MEMORY DISTORTION VIA IMAGINATION: NEURAL CORRELATES AND FORENSIC APPLICATIONS

Phot Dhammapeera

A thesis submitted for the degree of Doctor of Philosophy

at the University of Kent, Canterbury

School of Psychology

University of Kent

February 2019

Declaration

I, Phot Dhammapeera, declare that the work presented in this thesis is my own. The work presented is original completed under the supervision of Dr Zara M Bergström. I have not been awarded a degree by submitting the work included in this thesis for higher degree at any other institution.

Acknowledgements

This thesis would not be possible without support from people around me. Firstly, I would like to say thank you my supervisor, Dr Zara M Bergström, for introducing me to EEG research and supporting me throughout my PhD. I am grateful for guidance and encouragement. Without her tremendous support and dedication, I cannot imagine how my PhD life would be.

I would like to thank all my friends and colleagues who have supported me for advice and encouragement during my PhD. I also would like to thank all participants who participated in my studies. Their participation means so much to me.

Most importantly, I would like to say thank you to all my family members, especially from my parents and brother, for their great support from 5,874 miles away. Because their love, faith, and encouragement, I can be who I am today. Thank you for believing in me. Last but not least, thank you to my partner, Nash, for always be right next to me through good and bad times.

Abstract

Our memory is vulnerable to changes so that the way we remember past events can become distorted over time. One way that memory distortions may occur is if we rehearse and imagine an alternative scenario to what really happened. Such counterfactual imagination may distort the true memory and create a false memory of the imagined event. This has crucial implications, especially in forensic settings because guilty suspects may adopt this technique as a countermeasure, in an attempt to evade blame. The research presented in this thesis investigated the effect of counterfactual imagination on memory detection tests—the Autobiographic Implicit Association Test (aIAT) and the Concealed Information Test (CIT)—using both behavioural measures and EEG methods. It also investigated the neurocognitive mechanisms underlying counterfactual imagination effects on memory, and whether counterfactual imagination actually impairs the true memory of the original event. Results from five experiments supported the view that counterfactual imagination can distort our memory, leading to significant effects on forensic memory detection in some circumstances and with some types of tests. Specifically, it was found that the aIAT is very susceptible to the effects of counterfactual imagination, while the CIT is more resistant to this countermeasure. Furthermore, I found that counterfactual imagination impaired both recall and recognition of true actions, and I describe novel EEG effects that were associated with counterfactual imagination and subsequent memory impairments, thus providing new evidence of the neurocognitive mechanisms that underlie counterfactual imagination effects on memory. Altogether, my research makes an original contribution to improve our understanding of counterfactual imagination and memory distortion and suggests that forensic memory detection tests should be used with caution.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

In our daily life, we encounter different events each day. We remember some of these events while we forget some others. For the events that we can remember, we sometimes accurately remember very precise details of the event, while other event memories are distorted so that we misremember details. There are various ways that memory can be distorted. According to Schacter (1999), there are three key sins of memory distortion: misattribution, suggestibility, and bias. Misattribution refers to when information is retrieved to be from an incorrect source. For example, we may mistakenly remember that an imagined event was true. Suggestibility refers to when we incorporate information that was given from external sources, for example, when someone else gives us misleading information about past events. Bias refers to when memory is distorted as a result of pre-existing knowledge, beliefs, and feelings shaping our recollection of past experience (Schacter, 1999).

Memory distortion has long been an interesting topic for cognitive psychologists to try to understand its causes and consequences. More recently, the topic of memory distortion has been addressed by cognitive neuroscientists, who have investigated the neural processes underlying memory distortion and the differences between true and false memories (Addis, Wong, & Schacter, 2007; Schacter & Slotnick, 2004). These researchers believe that memory distortion is an adaptive cognitive process (Schacter, 2012; Schacter, Guerin, & St. Jacques, 2011), rather than just showing the flaws of memory (Clancy, Schacter, McNally, & Pitman, 2000). They have argued that memory distortion is a consequence of the efficiency and flexibility of memory.

