new paradigms (Deffenbacher et al., 2004; Sauerland, et al., 2016). As such, while different stress induction methods have been compared on memory tasks using word lists, no such study has been conducted using line-up outcome measures (Kirschbaum et al., 1993). Similarly, while there have been two meta-analyses comparing simultaneous and sequential line-ups, these line-ups have not been compared in interaction with stress (Steblay et al., 2001; McQuiston-Surrett et al., 2006). As the current review found that stressed participants performed better on sequential line-ups, it may suggest differences in willingness to make a decision. While stressed participants were only slightly more likely to get a successful hit on sequential line-ups, they were significantly less likely to identify a foil. The logical conclusion is that stressed participants with weaker memory traces were less likely to make a decision than non-stressed participants, but only if the line-up was presented sequentially. This suggests that the effect of sequential line-ups making participants more conservative was more pronounced on stressed participants (McQuiston-Surrett et al., 2006). The counter view, using the relative *vs* absolute judgement strategies would suggest that when using relative decision-making strategies, both high and low stress participants are equally prone to error and that high stress participants are less prone to error when making absolute judgements. This would suggest that the threshold of similarity is higher for stressed participants, making them less likely to make an identification (Wells, 1984). With the current data, it is not possible to determine which of these is the case. There are too few studies using sequential line-ups in the stress literature, a sparsely populated field itself and so only a tentative conclusion can be drawn. Namely, that sequential line-up presentation appears to favour stressed participants.

Multilevel Linear Modelling. The results of the MLMs showed some interesting

results using the continuous scale of stress. While no significant effects were found using the dichotomous stress variable, in either the MLM or meta-analyses, a significant non-linear relationship was found for hits and false alarms on TP models. These effects had opposite signs indicating the same pattern of performance (first improving and then deteriorating with stress) as hits and false alarms are opposite results once a participant commits to making a decision. This finding is in line with the Yerkes-Dodson law which predicts that performance will first increase with stress, until it reaches a maximum at which point performance should decrease (Yerkes-Dodson, 1908). The magnitude of the effect for false alarms was a little smaller, yet significant despite a smaller sample, which may reflect that some studies allowed participants not to make an identification. This relationship was not seen for rejections or *don't know* responses on the TP line-ups. It was also not seen on any of the TA line-ups, when results were collapsed across line-ups or on the signal detection measures. This may be due to the smaller sample sizes in the other analyses or the very high ICC values seen in some, as high heterogeneity between studies on some variables produces a great range in results. The stress scale as a continuous measure was a better predictor of two of the outcome variables considered and no worse at predicting the other variables, supporting Hypothesis 4 which suggested that more information on the stress-memory relationship could be gained by considering stress as a continuous variable.

Previous Studies. While some significant moderator variables were found in this study, differences in effect were still seen which were not explained by the results of this study. The values of residual heterogeneity between studies suggests that there are other factors which differ between studies which were not captured in this analysis. As such, other important moderator variables may exist, which when included would remove or reduce this residual heterogeneity. In contrast to the meta-analysis by Shields et al. (2016), who studied the relationship between stress and memory in non-eyewitness studies; few significant

moderator variables were found. This could be due to the sample size used in this meta-analysis of stress and the eyewitness literature in contrast to that used in the review of stress and memory more generally (15 articles including 17 studies, 31 effects and 2011 participants, compared with 113 studies and 6216 participants in the Shields et al. (2016) study). However, the study by Shields et al. (2016) also included effects of stress post-encoding and at retrieval, allowing for more potential effects. The main differences found in that study were that delay between encoding and retrieval, as well as relatedness between the stressor and the memory event were found to be significant moderators (Shields eta al., 2016). This should be examined further in the eyewitness literature. If the lack of significance in the current study is not due to a limitation of statistical power, which is unlikely given the number of articles and effects included, the best explanation in theory might be that the salience of the eyewitness events reduces the impairing effect of stress at, or prior to encoding.

The role of attention may also explain some of the differences seen between the literature on eyewitness memory and the more general body of research on stress and memory. As moderately emotionally arousing events will be better remembered as they attract attention, this may explain the better performance by stressed participants in the eyewitness, as opposed to general memory literature. (De Quervain et al., 2009). The nature of the eyewitness task may engage participants' attention more intensely as the events used by studies in this meta-analysis were generally engaging. The benefit from the focused attention may counteract the defects in memory caused by the stress response. Additionally, as these studies used line-up tasks, they may have been testing for central information only. Eyewitness tasks that ask for details of an event as well as using a face recognition task might find differences in performance across groups between these tasks. The previous meta-analysis in this area found negative effects of stress for both face recognition and recall of

details in eyewitness tasks (Deffenbacher et al., 2004). Differences in the results of the current study and the previous meta-analysis will be discussed in the following section. As the current study only considered line-up tasks, there will be no comparison on the other side of the eyewitness tasks, namely witness memory of details of the events. Future research should be conducted to update the state of that body of literature.

Differences between the current study and the previous meta- analysis. One note-worthy difference is that one particularly large effect that had been included in the meta-analysis by Deffenbacher et al. (2004) was excluded from the current analysis. This study, by Buckhout et al., (1974), yielded a very large effect size (h = -3.02) in their analysis. The study by Buckhout et al. (1974) was excluded as it divided participants first by line-up choice and then provided average stress scores. Although a difference in stress was shown between those who scored successful hits and those who had chosen foils, it created groups unlike any other study in the literature. The mean stress scores reported, using a self-report measure with no measure of variance given, were 5.43 for the hit group and 4.84 for the false alarm group (Buckhout et al., 1974). This then, would appear to be a small difference in stress resulting in a large difference in performance. By then using this effect to dichotomise between high and low stress participants, an artificially high line-up performance effect size occurs as the 'low stress' group is in fact the perfect hit group and the high stress group is the false alarm group. This effect, greater than any other in the study by a factor of 2.5, likely skewed the results of that study. Deffenbacher et al. (2004) did note that this was a particularly large effect size but attributed this to the use of a "rather realistic, live staged crime, rather than a filmed one" (Deffenbacher et al., 2004, p. 697). Their study subsequently found a moderation effect for live crimes as well. However, the present study included 20 effects which used a live encoding effect, including staged crimes, swimming lessons for children and university students being accused of plagiarism (Sauerland et al., 2016; Fitzgerald et al., 2012; Johnson

et al., 2019). While these studies vary in intensity, they are all live and more realistic than a filmed event. Yet, no effect for type of encoding was found in the current review. As such, it remains likely that the method for dividing the groups on decision outcome and then showing a difference in stress is the most likely reason for the extreme effect size found by Buckhout et al. (1974). The best standard of first creating groups, then inducing stress and finally measuring performance should be adhered to in order to ensure accurate and comparable results.

Future research. As this body of eyewitness research aims to induce stress and study its effect, greater attention should be paid to the manipulation checks used and the recording of their results. While individual studies may not find much benefit in a more robust manipulation check, when reviewing the literature in the future it would be greatly beneficial to have a more detailed breakdown of the results of the stress induction. Rather than only using two stress groups and making the stress variable discrete, using a greater number of groups or keeping stress intensity as a continuous variable would offer more insight into the stress-performance relationship. As differences have been found in previous stress research, outside the eyewitness literature on the effects of stress on implicit and explicit memory, future studies should consider asking participants about their line-up decision strategies (Sandi, 2013). A decision-making strategy might serve as a reasonable proxy for memory system used. This could be an interesting moderator variable to consider in studies on stress and line-up performance. The use of SDT measures would help in this regard as they provide greater insight into decision making strategies than ratio-based measures (Wixted & Mickes, 2014). For this to be done effectively, studies should employ both TP and TA line-ups and allow participants the full range of recommended decisions, including line-up rejections and *Don't know* responses. As an empirical study to further investigate the interaction between stress and line-up presentation method is needed, such measures may add to the

understanding of how line-up presentation affects decision making, willingness to choose and sensitivity.

Another variable relating to the outcome not mentioned in the literature, is task difficulty. As the cue-availability theory and studies on attention in general, indicate that task difficulty affects attention narrowing and memory this variable should be measured (Easterbrook, 1959, Plieger et al., 2016). The studies in this analysis showed a high degree of heterogeneity, which remained when moderators were considered. This indicates that other variables which influence the outcome were not considered. One such variable may be task difficulty. As task difficulty might affect performance at a given level of stress or might interact with stress reducing performance to a greater extent at higher levels of stress, this should also be considered. This variable is difficult to determine post-hoc and so should be measured during experiments.

Future research should also continue to investigate a potential interaction between stress and time delay between the crime and line-up task. Half of the studies used in this study employed delays of less than 24 hours, with the rest ranging from 24 yours to a year's follow up. As such there is a skew in the delays used in the current set of studies with the mean delay being far greater than the median. More studies with a moderate delay, between a day and a week would be helpful in further unpacking the effect of delay between encoding and recognition and how it affects high and low stress participants differently. As witnesses must provide evidence at varying times after a crime, either in court or to the police, it is important to know how stress affects the well studied forgetting curve.

Of great importance, and an overarching limitation of stress induced in laboratory settings in comparison to during real crimes, is the intensity and consequence of stress experienced. While the pairing of stressor and perpetrator has been discussed, a range of outcomes and reactions exist in real world settings that cannot be simulated in laboratory

settings. As real-world eyewitnesses are often in imminent danger the level of fear experienced is beyond that which can be achieved in experiments. Eyewitnesses may also attempt to escape a crime scene, exhibiting a 'flight' response which experiment participants are unlikely to consider. While the physiological and neurological effects of stress are relevant to the strength of encoding and later recognition, the lack of real threat and subsequent behaviour might have an effect beyond that which can be measured in experiments such as those considered in this analysis. While the role of attention has been discussed as possibly narrowing participants' focus on the perpetrator, a witness may focus their attention on escape or internal thoughts of fear. In such a situation attention narrowing might result in the exclusion of perpetrator features rather than a focus on the perpetrator. Future research should explore reactions of real participants to better understand the phenomena the laboratory experiments attempt to simulate.

Conclusion. Within this body of literature, attempts have been made to induce stress incidentally as part of the mock crime or memory event as well as induce a level of stress high enough to be comparable to real criminal events. These two aspects of stress and eyewitness experiments are both considered crucial. If the stress induction is too weak, the small stress event will not accurately match the physiological reaction and subsequent chemical processes in the brain that have been studied in the neuropsychological literature. As these processes which release cortisol affect regions in the brain responsible for memory, they are often thought to be the primary reason why stress affects memory. However, stress also affects attentional processes which play a role in the quality of encoded memories. As such, it is important to link the stressor and the memory event. While this is a difficult balance to achieve within any one study, meta-analytic methods allow for comparisons of effects across studies. As such, future research might attempt to induce higher levels of stress even if these methods are not as intrinsically linked to the memory event. Alternatively,

enhancing the incidental stress reaction, through methods such as cortisol shots or the ingestion of tablets should be considered. While these might cost individual studies some ecological validity, it may allow for more robust expressions of the stress response while allowing reviews to determine which of the many ecological considerations are most important to replicate.