The focus of this thesis is on memory distortion that occurs as a result of incorporating false information into memory after an event has occurred. However,

my research is concerned not only with suggestibility from external sources, but also with how our own internal processing of false information may contribute to memory distortions. I investigated how counterfactual imagination of a false version of a past event influence memory for that false information, and memory for what truly happened in the original event. To address these questions, I used a combination of behavioural and cognitive neuroscience techniques, specifically Event-Related Potentials (ERPs). I investigated these issues in both applied forensic contexts and in terms of theories of underlying cognitive and neural mechanisms. In the first chapter of the thesis, I review relevant literature on how memory distortion may occur as a result of counterfactual imagination and how researchers can use Event-Related Potentials to investigate the neurocognitive mechanisms of memory. I also review literature on how memory distortion as a result of counterfactual imagination may affect forensic applications, and also provide an overview what will be covered in this thesis.

1.1. Memory Distortion and Imagination

It has long been established that imagination can have distorting effects on memory. When people imagine a novel event, they may later believe that the event actually occurred in the past, which in fact it did not (Goff & Roediger, 1998; Marsh, Pezdek, & Lam, 2014). This phenomenon is known as imagination inflation. It has been suggested that memory distortion related to imagination inflation is due to memories for imagined events sharing similar characteristics with memories for actual perceived events, which can lead to 'reality monitoring errors' so that participants are not able to successfully apply retrieval monitoring processes to distinguish between true and imagined versions of the past (de Brigard, 2017;

Mitchell & Johnson, 2009). Consistent with this view, neuroimaging research has found extensive overlap in neural processes between true and false memories (Addis, 2018; Addis et al., 2007) Thus, people may sometime misattribute memories of imagined events as having been perceived because imagined events share common features and are too similar to memories of real experiences.

One example of when imagination inflation may occur is as a result of counterfactual thinking. Counterfactual thinking refers to when people mentally simulate (i.e. imagine) an alternative version of reality, which could involve imagination of an alternative outcome of a past event (episodic counterfactual thinking), that could be emotionally neutral, positive (upward counterfactual thinking) or negative (downward counterfactual thinking). Previous research has found that imagining a counterfactual event can distort memory and cause people to falsely remember that the counterfactual event was true (Gerlach, Dornblaser, & Schacter, 2014; Petrocelli & Crysel, 2009; Petrocelli & Harris, 2011). For example, De Brigard and colleagues (de Brigard, Szpunar, & Schacter, 2013) found that episodic counterfactual thinking of a childhood event that would have been possible to occur can lead participants to remember that it did occur, showing that counterfactual thinking can affect our memory. In other research, it was found that counterfactual thinking not only has an effect on our memory, but also on learning and decision making. Petrocelli, Rubin, and Stevens (Petrocelli, Rubin, & Stevens, 2016) suggested that when people think counterfactually about an event involving gambling (i.e. upward counterfactual thinking), they are more likely to overestimate their wins compared to losses, leading to overconfidence in their judgements when betting and gambling. As a consequence, counterfactual thinking distorted memory of their actual wins, and they were more likely to invest more.

One explanation for how counterfactual imagination affects memory involves interference between competing memories, coupled with source monitoring errors. According to interference theory (specifically retroactive interference), learning new material after encoding can lead the newly encoded memory to become stronger than the old memory, so that the new memory competes with the old memory and blocks access to it when it is cued with a shared retrieval cue (for review Anderson & Neely, 1996; Camp, Pecher, & Schmidt, 2007). In a counterfactual thinking situation, people may have an original memory of how an event really happened, but when that memory is cued, the more recently encoded counterfactual version of that event might be retrieved instead. Previous research suggested that newly acquired memories often have a stronger association with the shared cue than old memories, and can interfere with and block access to the old memory during competition. As a consequence, the new memory is highly accessible, and the old memory is less recallable. This theoretical account thus suggests that the true memory is still intact and stored in memory, but not possible to access due to interference from the new memory.

This interference account described above might also involve source monitoring errors, whereby people fail to adequately detect if a memory stems from perception or imagination. In some cases, people may be able to distinguish truly perceived from imagined events by using reality monitoring processes, whereas other times they may fail to do so because memory for the imagined event is too similar to a perceived event (Lyle & Johnson, 2006; Mitchell & Johnson, 2009). Research suggests that repeatedly retrieving an imagined event can cause people to believe it is a perceived memory (Suengas & Johnson, 1988). Repeated retrieval can eventually increase the accessibility of the imagined event and increase confusion

between a true and imagined memory (Jacques, Szpunar, & Schacter, 2017; Johnson, 1997). Thus, a combination of interference between competing true and false memories coupled with reality monitoring errors might cause memories of counterfactual imagination to be mistaken for true memories.

An alternative theoretical explanation for how counterfactual imagination affects memory stems from research on the retrieval-induced forgetting (RIF) phenomenon. RIF refers to the finding that repeatedly retrieving target information can eventually inhibit non-target information that is associated to the same retrieval cue (Anderson, 2003; Anderson, Bjork, & Bjork, 1994). In the classic studies of RIF, participants were asked to learn list of categories-exemplar pairs (e.g. fruits-orange, fruits-banana, drinks-scotch, drinks-water). Then, they performed retrieval practice on half of the exemplars (i.e. orange, banana, scotch, or water) of half of previously learned categories (i.e. fruits and drinks) by using cued stem-recall tests (Anderson, 2003; Anderson et al., 2004; Levy & Anderson, 2002). For example, they were asked to retrieve "orange" when given "fruit-or___?" as a cue, and this process was repeated three times for each retrieval cue. After that, participants were given a cued recall test for all categories-exemplar pairs (e.g. fruits-___?) assessing performance for each stimulus type: practiced exemplars (e.g. orange), unpractised exemplars from the same category (e.g. banana), and a baseline condition composed of unpractised exemplars from an unpractised category (e.g. scotch) that was assessing simple forgetting over time as a result of lack of practice. It was found that retrieval practice enhanced recall performance of practiced exemplars when compared to baseline exemplars, but unpractised exemplars from practiced categories were impaired compared to baseline, demonstrating a *below baseline impairment* that is therefore not just due to a lack of rehearsal, but must be due to an additional process

that impairs recall further. A large body of evidence suggests that this additional process is *inhibition* of the memories of unpractised exemplars from practiced categories that causes those memories to become inaccessible (Anderson, 2003; Anderson et al., 1994; Storm & Levy, 2012)

Research on RIF has thus built on interference theory by adding an additional mechanism of inhibition, arguing that inhibition is sometimes necessary to facilitate selective retrieval of some memories in the face of interference from other memories (Anderson, 2003; Levy & Anderson, 2002). For instance, when participants practice retrieval of some exemplars from a category, this can create interference between target exemplars and other non-target exemplars that are still activated by the shared retrieval cue (the category). Accordingly, inhibition therefore overrides retrieval of non-target exemplar memories in order to facilitate retrieval of the target exemplar (Anderson et al., 2004). This inhibition process impairs the memory of unpractised exemplars from practiced categories (Hellerstedt & Johansson, 2013), but does not impair unpractised exemplars from the baseline condition, because those categories were not shown in the retrieval practice phase and therefore did not elicit interference between competing exemplar memories. Applying this theory to counterfactual imagination effects on memory, it is possible that repeatedly thinking counterfactually can both produce a false memory that the imagined event has occurred, while also inhibiting the true memory of what really happened. Thus, when repeatedly imagining a counterfactual event, the original memory of the event may be weakened and forgotten. However to my knowledge, research on counterfactual imagination effects of memory has not addressed whether these effects are due to inhibition or interference, or both, as investigated in this thesis.