As research is cumulative, attempting to answer the same question with a variety of methods, where these are carefully documented and outcomes consistently measured, will allow for interesting comparisons across studies to draw conclusions beyond the scope of any one study. The review contained in this thesis identified several areas of under reporting in this literature. Improvements in this regard by future studies will allow for more depth in comparing results allowing for a clearer understanding of the effects of stress at encoding on subsequent line-up performance. It also showed that there is still a gap in the literature regarding stress and eyewitness performance as this effect is neither simple nor clear.

This analysis showed that studies have found a varied effect of stress at encoding on line-up performance. Some of these differences were accounted for by moderator variables which interacted with stress. Further variation between studies was shown to be because of varying levels of stress dichotomised into high and low. By using continuous measures of stress, the MLMs presented here show that the stress-performance relationship can be better explained. Future studies should take this into consideration, as a better representation of the underlying stress construct may better explain the relationship in general as well as the interactions between stress and other variables.

As eyewitnesses will continue to experience stress and affect criminal cases with their testimony, it is important to have a clear understanding of factors which affect recognition of perpetrators. While this review does not definitively answer the question, it is hoped that it will provide future researchers with some direction when attempting to do so. As the pool of

knowledge grows along with the tools to undertake such research more satisfactory answers are likely to be found. However, it is important to take stock of what has been done and to build on the current state of the science. Although this review was often critical, much good work has been done. Despite the lack of clarity in the existing literature, existing research has provided a solid base upon which to build. By continuing to build on this body of research through clear measurement and recording of critical variables, a clear and definitive answer as to the effect of stress at encoding on witness memory and line-up performance can be achieved.

References

Arnsten, A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature reviews neuroscience*, *10*(6), 410-422. https://doi.org/10.1038/nrn2648

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, *2*(4), 89-195. https://doi.org/10.1016/b978-0-12-121050-2.50006-5

Bäumler, G., & Lienert, G. A. (1993). Re-evaluation of the Yerkes-Dodson law by nonparametric tests of trend. *Studia Psychologica*. *35*(4-5), 431-436.

Bird, C. M. (2017). The role of the hippocampus in recognition memory. *Cortex, 93*, 155-165. https://doi.org/10.1016/j.cortex.2017.05.016

*Buckhout, R., Alper, A., Chern, S., Silverberg, G., & Slomovits, M. (1974). Determinants of eyewitness performance on a lineup. *Bulletin of the Psychonomic Society,* 4, 191–192. https://doi.org/10.3758/bf03334241

Brewin, C. R., Rose, S., Andrews, B., Green, J., Tata, P., McEvedy, C., Turner, S., & Foa, E. B. (2002). Brief screening instrument for post-traumatic stress disorder. *The British Journal of Psychiatry*, 181(2), 158-162. https://doi:10.1192/bjp.181.2.158

*Brown, J. M. (2003). Eyewitness memory for arousing events: Putting things into context. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition,* 17(1), 93-106. https://doi.org/10.1002/acp.848

Christianson, S. Å. (1992). Emotional stress and eyewitness memory: a critical

review. *Psychological bulletin, 112*(2), 284. https://doi:10.1037/0033-2909.112.2.284

Clark, S. E., Benjamin, A. S., Wixted, J. T., Mickes, L., & Gronlund, S. D. (2015). Eyewitness identification and the accuracy of the criminal justice system. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 175-186. https://doi.org/10.1016/bs.plm.2015.03.003

Clifford, B. R., & Hollin, C. R. (1981). Effects of the type of incident and the number of perpetrators on eyewitness memory. *Journal of Applied Psychology,* 66, 364–370. https://doi.org/10.1037/0021-9010.66.3.364

*Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identification. *Law and Human Behavior, 11*(3), 233-258. https://doi.org/10.1007/bf01044644

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 and Cognition. 8(*3). https://doi.org/10.1016/j.jarmac.2019.05.006

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*(6), 687. https://doi:10.1007/s10979-004-0565-x

Dew, I. T., & Cabeza, R. (2011). The porous boundaries between explicit and implicit memory: behavioral and neural evidence. *Annals of the New York Academy of Sciences, 1224*(1), 174-190. https://doi.org/10.1111/j.1749-6632.2010.05946.x

De Quervain, D. J.-F., Aerni, A., Schelling, G., & Roozendaal, B. (2009). Glucocorticoids and the regulation of memory in health and disease. Frontiers in Neuroendocrinology,

30(3), 358–370. https://doi:10.1016/j.yfrne.2009.03.002

Diamond, D. M., Campbell, A. M., Park, C. R., Halonen, J., & Zoladz, P. R. (2007). The temporal dynamics model of emotional memory processing: a synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law. *Neural plasticity, 2007*. https://doi:10.1155/2007/60803

Drexler, S. M., & Wolf, O. T. (2017). Stress and memory consolidation. *Studies in Neuroscience, Psychology and Behavioral Economics*, 285–300. https://doi:10.1007/978-3-319-45066-7_17

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological review, 66*(3), 183. https://doi.org/10.1037%2Fh0047707

Fitzgerald, R. J., & Price, H. L. (2015). Eyewitness identification across the life span: A meta-analysis of age differences. *Psychological Bulletin*, 141(6), 1228-1332. https://doi: 10.1037/bul0000013

*Fitzgerald, R. J., Price, H. L., & Connolly, D. A. (2012). Anxious and nonanxious children's face identification. *Applied Cognitive Psychology*, 26(4), 585-593. https://doi.org/10.1002/acp.2833

Fitzgerald, R. J., Price, H. L., & Valentine, T. (2018). Eyewitness identification: Live, photo, and video lineups. *Psychology, Public Policy, and Law*, 24(3), 307. https://doi.org/10.1037/law0000164

*Gering, M.A., & Tredoux, C.G (2018). Stress has no effect on witness memory or face identification. *Unpublished manuscript*.

Godoy, L. D., Rossignoli, M. T., Delfino-Pereira, P., Garcia-Cairasco, N., & de Lima Umeoka, E. H. (2018). A comprehensive overview on stress neurobiology: basic concepts and clinical implications. *Frontiers in behavioral neuroscience*, *12*, 127.

*Goodman, G. S., Hirschman, J. E., Hepps, D., & Rudy, L. (1991). Children's memory for stressful events. *Merrill-Palmer Quarterly* (1982-), 109-157.

Gray, J. D., Kogan, J. F., Marrocco, J., & McEwen, B. S. (2017). Genomic and epigenomic mechanisms of glucocorticoids in the brain. *Nature Reviews Endocrinology*, *13*(11), 661. https://doi.org/10.1038/nrendo.2017.97

Gronlund, S. D., Mickes, L., Wixted, J. T., & Clark, S. E. (2015). Conducting and eyewitness lineup: How the research got it wrong. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 63, pp. 1-43). New York: Academic Press. https://doi.org/10.1016/bs.plm.2015.03.003

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. https://doi:10.1177/ 0963721413498891.

Hanoch, Y., & Vitouch, O. (2004). When less is more: Information, emotional arousal and the ecological reframing of the Yerkes-Dodson law. *Theory & Psychology*, *14*(4), 427-452. https://doi.org/10.1177%2F0959354304044918

Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: the role of active forgetting in memory. *Trends in cognitive sciences*, *17*(3), 111-120. https://doi.org/10.1016/j.tics.2013.01.001

Henry, M., Wolf, P. S., Ross, I. L., & Thomas, K. G. F. (2015). Poor quality of life, depressed mood, and memory impairment may be mediated by sleep disruption in patients with

Addison's disease. *Physiology and Behavior, 151,* 379-385. https://doi: 10.1016/j.physbeh.2015.08.011

Het, S., Ramlow, G., & Wolf, O. T. (2005). A meta-analytic review of the effects of acute cortisol administration on human memory. *Psychoneuroendocrinology*, 30, 771–784. https://doi:10.1016/j.psyneuen.2005.03.005.

Hope, L., Gabbert, F., & Fisher, R. P. (2011). From laboratory to the street: Capturing witness memory using the Self-Administered Interview. *Legal and Criminological Psychology, 16*(2), 211-226. https://doi: 10.1111/j.2044-8333.2011.02015.x

Hope, L., & Wright, D. (2007). Beyond unusual? Examining the role of attention in the weapon focus effect. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 21*(7), 951-961. https://doi.org/10.1002/acp.1307

*Hosch, H. M., & Bothwell, R. K. (1990). Arousal, description and identification accuracy of victims and bystanders. *Journal of Social Behavior and Personality, 5*(5), 481. https://doi.org/10.1111/j.1559-1816.1984.tb02257.x

Hoscheidt, S. M., LaBar, K. S., Ryan, L., Jacobs, W. J., & Nadel, L. (2014). Encoding negative events under stress: High subjective arousal is related to accurate emotional memory despite misinformation exposure. *Neurobiology of Learning and Memory, 112*, 237–247. https://doi:10.1016/j.nlm.2013.09.008

*Houston, K. A., Clifford, B. R., Phillips, L. H., & Memon, A. (2013). The emotional eyewitness: The effects of emotion on specific aspects of eyewitness recall and recognition performance. *Emotion, 13*(1), 118. https://doi.org/10.1037/a0029220

*Hulse, L. M., & Memon, A. (2006). Fatal impact? The effects of emotional arousal and

weapon presence on police officers' memories for a simulated crime. *Legal and Criminological Psychology, 11*(2), 313-325. https://doi.org/10.1348/135532505x58062

Joëls, M., Fernandez, G., & Roozendaal, B. (2011). Stress and emotional memory: a matter of timing. *Trends in cognitive sciences, 15*(6), 280-288. https://doi.org/10.1016/j.tics.2011.04.004

*Johnson, T., Gering, M.A., Nortje, A., & Tredoux, C.G. (2019), The effects of stress on eyewitness memory and suspect identification in photographic lineups Unpublished manuscript.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1476), 2109-2128.

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*(1-2), 76-81. https://doi.org/10.1159/000119004

Kirschbaum, C., Wolf, O. T., May, M., Wippich, W., & Hellhammer, D. H. (1996). Stress-and treatment-induced elevations of cortisol levels associated with impaired declarative memory in healthy adults. *Life sciences, 58*(17), 1475-1483. https://doi.org/10.1016/0024-3205(96)00118-x

Krix, A. C., Sauerland, M., Raymaekers, L. H., Memon, A., Quaedflieg, C. W., & Smeets, T. (2016). Eyewitness evidence obtained with the Self-Administered Interview© is unaffected by stress. *Applied Cognitive Psychology, 30*(1), 103-112. https://doi:10.1002/acp.3173

Levine, L. J., & Edelstein, R. S. (2009). Emotion and memory narrowing: A review and goal-relevance approach. *Cognition & Emotion, 23(*5), 833–875. https://doi:10.1080/02699930902738863

*Lindberg, M. A., Jones, S., Collard, L. M., & Thomas, S. W. (2001). Similarities and differences in eyewitness testimonies of children who directly versus vicariously experience stress. *The Journal of genetic psychology, 162*(3), 314-333. https://doi.org/10.1080/00221320109597486

Lupien, S. J., Maheu, F., Tu, M., Fiocco, A., & Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and cognition, 65*(3), 209-237. https://doi.org/10.1016/j.bandc.2007.02.007

*Maass, A., & Köhnken, G. (1989). Eyewitness identification: Simulating the "weapon effect". Law and Human Behavior, 13(4), 397-408. https://doi.org/10.1007/bf01056411

Marr, C., Otgaar, H., Sauerland, M., Quaedflieg, C. W., & Hope, L. (2020). The effects of stress on eyewitness memory: A survey of memory experts and laypeople. *Memory & Cognition*, 1-21. https://doi.org/10.3758/s13421-020-01115-4

McQuiston-Surrett, D., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. Simultaneous Lineups: A Review of Methods, Data, and Theory. *Psychology, Public Policy, and Law, 12*(2), 137. https://doi.org/10.1037/1076-8971.12.2.137

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & cognition, 33*(5), 783-792. https://doi.org/10.3758/bf03193074

Meyer, T., Smeets, T., Giesbrecht, T., Quaedflieg, C. W., & Merckelbach, H. (2013). Acute stress differentially affects spatial configuration learning in high and low cortisol-responding healthy adults. *European Journal of Psychotraumatology, 4*(1), 1-9. https://doi:10.3402/ejpt.v4i0.19854

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews, 4*(1), 1-9. https://doi.org/10.1186/2046-4053-4-1

*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. https://doi:10.1016/j.ijlp.2004.03.004.

Murison, R. (2016). The neurobiology of stress. In *Neuroscience of pain, stress, and emotion* (pp. 29-49). Academic Press. https://doi.org/10.1016/B978-0-12-800538-5.00002-9

Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PloS one, 10*(7), e0120644. https://doi.org/10.1371/journal.pone.0120644

Nixon, P. G. F. (1982). Stress and the cardiovascular system. *The Practitioner*, *226*, 1589-1598

Packard, M. G., Cahill, L., & McGaugh, J. L. (1994). Amygdala modulation of hippocampal-dependent and caudate nucleus-dependent memory processes. *Proceedings of the National Academy of Sciences, 91*(18), 8477-8481. https://doi.org/10.1073/pnas.91.18.8477

Pickel, K. (2015). Eyewitness Memory. In Fawcett, J., Risko, E., & Kingstone, A. (Eds.), *The Handbook of Attention*. MIT Press, 485-502. ISBN: 9780262029698

Plieger, T., Felten, A., Diks, E., Tepel, J., Mies, M., & Reuter, M. (2016). The impact of acute stress on cognitive functioning: a matter of cognitive demands? *Cognitive Neuropsychiatry, 22*(1), 69–82. https://doi:10.1080/13546805.2016.1261014

*Read, J. D., Yuille, J. C., & Tollestrup, P. (1992). Recollections of a robbery. *Law and Human Behavior, 16*(4), 425-446. https://doi.org/10.1007/bf02352268

Roozendaal, B., McEwen, B. S., & Chattarji, S. (2009). Stress, memory and the amygdala. *Nature Reviews Neuroscience, 10*(6), 423. https://doi.org/10.1038/nrn2651

*Rush, E. B., Quas, J. A., Yim, I. S., Nikolayev, M., Clark, S. E., & Larson, R. P. (2014). Stress, interviewer support, and children's eyewitness identification accuracy. *Child Development, 85*(3), 1292-1305. https://doi.org/10.1111/cdev.12177

Sandi, C. (2013). Stress and cognition. Wiley Interdisciplinary Reviews: Cognitive Science, 4(3), 245-261. https://doi: 10.1002/wcs.1222

*Sauerland, M., Raymaekers, L. H., Otgaar, H., Memon, A., Waltjen, T. T., Nivo, M., Slegers, C., Broers, N.J., & Smeets, T. (2016). Stress, stress-induced cortisol responses, and eyewitness identification performance. *Behavioral sciences & the law, 34*(4), 580-594. https://doi:10.1002/bsl.2249

Senn, S. (2005). Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. *Proceedings of the International Statistical Institute, 55th Session, Sydney*.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., &

Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *Bmj, 349*. https://doi.org/10.1136/bmj.g7647

Shields, G. S., Sazma, M. A., McCullough, A. M., & Yonelinas, A. P. (2017). The effects of acute stress on episodic memory: A meta-analysis and integrative review. *Psychological bulletin, 143*(6), 636-777. https://doi.org/10.1037/bul0000100

Shields, G. S., Sazma, M. A., & Yonelinas, A. P. (2016). The effects of acute stress on core executive functions: A meta-analysis and comparison with cortisol. *Neuroscience & Biobehavioral Reviews, 68*, 651–668. https://doi:10.1016/j.neubiorev.2016.06.038

Smeets, T., Cornelisse, S., Quaedflieg, C. W., 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*(12), 1998-2008.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). Manual for the State-Trait Anxiety Inventory. *Palo Alto, CA: Consulting Psychologists Press.* https://doi:10.1037/t06496-000

Spielberger, C. D., & Vagg, P. R. (1984). Psychometric properties of the STAI: A reply to Ramanaiah, Franzen, and Schill. *Journal of Personality Assessment, 48*, 95-97. https://doi: 10.1207/s15327752jpa4801_16

Steblay, N. M. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior, 16*(4), 413-424. https://doi.org/10.1007/bf02352267

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law*

and human behavior, *25*(5), 459-473. https://doi.org/10.1023/a:1012888715007

Suero, M., Privado, J., & Botella, J. (2017). Methods to estimate the variance of some indices of the signal detection theory: A simulation study. *Psicologica: International Journal of Methodology and Experimental Psychology*, *38*(1), 149-175. https://doi.org/10.1016/j.jmp.2018.02.001

Thorley, C., Dewhurst, S. A., Abel, J. W., & Knott, L. M. (2015). Eyewitness memory: The impact of a negative mood during encoding and/or retrieval upon recall of a non-emotive event. *Memory, 24(*6), 838-852. https://doi:10.1080/09658211.2015.1058955

Troyer, A. K., & Craik, F. I. M. (2000). The effect of divided attention on memory for items and their context. *Canadian Journal of Experimental Psychology, 54*(3), 161–170. https://doi.org/10.1037/h0087338

Vakil, E., Wasserman, A., & Tibon, R. (2018). Development of perceptual and conceptual memory in explicit and implicit memory systems. *Journal of Applied Developmental Psychology, 57*, 16-23. https://doi.org/10.1016/j.appdev.2018.04.003

*Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. Applied Cognitive Psychology: *The Official Journal of the Society for Applied Research in Memory and Cognition, 23*(2), 151-161. https://doi.org/10.1002/acp.1463

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, *36*(3), 1-48. https://doi.org/10.18637/jss.v036.i03

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, *14*, 89 –103. https://doi.org/10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*(1), 3. https://doi.org/10.1037/lhb0000359

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review*, *114*(1), 152. https://doi.org/10.1037/0033-295x.114.1.152

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*(2), 262. https://doi.org/10.1037/a0035940

Wolf, O. T., Atsak, P., De Quervain, D. J., Roozendaal, B., & Wingenfeld, K. (2016). Stress and memory: a selective review on recent developments in the understanding of stress hormone effects on memory and their clinical relevance. *Journal of neuroendocrinology, 28*(8). https://doi.org/10.1111/jne.12353

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*, *18*(5), 459-482. https://doi.org/10.1002/cne.920180503

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1341. https://doi.org/10.1037/0278-7393.20.6.1341

Appendix A

UNIVERSITY OF CAPE TOWN

Department of Psychology

University of Cape Town Rondebosch 7701 South Africa
Telephone (021) 650 3417
Fax No. (021) 650 4104

26 September 2019

Milton Gering
Department of Psychology
University of Cape Town
Rondebosch 7701

Dear Milton

I am pleased to inform you that ethical clearance has been given by an Ethics Review Committee of the Faculty of Humanities for your study, *The effect of stress at encoding on witness recall memory and face identification.* The reference number is PSY2019-055.

I wish you all the best for your study.

Yours sincerely

University of Cape Town
PSYCHOLOGY DEPARTMENT
Upper Campus
Rondebosch

Lauren Wild (PhD)
Associate Professor
Chair: Ethics Review Committee

Appendix B

PRISMA-P (Preferred Reporting Items for Systematic review and Meta-Analysis Protocols) 2015 checklist: recommended items to

address in a systematic review protocol*

Section and topic Item No Checklist item

ADMINISTRATIVE INFORMATION

Title:

Identification 1a Identify the report as a protocol of a systematic review

Update 1b If the protocol is for an update of a previous systematic review, identify as such

Registration 2 If registered, provide the name of the registry (such as PROSPERO) and registration number

Authors:

Contact 3a Provide name, institutional affiliation, e-mail address of all protocol authors; provide physical mailing address of

corresponding author

Contributions 3b Describe contributions of protocol authors and identify the guarantor of the review

Amendments 4 If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes;

otherwise, state plan for documenting important protocol amendments

Support:

Sources 5a Indicate sources of financial or other support for the review

Sponsor 5b Provide name for the review funder and/or sponsor

Role of sponsor or funder 5c Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol

INTRODUCTION

Rationale 6 Describe the rationale for the review in the context of what is already known

Objectives 7 Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)

METHODS

Eligibility criteria 8 Specify the study characteristics (such as PICO, study design, setting, time frame) and report characteristics (such as years considered, language, publication status) to be used as criteria for eligibility for the review

Information sources 9 Describe all intended information sources (such as electronic databases, contact with study authors, trial registers or other grey literature sources) with planned dates of coverage

Search strategy 10 Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated

Study records:

Data management 11a Describe the mechanism(s) that will be used to manage records and data throughout the review

Selection process 11b State the process that will be used for selecting studies (such as two independent reviewers) through each phase of the review (that is, screening, eligibility and inclusion in meta-analysis)

Data collection process 11c Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators

Data items 12 List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications

Outcomes and prioritization 13 List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale

Risk of bias in individual studies 14 Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis

Data synthesis 15a Describe criteria under which study data will be quantitatively synthesised

15b If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data and methods of combining data from studies, including any planned exploration of consistency (such as I2, Kendall's τ)

15c Describe any proposed additional analyses (such as sensitivity or subgroup analyses, meta-regression)

15d If quantitative synthesis is not appropriate, describe the type of summary planned

Meta-bias(es) 16 Specify any planned assessment of meta-bias(es) (such as publication bias across studies, selective reporting within studies)

Confidence in cumulative evidence 17 Describe how the strength of the body of evidence will be assessed (such as GRADE)