1.2. Using ERPs to Investigate Episodic Memory

Apart from behavioural measures, measures of brain activity like functional magnetic resonance imaging (fMRI) and Event-related potentials (ERP) measures can also be used to investigate effects of counterfactual imagination on memory. Evidence from fMRI has suggested that retrieving a memory of the past shares overlapping neural mechanisms with imagining a counterfactual event, since activity in left posterior inferior parietal and ventrolateral frontal cortices was similar during autobiographical memory retrieval and counterfactual imagination (Jacques, Carpenter, Szpunar, & Schacter, 2018). fMRI research has also revealed that distinguishing between memories of performed and imagined actions (reality monitoring) involves brain activation in the prefrontal cortex and motor regions (Brandt et al., 2013). However, this thesis is focused on ERPs as a method to investigate the effect of counterfactual imagination in episodic memory, which to my knowledge has not been addressed previously. The ERP method is a powerful technique that is extensively used in the literature to investigate the neurocognitive processes that underlie memory (Luck, 2014). Unlike fMRI that measures oxygenation levels of blood flow (i.e. a haemodynamic technique), ERPs refers to waveforms of electrical activity voltages that are elicited by some experimental event. ERPs are extracted by averaging together multiple trials of recorded continuous electroencephalogram (EEG) that is time-locked to some events, such as exposure to experimental stimuli or responses to stimuli. Then, these waves are used to examine brain activity changes over time and differences between experimental conditions with regards to ERP components—that are ERP modulations that are defined in terms of their amplitude, timing, and scalp location.

The ERP method is non-invasive and relatively less expensive than other

neuroimaging techniques such as fMRI and magnetoencephalography (MEG). Although ERPs have low spatial resolution such that it is not possible to measure precise neuroanatomical information with ERPs, they have high temporal resolution such that it can monitor and record small changes in neural activity in terms of milliseconds. ERPs therefore allow researchers to examine real-time changes in brain activity as soon as the stimulus onsets with randomized trial orders (Voss & Paller, 2017). Furthermore, compared to fMRI, ERPs are more direct measures of neural activity. The interpretation of fMRI can be complicated due to oxygenation changes associated with vasculature, especially if the study requires comparing between groups of participants. Thus, ERPs is a suitable method to use when investigating changes between groups in neural correlates of cognition that are related to particular events.

ERPs has been used to investigate various processes such as emotion, language, and visual sensory responses, including memory. Early ERP components in the first few hundred milliseconds after stimulus presentation are thought to indicate perceptual processes elicited by the stimuli, such as for example, the N1 that reflects visual processing. After a few hundred milliseconds, ERPs start to index activation of information in memory, such as the N400 that reflects activation of semantic information (Kutas & Federmeier, 2000), or processes involved in stimulus evaluation or categorisation, such as the P300 (see Luck, 2014). Later ERP effects are often interpreted as related to decision making and response monitoring, such as the late posterior negative component (LPN) that has been related to response monitoring or evaluation of retrieved information in memory retrieval tasks (Mecklinger, Rosburg & Johansson, 2016). Relevant to the current thesis, ERPs enables us to examine the neurocognitive processes that underlie different responses

in memory, and we can use ERPs to detect how much people can remember (Bergstrom, Velmans, de Fockert, & Richardson-Klavehn, 2007). ERP studies in memory research are typically designed to investigate the electrophysiological correlates of memory-relevant cognitive processes, rather than simply studying the characteristics of a specific ERP component. In episodic memory paradigms, it is often recorded during study phases to investigate encoding and during test phases to investigate retrieval (reviewed in Voss & Paller, 2017).