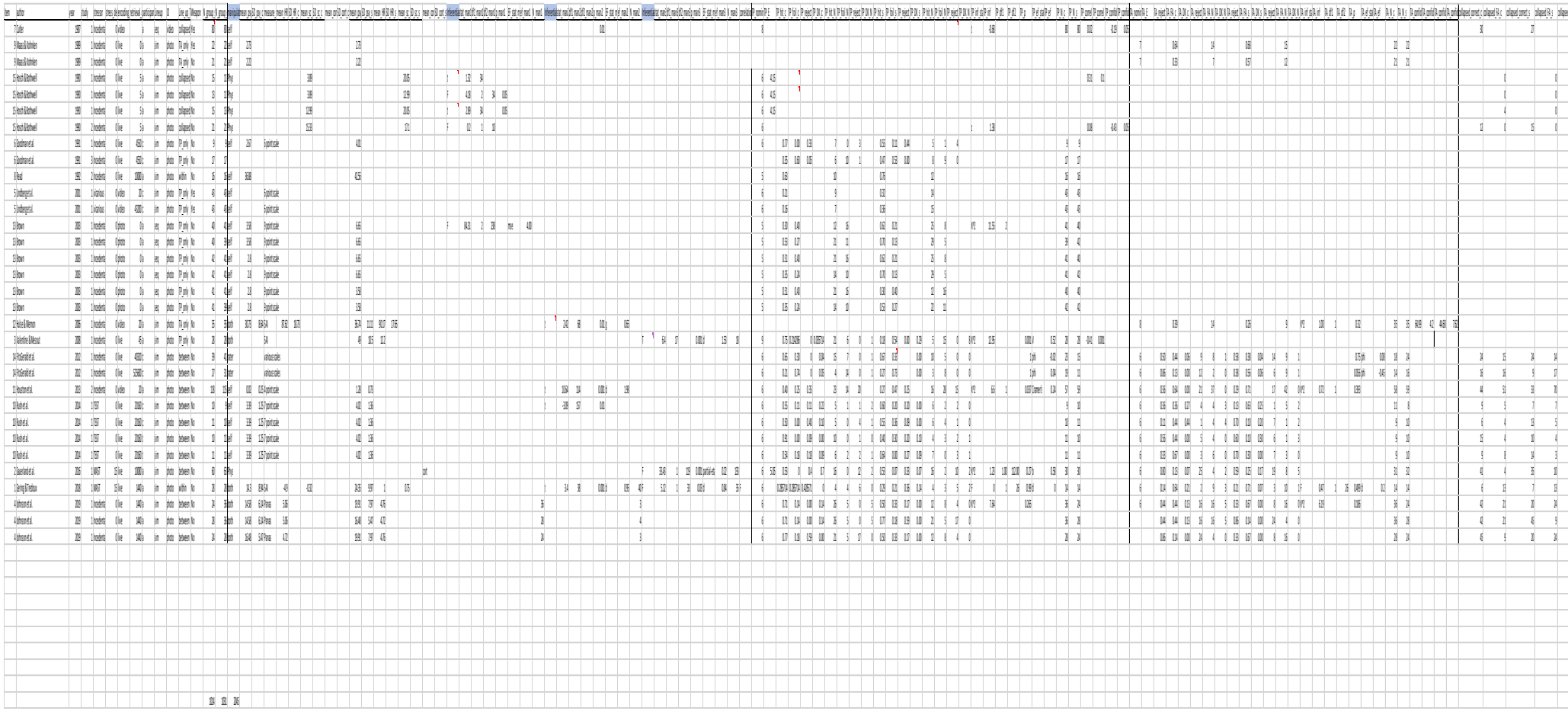
* It is strongly recommended that this checklist be read in conjunction with the PRISMA-P Explanation and Elaboration (cite when available) for important

clarification on the items. Amendments to a review protocol should be tracked and dated. The copyright for PRISMA-P (including checklist) is held by the PRISMA-P Group and is distributed under a Creative Commons Attribution Licence 4.0.

From: Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart L, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015 Jan 2;349(jan02 1):g7647.

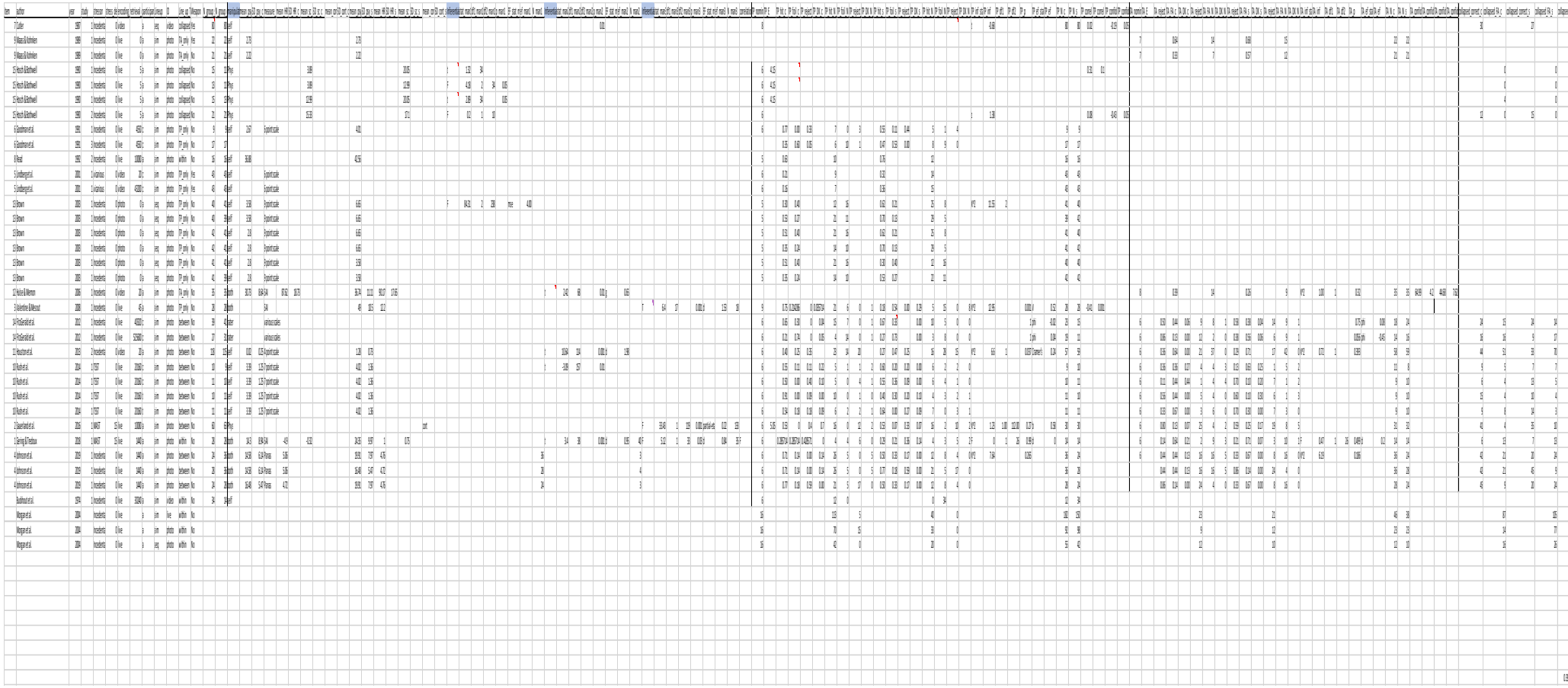
Appendix C

excel data file “Stress_meta_final.xlsx”



Appendix D

excel data file “Stress_meta_final_extra.xlsx”



Appendix E

R code

````
---
title: "A Meta Analysis of the Effect of Stress on Lineup Performance"
output: html_notebook
author: "Milton Gering"
---

```{r global_options, include=FALSE}
knitr::opts_chunk$set(fig.width=11, fig.height=4, out.height=3, fig.path='Figs/', fig.align = "center",
                      echo=FALSE, warning=FALSE, message=FALSE, error=FALSE)
```

```{r }
library(pacman)

#install.packages("bbmle")
p_load(tidyverse,knitr,foreign,
       tidyr,readxl,psych,janitor,ggplot2,dplyr,plyr,caret,jtools,haven,skimr,car,vcd,gridExtra,sjPlot,MASS,e1
       071,Metrics, PredPsych, boot, gbm, Hmisc, data.Table, lme4, merTools, lmerTest, nlme, semPlot,
       lavaan, lavaanPlot, DiagrammeR, metafor, stats4, bbmle)
```

```{r}
data<-read_xlsx("Stress_meta_final.xlsx")
data<-data %>%
  mutate(encoding_scenario = as.factor(encoding_scenario)) %>%
  mutate(stressor= as.factor(stressor)) %>%
  mutate(participant_age = as.factor(participant_age)) %>%
  mutate(Lineup = as.factor(Lineup)) %>%
  mutate(Weapon = as.factor(Weapon)) %>%
  mutate(Line_up_TA_TP = as.factor(Line_up_TA_TP))

data_sim<- dplyr::filter(data, Lineup=="sim")
```
````

TP hits

```{r}
hit_dat<-escalc(measure = "OR", ai = TP_hit_N_c, bi = TP_N_c - TP_hit_N_c, ci = TP_hit_N_s, di = TP_N_s - TP_hit_N_s, data = data)


hit_res<-rma(yi, vi, data = hit_dat)
hit_res


```

```{r}
forest.rma(hit_res, slab = paste(hit_dat$author, hit_dat$year, sep = ", "), xlim=c(-14, 10), ylim=c(-0.5,27))
text(-13, 26, "Author(s) and Year",    pos=4)
text( 10,26, "Observed LOR [95% CI]",  pos=2)
```

TP hits sim lineup only

```{r}
inf<- influence(hit_res)
plot(inf)
```

```{r}
data_hit_out<- dplyr::filter(data, TP_hit_N_c!="NA")


hit_dat_out<-escalc(measure = "OR", ai = TP_hit_N_c, bi = TP_N_c - TP_hit_N_c, ci = TP_hit_N_s, di = TP_N_s - TP_hit_N_s, data = data_hit_out)


hit_res_out<-rma(yi, vi, data = hit_dat_out)
inf<- influence(hit_res_out)
plot(inf)
```

```{r}
hit_dat_out = hit_dat_out[-c(12),]
hit_res_out<-rma(yi, vi, data = hit_dat_out)
hit_res_out
forest(hit_res_out)
```

```{r}
```

```
hit_dat_sim<-escalc(measure = "OR", ai = TP_hit_N_c, bi = TP_N_c - TP_hit_N_c, ci = TP_hit_N_s, di =
        TP_N_s - TP_hit_N_s, data = data_sim)


hit_res_sim<-rma(yi, vi, data = hit_dat_sim)

hit_res_sim
```

```
TP false alarms

```{r}

TP_FA_dat<-escalc(measure = "OR", ai = TP_foil_N_c, bi = TP_N_c - TP_foil_N_c, ci = TP_foil_N_s, di =
        TP_N_s - TP_foil_N_s, data = data)


TP_FA_res<-rma(yi, vi, data = TP_FA_dat)

TP_FA_res
```

```
```{r}

forest(TP_FA_res, slab = paste(TP_FA_dat$author, TP_FA_dat$year, sep = ", "), xlim=c(-16, 10), ylim=c(-
        0.5,24))

text(-15, 23, "Author(s) and Year",    pos=4)

text( 10,23, "Observed LOR [95% CI]",  pos=2)
```

```{r}

inf<- influence(TP_FA_res)

plot(inf)
```

```{r}

funnel(TP_FA_res)
```

TP rejections

```{r}

TP_reject_dat<-escalc(measure = "OR", ai= TP_reject_N_c, bi = TP_N_c - TP_reject_N_c, ci =
        TP_reject_N_s, di = TP_N_s - TP_reject_N_s, data = data)


TP_reject_res<-rma(yi, vi, data = TP_reject_dat)

TP_reject_res
```

```
```

```{r}
forest(TP_reject_res, slab = paste(TP_reject_dat$author, TP_reject_dat$year, sep = ", "), xlim=c(-25, 13),
       ylim=c(-2,18))
text(-25, 17, "Author(s) and Year",    pos=4)
text( 13,17, "Observed LOR [95% CI]",  pos=2)


```
```

```{r}
inf<- influence(TP_reject_res)
plot(inf)
```
```

```{r}
TP_reject_out = TP_reject_dat[-c(31),]


reject_res_out<-rma(yi, vi, data = TP_reject_out)
reject_res_out
forest(reject_res_out)
```


TP Don't know

```{r}
TP_DK_dat<-escalc(measure = "OR", ai = TP_DK_N_c, bi = TP_N_c - TP_DK_N_c, ci = TP_DK_N_s, di =
       TP_N_s - TP_DK_N_s, data = data)


TP_DK_res<-rma(yi, vi, data = TP_DK_dat)
TP_DK_res
```
```

```{r}
inf<- influence(TP_DK_res)
plot(inf)
```
```

```{r}
TP_DK_out = TP_DK_dat[-c(20),]
```

```
reject_DK_out<-rma(yi, vi, data = TP_DK_out)

reject_DK_out

forest(reject_DK_out)
```

```{r}
forest(TP_DK_res, slab = paste(TP_DK_dat$author, TP_DK_dat$year, sep = ", "), xlim=c(-25, 13), ylim=c(-2,15))

text(-25, 15, "Author(s) and Year",    pos=4)

text( 13,15, "Observed LOR [95% CI]",  pos=2)
```

```{r}
funnel(reject_DK_out)
```

TA rejections

```{r}
TA_reject_dat<-escalc(measure = "OR", ai = TA_reject_N_c, bi = TA_N_c - TA_reject_N_c, ci = TA_reject_N_s, di = TA_N_s - TA_reject_N_s, data = data)


TA_reject_res<-rma(yi, vi, data = TA_reject_dat)

TA_reject_res
```

```{r}
forest(TA_reject_res, slab = paste(TA_reject_dat$author, TA_reject_dat$year, sep = ", "), xlim=c(-25, 13), ylim=c(-2,15))

text(-25, 15, "Author(s) and Year",    pos=4)

text( 13,15, "Observed LOR [95% CI]",  pos=2)
```

```{r}
inf<- influence(TA_reject_res)

plot(inf)
```

```{r}
funnel(TA_reject_res)
```

TA False alarms

```{r}
```

```
TA_FA_dat<-escalc(measure = "OR", ai = TA_FA_N_c, bi = TA_N_c - TA_FA_N_c, ci = TA_FA_N_s, di =
        TA_N_s - TA_FA_N_s, data = data)



TA_FA_res<-rma(yi, vi, data = TA_FA_dat)

TA_FA_res
```

```{r}

forest(TA_FA_res, slab = paste(TA_FA_dat$author, TA_FA_dat$year, sep = ", "), xlim=c(-25, 13), ylim=c(-
        2,18))

text(-25, 18, "Author(s) and Year",    pos=4)

text( 13,18, "Observed LOR [95% CI]",  pos=2)
```

```{r}

inf<- influence(TA_FA_res)

plot(inf)
```

```{r}

funnel(TA_FA_res)
```

TA don't know

```{r}

TA_DK_dat<-escalc(measure = "OR", ai = TA_DK_N_c, bi = TA_N_c - TA_DK_N_c, ci = TA_DK_N_s, di =
        TA_N_s - TA_DK_N_s, data = data)


TA_DK_res<-rma(yi, vi, data = TA_DK_dat)

TA_DK_res
```

```{r}

forest(TA_DK_res, slab = paste(TA_DK_dat$author, TA_DK_dat$year, sep = ", "), xlim=c(-25, 13), ylim=c(-
        2,15))

text(-25, 15, "Author(s) and Year",    pos=4)
```

```
text( 13,15, "Observed LOR [95% CI]",  pos=2)
```

```
```{r}
inf<- influence(TA_DK_res)
plot(inf)
```
```{r}
TA_DK_out = TA_DK_dat[-c(27),]

TA_DK_out<-rma(yi, vi, data = TA_DK_out)
TA_DK_out
forest(TA_DK_out)
```
```{r}
funnel(TA_DK_res)
```
```{r}
residuals.rma(TA_DK_res)
```
collapsed

```{r}
collapsed_correct_dat<-escalc(measure = "OR", ai = collapsed_correct_c, bi = collapsed_N_c
 - collapsed_correct_c, ci = collapsed_correct_s, di = collapsed_N_s
 - collapsed_correct_s, data = data)
collapsed_correct_res<-rma(yi, vi, data = collapsed_correct_dat)
collapsed_correct_res

```
```{r}
forest(collapsed_correct_res, slab = paste(TA_FA_dat$author, TA_FA_dat$year, sep = ", "), xlim=c(-25, 13),
       ylim=c(-2,17))
text(-25, 17, "Author(s) and Year",    pos=4)
text( 13,17, "Observed LOR [95% CI]",  pos=2)
```
```

```
```

```{r}
inf<- influence(collapsed_correct_res)
plot(inf)
```

```{r}


funnel(collapsed_correct_res)
```

```{r}
collapsed_correct_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup,
        stressor, participant_age, Weapon, Line_up_TA_TP), data = collapsed_correct_dat)
collapsed_correct_res_mod
```

```{r}
collapsed_FA_dat<-escalc(measure = "OR", ai = collapsed_FA_c, bi = collapsed_N_c
 - collapsed_FA_c, ci = collapsed_FA_s, di = collapsed_N_s
 - collapsed_FA_s, data = data)


collapsed_FA_res<-rma(yi, vi, data = collapsed_FA_dat)
collapsed_FA_res
```

```{r}
forest(collapsed_FA_res, slab = paste(collapsed_FA_dat$author, collapsed_FA_dat$year, sep = ", "), xlim=c(-
        25, 15), ylim=c(-2,18.5))
text(-25, 18.5, "Author(s) and Year",    pos=4)
text( 15,18.5, "Observed LOR [95% CI]",  pos=2)


```

```{r}
inf<- influence(collapsed_FA_res)
plot(inf)
```

```{r}
funnel(collapsed_FA_res)
```

```{r}
collapsed_FA_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = collapsed_FA_dat)

collapsed_FA_res_mod
```

Moderator analysis:

```{r}
hit_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = hit_dat)

hit_res_mod
```

```{r}
hit_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age), data = hit_dat_out)

hit_res_mod
```

```{r}
forest(hit_res_mod, slab = paste(hit_dat_out$author, hit_dat_out$year, sep = ", "), xlim=c(-16, 10), ylim=c(-
        0.5,27))

text(-15, 27, "Author(s) and Year",    pos=4)

text( 10,27, "Observed OR [95% CI]",  pos=2)
```

```{r}
TP_FA_res_mod<-rma(yi, vi, mods = cbind(Lineup), data = TP_FA_dat)

TP_FA_res_mod
```

```{r}
TP_reject_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = TP_reject_out)

TP_reject_res_mod
```

```{r}
```

```
TP_DK_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = TP_DK_out)

TP_DK_res_mod

```

```{r}

TA_reject_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = TA_reject_dat)

TA_reject_res_mod

```

```{r}

TA_FA_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = TA_FA_dat)

TA_FA_res_mod

```

```{r}

TA_DK_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = TA_DK_dat)

TA_DK_res_mod

```,

D' prime

```{r}

 metatds <- function(nr = 100, ns = 100, pi_fa = 0.50, pi_a = 0.50)


   {
   #Variance d' Gourevitch & Galanter (1967)
   var_gg <- ((pi_fa*(1-pi_fa))/(nr*dnorm(qnorm(pi_fa))^2)) + ((pi_a*(1-pi_a))/(ns*dnorm(qnorm(pi_a))^2))


   #Variance d' and Expected Value d' Miller (1996)
   fre_fa <- c(0.5,(1:(nr-1)), nr-0.5)
   fre_a <- c(0.5,(1:(ns-1)), ns-0.5)
   prop_fa <- fre_fa/nr
   prop_a <- fre_a/ns
   z_fa <- qnorm(prop_fa, mean = 0, sd = 1)
   z_a <- qnorm(prop_a, mean = 0, sd = 1)
   prob_fa <- dbinom(0:nr,nr,pi_fa)
   prob_a <- dbinom(0:ns,ns,pi_a)
```

```
  v_esp_zfa <- sum(z_fa*prob_fa)
  v_esp_za <- sum(z_a*prob_a)
  v_esp_miller <- v_esp_za - v_esp_zfa
  var_zfa <- sum((z_fa*z_fa)*prob_fa)-(v_esp_zfa*v_esp_zfa)
  var_za <- sum((z_a*z_a)*prob_a)-(v_esp_za*v_esp_za)
  var_miller <- var_za + var_zfa
  mestim <- list(Val_Esp  = v_esp_miller,Varianza=var_miller)
   esti <- list(Var_GG  =var_gg, Miller=mestim)
   return(esti)
 }
#metatds(100,100,.2,.7)
metatds(37,27,.73,.4)
```

```{r}
 data_tp_d <-dplyr::filter(data,is.na(d_prime_TP_c)==FALSE)
data_ta_d<-dplyr::filter(data,is.na(d_prime_TA_c)==FALSE)
#data<-dplyr::filter(data,is.na(TA_N_c)==FALSE)
#data<-dplyr::filter(data,is.na(TP_N_c)==FALSE)
  x<- metatds(data_tp_d$d_TP_N_c,data_tp_d$d_TP_N_c, data_tp_d$TP_foil_c+0.001,
           data_tp_d$TP_hit_c+0.001)

  y <- metatds(data_tp_d$d_TP_N_s,data_tp_d$d_TP_N_s, data_tp_d$TP_foil_s+0.001,
           data_tp_d$TP_hit_s+0.001)

  data_tp_d<-data_tp_d %>%
  mutate(d_prime_var_TP_c = x$Var_GG) %>%
  mutate(d_prime_var_TP_s = y$Var_GG)

  w <- metatds(data_ta_d$d_TA_N_c,data_ta_d$d_TA_N_c, data_ta_d$TA_FA_c+0.001,
           data_ta_d$TA_reject_c+0.001)
  z <- metatds(data_ta_d$d_TA_N_s,data_ta_d$d_TA_N_s, data_ta_d$TA_FA_s+0.001,
           data_ta_d$TA_reject_s+0.001)

   data_ta_d<-data_ta_d %>%
   mutate(d_prime_var_TA_c = w$Var_GG)%>%
   mutate(d_prime_var_TA_s= z$Var_GG)
```

```
```

```{r}
d_TP_dat<-escalc(measure = "MD", m1i = d_prime_TP_c, m2i = d_prime_TP_s , sd1i =
        sqrt(d_prime_var_TP_c), sd2i = sqrt(d_prime_var_TP_s) ,n1i = TP_N_c, n2i=TP_N_s, data =
        data_tp_d)


d_TP_res<-rma(yi, vi, data = d_TP_dat)
d_TP_res