When investigating encoding processes, ERPs are typically recorded during study phases where experimental stimuli are first presented. ERPs can then be separately averaged based on later performance in a subsequent test, to examine brain activity during encoding that predicts whether or not participants can later remember the stimuli. Any resulting differences in brain activity between those ERP conditions is often referred to as Dm (difference due to memory) effects (Paller, 1990; Paller, Kutas, & Mayes, 1987), or are also known as subsequent memory effects (Paller & Wagner, 2002). Such differences between the average ERP for later remembered vs. forgotten trials are used to investigate processes relating to successful encoding, such as when in time those processes are active, across which scalp locations, and whether they differ dependent on the type of subsequent memory experience (for example, whether participants will later recollect information or will only recognise stimuli as familiar, without recollection of context, see Paller & Wagner, 2002). Subsequent memory effects also vary depending on the type of later test, for example they are typically found to be more robust when memory is later tested with recall rather than recognition tests. In ERPs, subsequent memory effects are often expressed as positive potentials that are maximal at parietal sites during 400-800ms after a stimulus is presented, that are

larger for later remembered than later forgotten stimuli (Gonsalves et al., 2004; Voss & Paller, 2017). This positivity has been related to the P300 or late-positive complex ERPs component, which may index attentional processes (Fabiani, Karis, & Donchin, 1986). However, other subsequent memory effects have also been observed, including effects that onset even before a stimulus has been presented, thus potentially indexing preparatory processes that facilitate encoding (e.g. Otten, Quayle, Akram, Ditewig, & Rugg, 2006).

ERPs can also be recorded during memory tests to investigate retrieval processes that underlie memory decisions, such as the activation of a memory trace and the decision related processes that people engage before giving a response in a test. In recognition tests, researchers often study old-new ERP effects that comprise the difference between old (previously studied) and new (previously non-studied) stimuli. ERP research has revealed that old-new effects are typically expressed as increased ERP positivities for old compared to new items across frontal, central and parietal electrode sites, and comprise multiple sub-components that relate to familiarity, recollection and other decision-related processes such as post-retrieval monitoring. Familiarity (the sense that a stimulus is familiar without remembering any context) is typically associated with a frontal and central positivity around 300-500ms post-stimulus that modulates the N400 deflection, and is therefore referred to as the "FN400" effect. It can be difficult to distinguish between the FN400 and N400, and there is ongoing debate regarding how these effects relate to each other (see Voss & Paller, 2017). Nevertheless, many researchers argue that the FN400 component is distinct from N400 potentials related to semantic memory, and that FN400 is instead related to episodic familiarity (Bridger et al., 2012). Recollection (remembering the context where a stimulus was encountered) is associated with a

later positive peak that is largest at the left parietal electrode site (Curran, Tepe, & Piatt, 2006; Rugg, 1995; Sanquist, Rohrbaugh, Syndulko, & Lindsley, 1980), and is therefore referred to as the left parietal old-new effect. This effect is typically maximal between 500-800ms after stimulus onset (Curran, 2004; Rugg & Curran, 2007), and the magnitude of the left parietal old-new effect is correlated with how much information about the event a participant can remember (Duarte, Ranganath, Winward, Hayward, & Knight, 2004). Similar (but typically more broadly distributed) old-new ERP effects are found in cued recall tests when comparing old vs. new cues, and these effects are also more positive when recall is successful (see Allan & Rugg, 1997) and are therefore thought to index successful reactivation of episodic information.

In addition to the earlier old-new effects described above, ERPs related to later "post-retrieval" processes during memory tests have also been extensively studied in the literature. Researchers have found late and sustained frontal positive effects, often right-lateralised, that are sometimes enhanced for old compared to new items, but are sometimes also found for new items (Leynes, Cairns, & Crawford, 2005; Rosburg, Mecklinger, & Johansson, 2011; see for review Wilding & Ranganath, 2011). These effects are often observed from around 500ms onwards and last for several seconds after stimulus onset, and are therefore thought to occur too late to index memory reactivation. Furthermore, since the effects are also found for new items or in other types of tasks that do not involve episodic retrieval (Hayama, Johnson, & Rugg, 2008), they do not seem to depend on successful memory reactivation. Therefore, researchers have interpreted these effects to relate to the involvement of executive control processes that are recruited to meet task demands, and have linked these ERP effects with converging evidence from fMRI research,