```

```{r}
forest(d_TP_res, slab = paste(d_TP_dat$author, d_TP_dat$year, sep = ", "), xlim=c(-16, 10), ylim=c(-1.5,24))
text(-15, 23, "Author(s) and Year", pos=4)
text( 10,23, "Observed MD [95% CI]", pos=2)
```

```{r}
d_TP_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = d_TP_dat)
d_TP_res_mod
```


```{r}
inf<- influence(d_TP_res)
plot(inf)
```

```{r}
funnel(d_TP_res)
```

```{r}
d_TA_dat<-escalc(measure = "MD", m1i = d_prime_TA_c, m2i = d_prime_TA_s , sd1i =
        sqrt(d_prime_var_TA_c), sd2i = sqrt(d_prime_var_TA_s) ,n1i = TA_N_c, n2i=TA_N_s,data =
        data_ta_d)


d_TA_res<-rma(yi, vi, data = d_TA_dat)
d_TA_res
```

```{r}
```

```
forest(d_TA_res, slab = paste(d_TA_dat$author, d_TA_dat$year, sep = ", "), xlim=c(-16, 13), ylim=c(-
        1.5,15.5))

text(-15, 15.5, "Author(s) and Year",    pos=4)

text( 10,15.5, "Observed LOR [95% CI]",  pos=2)
```

```{r}
d_TA_res_mod<-rma(yi, vi, mods = cbind(encoding_scenario,retrieval_delay_mins,Lineup, stressor,
        participant_age, Weapon, Line_up_TA_TP), data = d_TA_dat)

d_TA_res_mod
```

```{r}
inf<- influence(d_TA_res)

plot(inf)
```

```{r}
funnel(d_TA_res)
```

```{r}
scale_data<-read_xlsx("stress_scale.xlsx")

scale_data <- scale_data%>%

  mutate(author = as.factor(author))
```

```{r}
model <- lmer(TP_hit ~ poly(Stress_Rating,2) + (1|author), data = scale_data)

summ(model)

summary(model)
```

```{r}
plot(model)
```

```{r}
linear <-lm(TP_hit~  Condition+ poly(Stress_Rating,2) , data = scale_data)

summ(linear)
```

```{r}
mode2 <- lmer(TP_foil ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
```

```
summ(mode2)
summary(mode2)
```

```{r}
mode3 <- lmer(TP_reject ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode3)
```

```{r}
mode4 <- lmer(TP_DK  ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode4)
```

```{r}
mode5 <- lmer(TA_reject ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode5)
```

```{r}
mode3 <- lmer(TA_FA  ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode3)
```

```{r}
mode3 <- lmer(d_TP ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode3)
```

```{r}
mode3 <- lmer(d_TA ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode3)
```

```{r}
mode3 <- lmer(collapsed_hit ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode3)
```

```{r}
mode3 <- lmer(collapsed_FA ~ poly(Stress_Rating,2) + (1|author), data = scale_data)
summ(mode3)
```

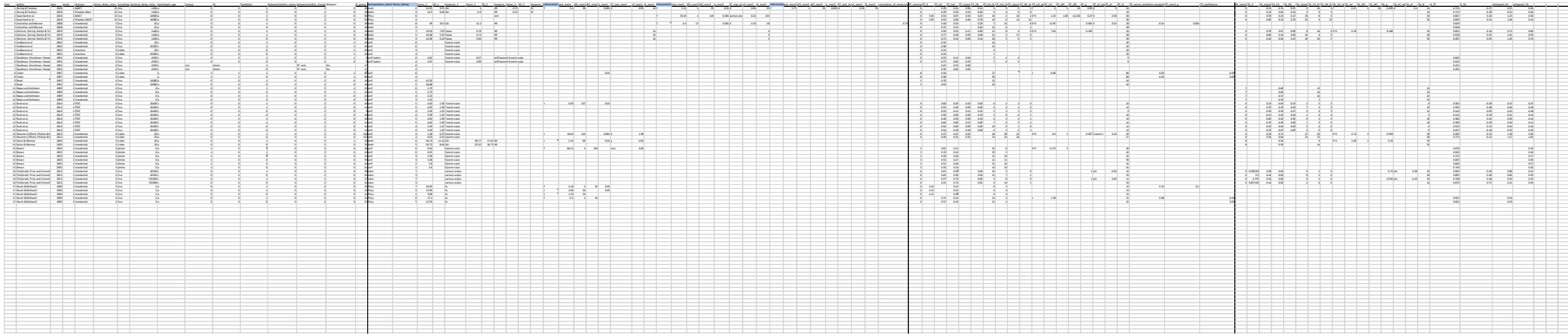
Appendix F

Stress scale coding sheet

1. Code each event on a scale from 1-9
2. As the brown Study uses a scale of 1-9, use rounded down values as a starting point.
3. Where a significant difference with cohens d = 1 between groups, difference of 2 can be used on the scale.
4. Where a significant difference with cohens d < 1 between groups, difference of 1 can be used on the scale.
5. Where a significant difference with cohens d > 1.5 between groups, difference of 3 can be used on the scale.
6. Procedures using the same stressor (eg. MAST) should be coded as equally stressful
7. A live event should be more stressful than the video equivalent.
8. Personal interaction with the perpetrator/target will be more stressful
9. Seeing a weapon will increase stress

Appendix G

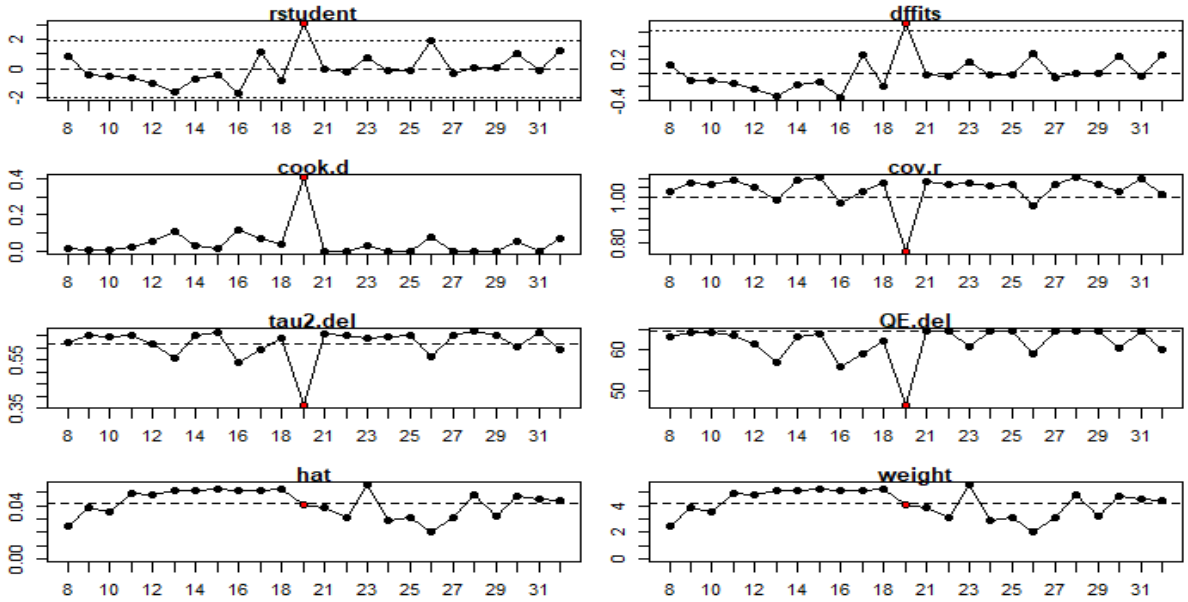
excel data file “Stress_scale.xlsx”



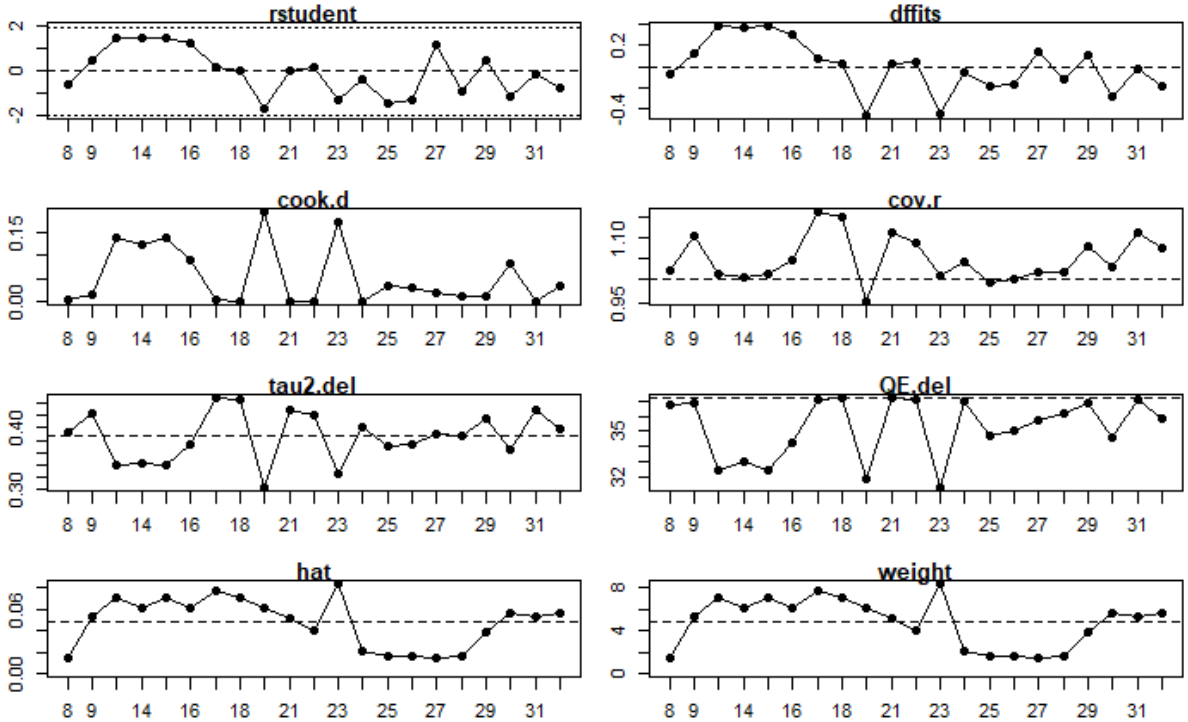
Appendix H:

Influence plots

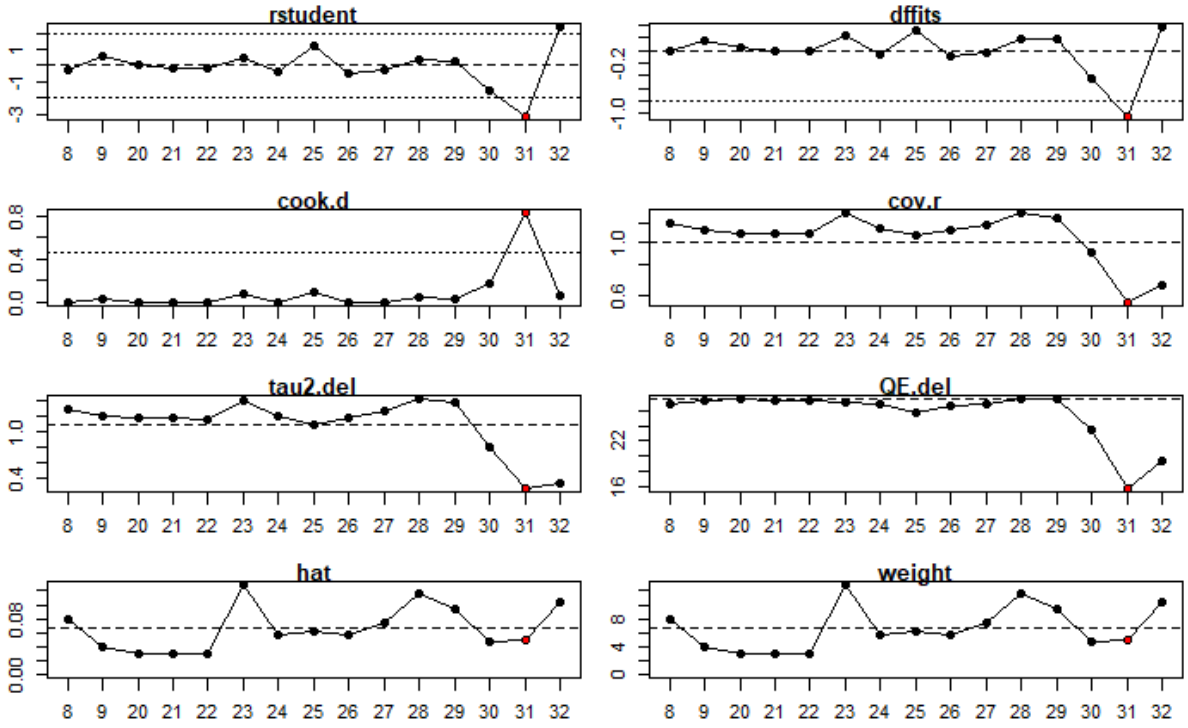
TP hits:



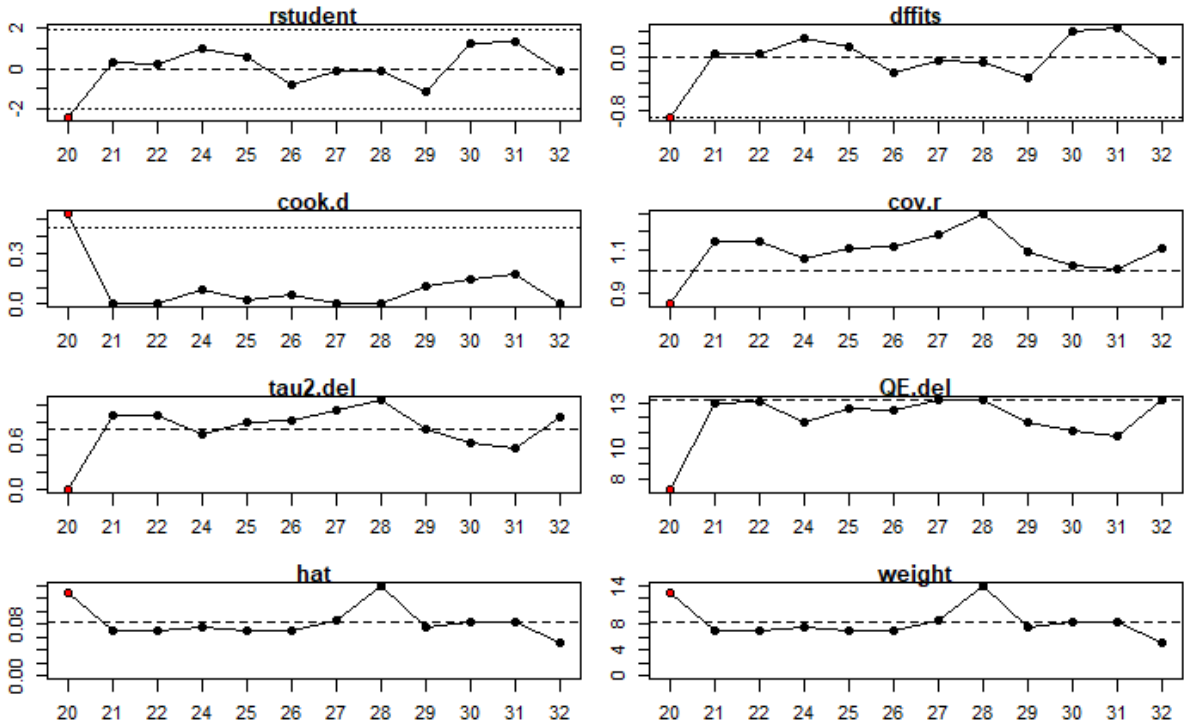
TP foil IDs



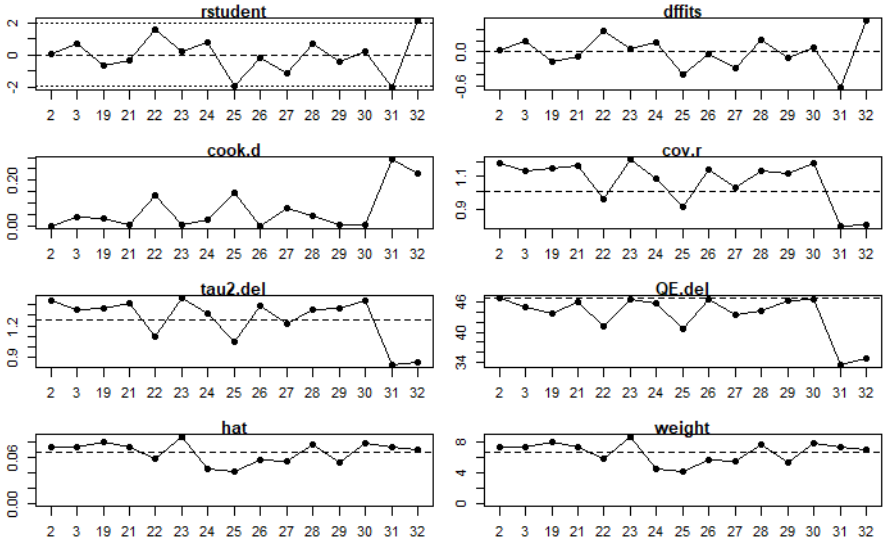
TP rejections



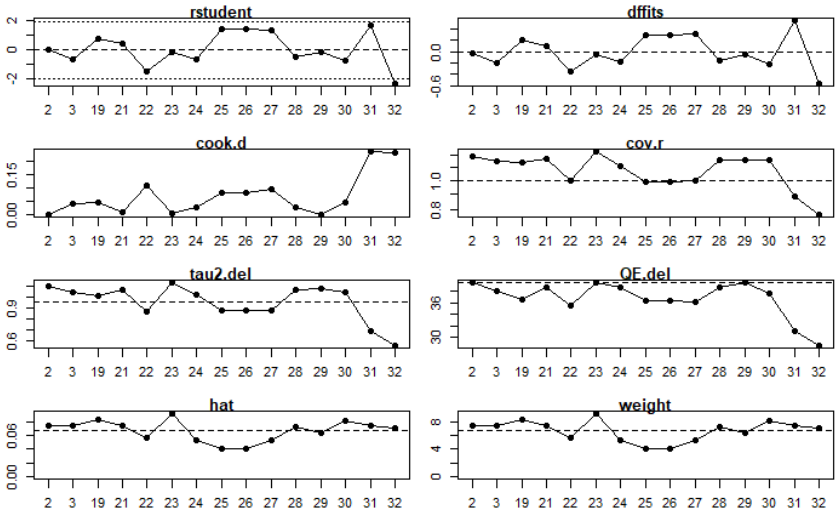
TP DK



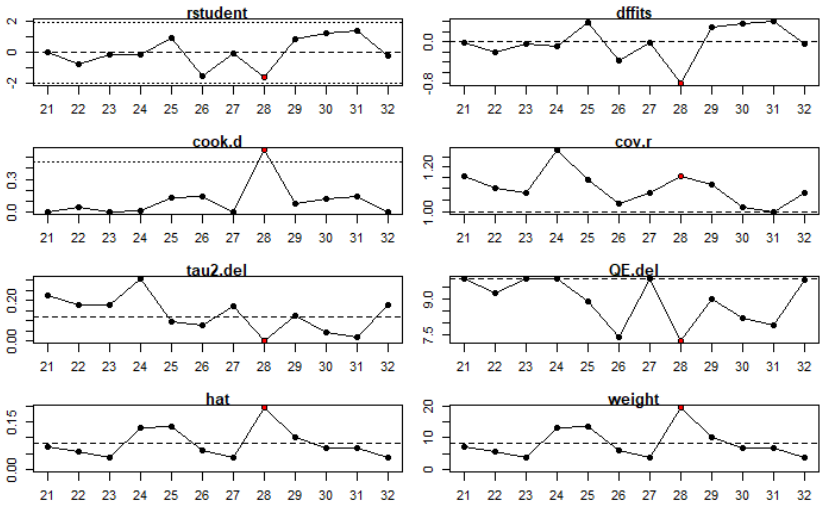
TA Rejections



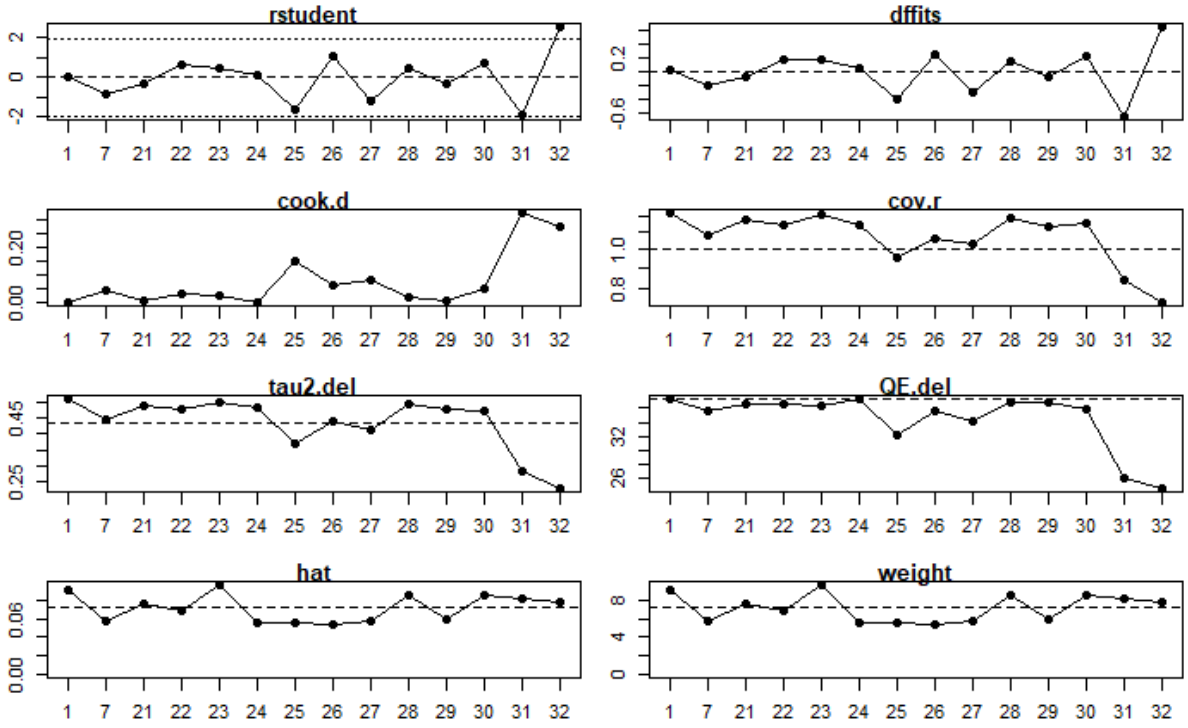
TA false alarms



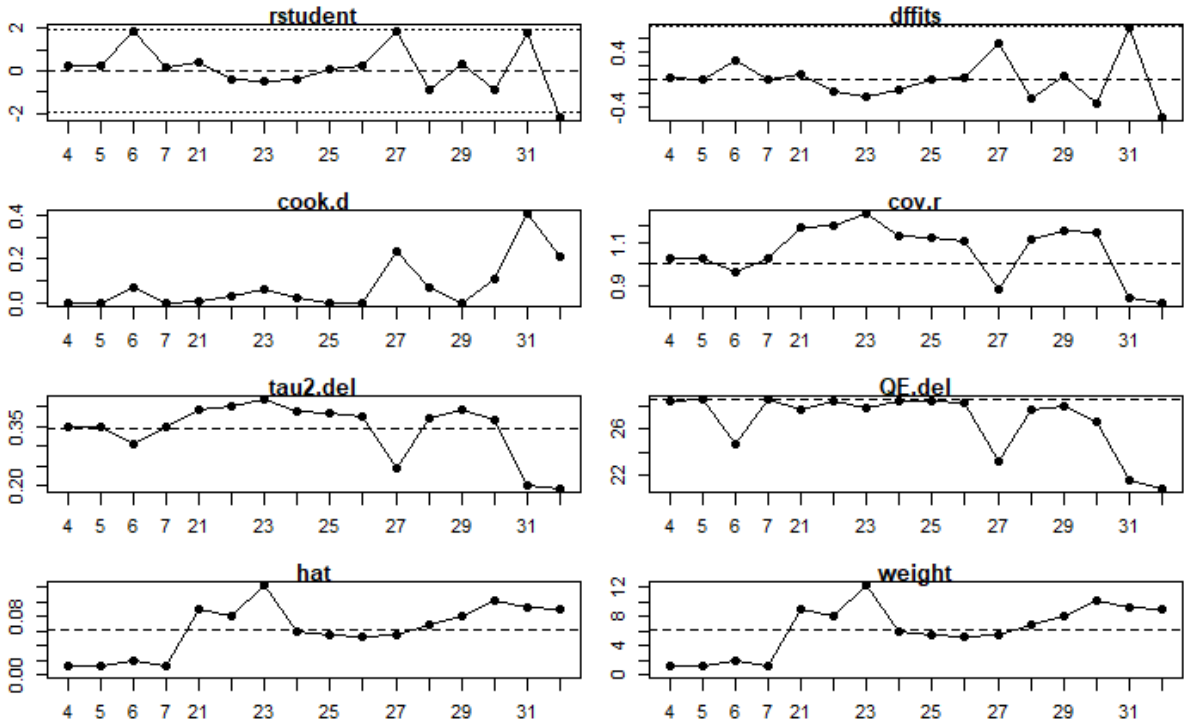
TA DK



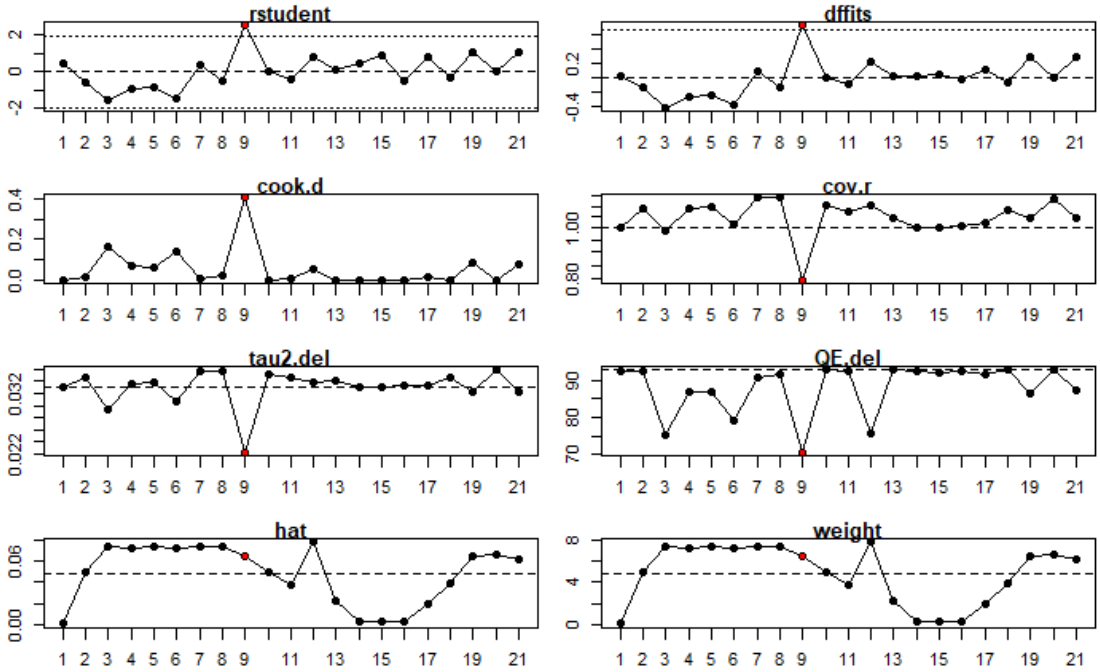
Collapsed Correct



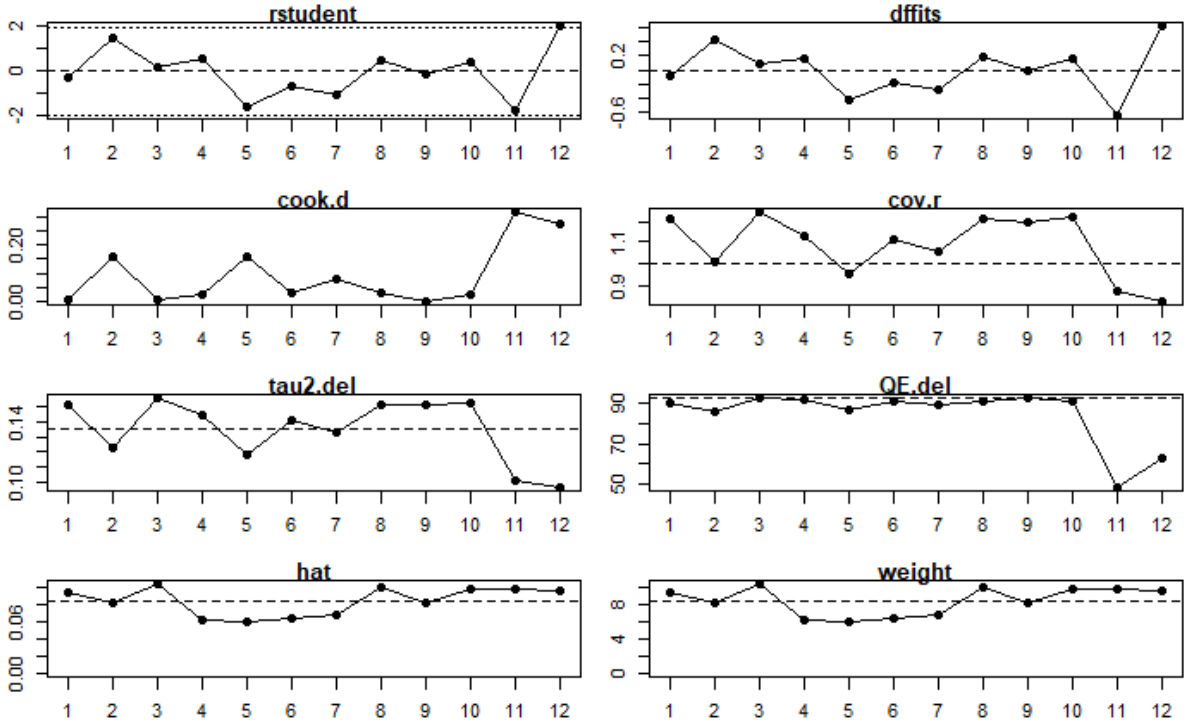
Collapsed false alarms



TP d’

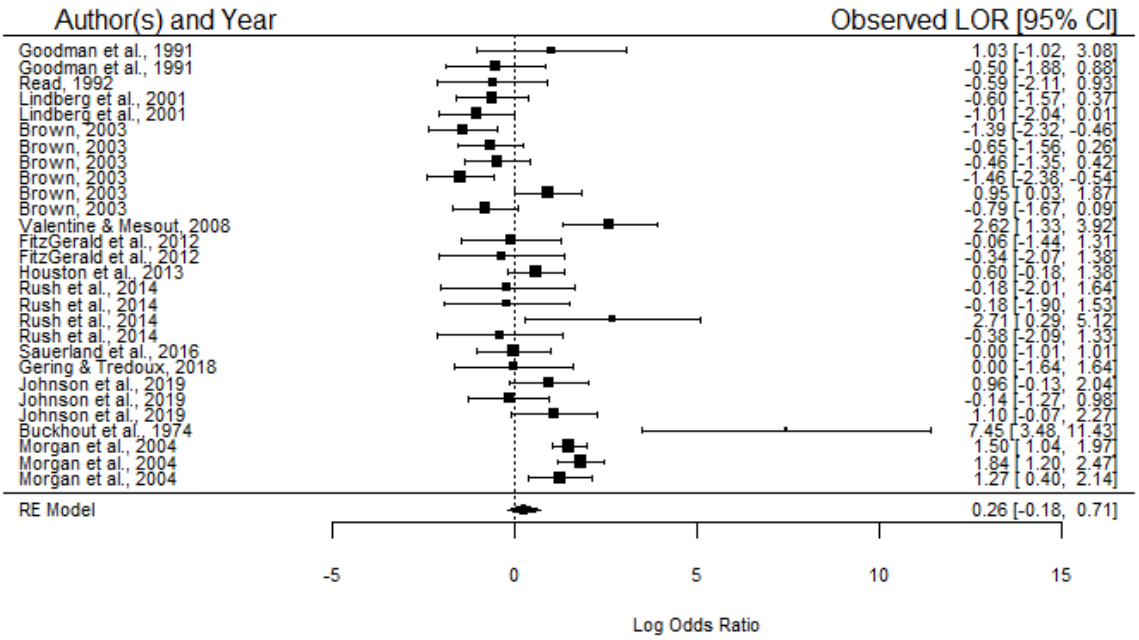

rstudent
dffits
cook.d
cov.r
tau2.del
QE.del
hat
weight

TA d’


rstudent
dffits
cook.d
cov.r
tau2.del
QE.del
hat
weight

Appendix I

Forrest plot including Morgen et al. (2004) and Buckhout et al (1972) studies



Note: only the forrest plot for hits is displayed. All other results can be generated by running the R code in appendix E using the data file in appendix D and changing the name of the data file read in from "Stress_meta_final.xlsx" to "Stress_meta_final_extra.xlsx"