where the PFC is thought to mediate cognitive control operations (Jacques et al., 2017; Ranganath, 2004). Previous research has revealed that frontal late ERP effects were modulated by relative memory accuracy such that a stronger activation indicated high task demand, when participants needed to distinguish between true and imagined memories (Rosburg et al., 2011). In addition, Dzulkifli and colleagues (2004) suggested that this late frontal ERPs effect can be modulated by task difficulty, and is larger when the task involve more cognitive demands. In addition to frontal positivities, researchers have also found late sustained negativities over the posterior scalp (i.e. LPN effects as mentioned earlier) that have been associated with post-retrieval processing when the task requires monitoring response conflict between an automatic response and a task-appropriate response (Hu et al., 2015; Johansson & Mecklinger, 2003).

Most relevant to the topic of this thesis that used ERPs instead of fMRI, researchers have used ERPs to study both the cognitive processes during imagination that gives rise to later false memories that imagined events were perceived, and the after-effects of such imagination on retrieval processes during a subsequent test (Gonsalves & Paller, 2000, 2002; see also Gonsalves et al., 2004). In the study most relevant to the current thesis, participants were shown either pictures of objects, or object words and asked to imagine what the object looked like. During the imagination trials, it was found that posterior ERPs around 400-800ms after the word onset were more positive for words that the participants later mistakenly remembered as having been shown as pictures. This effect was interpreted to be related to visual imagination, and it was argued that vivid imagination had resulted in encoding of memories that had similar characteristics as memories of perceived pictures, thereby causing later reality monitoring failures on the subsequent test.

Furthermore, during the later test, parietal old-new ERPs were more positive during retrieval of true memories than retrieval of false memories, suggesting that false memories were associated with reductions in recollection. Thus, one previous study has investigated imagination inflation effects and their influence on source monitoring errors with ERPs, but to my knowledge this is the only published paper that investigated this issue. Furthermore, prior research has not investigated the effects of counterfactual imagination on memory with ERPs, as addressed in this thesis (see Chapter 5).

1.3. Memory Distortion and Forensic Memory Detection

As described in the previous sections, brain processes underlying true and imagined memory appear to be very overlapping, and a large body of research has illustrated how malleable our memories are in that they can be changed and updated after the event. This adaptive function of memory can be beneficial in some cases, such as for those who suffers from trauma such as childhood abuse, and for everyone else who have negative memories of past experiences. People who suffered negative experiences can try to imagine an alternative version that might had happened instead of those events, and eventually make themselves feel better. However, the fact that memories are modified by imagination has a serious drawback in forensic settings. People who committed a crime may try to come up with an alternative false version of their past such as a false alibi, in order to try to escape blame. Such episodic counterfactual thinking may have severe consequences for investigators when trying to find out the true version of what happened during the crime. In this section, I will discuss these issues in relation to forensic memory detection, which is where memory tests are used in forensic settings to try to determine the presence of

incriminating memories in a suspect's brain.

Detecting lies is one of the major challenges in forensic settings. Deception is also common in everyday life and it is difficult to distinguish between truthfulness and deception just by observation. Therefore, researchers are developing psychophysiological measures that can detect if someone is trying to deceive. These methods can be classified into two different categories: detecting deception and detecting concealed information (Ben-Shakhar, 2012). Detecting deception methods are designed to rely on physiological responses to direct questions that elicit lies in guilty suspects. For example, the Control Questions Technique (CQT) is often used combined with a polygraph. CQT is a set of questions that requires yes-no responses and involves three different types of questions: relevant (e.g. What did you do last Saturday night?), irrelevant (e.g. Did you ever steal something?), and control (e.g. Have you ever lie to get out of trouble?) questions. It is believed that if a person is innocent, he/she will response more strongly to control questions because those questions are more important to them than crime-related questions. Therefore, if there are stronger physiological responses (i.e. heart rate, respiratory rate, and skin conductance measure) to the relevant questions compared to the control questions, this will be considered as evidence of deception. Yet, CQT is rarely use due to its validity and reliability issues. During this stressful situation, control questions can fail to elicit strong reactions to provide enough psychological counter to the impact of the false accusation that is caused by relevant questions. Studies have therefore found that the CQT is biased against the innocent because it has high false positive rates in detecting guilt (review in Ben-Shakhar, 2012). A false accusation from the CQT is detrimental for the innocent suspects, especially in the case that false detection results in severe real-life consequences.

Because of the problems with the CQT, some researchers have suggested that tests that detect concealed knowledge (a.k.a. memory detection test; Ben-Shakhar, 2012) are more valid and reliable and can be used as alternatives to lie detection techniques (Verschuere, Ben-Shakhar, & Meijer, 2011). These tests are based on the assumption that the true suspect will have unique knowledge of the crime that another person would not. In contrast to the lie detection test, memory detection is an indirect measure of guilt. It examines whether a suspect can remember or recognise crime-related details, which is inferred as indicative of guilt. The tests do not require any overt deceptive response from the criminal suspect but rather measure behavioural or physiological markers of memory while the suspect engages in a seemingly irrelevant task. Thus, non-verbal markers of memory such as memory-related brain activity (e.g. Allen, Iacono, & Danielson, 1992; Gamer, Klimecki, Bauermann, Stoeter, & Vossel, 2012; Rosenfeld, Angell, Johnson, & Qian, 1991; Van Hooff, Brunia, & Allen, 1996), physiological activity (Gamer, 2011; Lykken, 1959), or reaction times and accuracy on indirect tests of memory (Sartori et al., 2008) can been used to assess whether suspects are guilty by detecting if they have any concealed knowledge of the crime. Many of these methods can very accurately detect concealed information, at least in cooperative research participants with little motivation to hide their guilt (Granhag, Vrij, & Verschuere, 2015; Verschuere et al., 2011).

However, one prominent concern is that real criminals may use countermeasure strategies to attempt to hide their guilt (Bergstrom, Anderson, Buda, Simons, & Richardson-Klavehn, 2013; Hu, Bergstrom, Bodenhausen, & Rosenfeld, 2015), threatening the validity of these tests in real-life settings. Considering the important societal, legal and ethical implications of forensic memory detection, it is

therefore critical to evaluate whether memory detection tests are susceptible to countermeasures. It is also important to assess which types of countermeasures are likely to be successful in order to ensure that memory detection tests are optimally designed to withstand evasion attempts. In the following section, I will review two prominent memory detection tests: the autobiographical Implicit Association Test (aIAT) and the Concealed Information Test (CIT), that were both used in experiments reported in this thesis to investigate the effects of counterfactual imagination on the accuracy of forensic memory detection (see Chapters 2-4).

1.3.1. Autobiographical Implicit Association Test (aIAT)

In 2008, a newly developed autobiographical Implicit Association Test (aIAT) was introduced as an inexpensive and practical tool to indirectly assess concealed memories (Sartori et al., 2008). It is a reaction time based memory detection measure, which was developed from the original Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). The IAT measures implicit associations regarding race and various other social attitudes (Gray, MacCulloch, Brown, Smith, & Snowden, 2005; Gray, MacCulloch, Smith, Morris, & Snowden, 2003; Gregg, 2007). The original IAT involves classifying stimuli into four different categories (Greenwald et al., 1998). For example, in an IAT test of implicit race attitudes towards European American/African American, classification stimuli are drawn from two target concepts categories (European American vs. African American names) and two attribute categories (pleasant vs. unpleasant words). Participants are asked to respond by pressing keys corresponding to the classification category. In the double classification block, one concept and one attribute category are assigned to the same key response (e.g. left button for both European and

pleasant words, right for both African and unpleasant words); and in the reverse classification block, the classification pairs were reversed (e.g. left button for both European and unpleasant words, right for both African and pleasant words). It is expected that concepts that are implicitly associated will produce faster and more accurate responses when they share the same key in double classification blocks than when they are mapped onto opposite keys. Thus, this test will reveal which attribute has a stronger association with the target concepts.

Similar to the original IAT, the aIAT involves sorting stimuli into categories but with the goal to assess implicit associations between autobiographic events and the truth (Sartori et al., 2008). It is supposed to allow researchers to investigate which of two contrasting autobiographical events is encoded in respondent's brain as a true experience. The aIAT involves presenting individual with four different types of statement to be classified on two dimensions: logically true versus false, and belonging to one of two alternative autobiographical events. For logically true versus false sentences, the sentence should be absolutely true or false for the individual. On the other hand, one of the two autobiographical should be true for individual and other should be false to the individual. Similar to the original IAT, double classification blocks are conducted where true/false sentences are paired with sentences describing either autobiographical event, and the key assignment is reversed across blocks. Thus, like in the original IAT, it is expected that participants will give faster and more accurate responses when an autobiographical event that is true shares the same key with logically true sentences in double classification blocks than when a true event shares a key with logically false sentences. Therefore, this test could help to evaluate if an individual is guilty or innocent of a crime by testing if they have implicit associations between that crime and the truth.

According to Sartori and colleagues (Sartori et al., 2008), this method is very promising because the accuracy in detecting concealed autobiographical knowledge was very high - guilty versus innocent participants could be correctly classified as such at a 91% rate on average (Agosta, Pezzoli, & Sartori, 2013; Sartori et al., 2008). When compared true and false autobiographical events, the aIAT was very effective in detecting which of the two event is true (Agosta, Castiello, Rigoni, Lionetti, & Sartori, 2011; Hu & Rosenfeld, 2012; Hu, Rosenfeld, & Bodenhausen, 2012; Lanciano, Curci, Mastandrea, & Sartori, 2013; Marini, Agosta, Mazzoni, Barba, & Sartori, 2012; Sartori et al., 2008; Takarangi, Strange, Shortland, & James, 2013) . Furthermore, the aIAT has been able to differentiate between true and false memories, which participants believed were true, that would therefore not be possible to identify at an explicit level (Marini et al., 2012). In Marini et al.'s study, participants heard list of words and were asked to response in the aIAT whether they had heard a certain word or not. This included critical word lures that had a strong semantic association with encoded words, which participants therefore believe they had heard in the list. They found that the aIAT was not only detecting which of the two words were more associated with the truth, but was also better in detecting the true memory than false memory—lures that participant believed were true (Marini et al., 2012). Likewise, recent research suggested that the aIAT can distinguish between performed and imagined actions (Shidlovski, Schul, & Mayo, 2014; Takarangi, Strange, & Houghton, 2015; Takarangi et al., 2013). Nevertheless, the efficiency in detecting imagined actions was not as strong as performed actions (Shidlovski et al., 2014).

The aIAT efficiency is not limited to general autobiographical events, but it can also be used in forensic settings. The study of mock crimes, where participants

perform a simulated “crime” as a task (such as stealing something), found approximately 90% suspects correctly identified as guilty when the aIAT contrasted a true mock crime with a false event (Hu & Rosenfeld, 2012; Sartori et al., 2008; Takarangi et al., 2013). Moreover, the aIAT was found to still be effective in detecting true memory traces after one month. Hu and Rosenfeld (Hu & Rosenfeld, 2012) reported that the aIAT detection efficiency for participants who had a month delay between the mock crime and the aIAT was as good as in another group who did the test immediately after the mock crime, suggesting that the aIAT is sensitive in detecting any concealed memory. Furthermore, promising results from the aIAT was not only found in laboratory research, as this test was recently used in a real court case in Italy (Sirgiovanni, Corbellini, & Caporale, 2016). The aIAT was used as an additional evidence for a convicted criminal suspect who was diagnosed with dissociative amnesia, claiming that they suffered from memory loss. Thus, there are already criminal justice systems that consider the aIAT to be reliable enough to use in forensic settings.

The aIAT has several advantages when used as a memory detection test. First, it is flexible in that any type of information can be assessed easily by adjusting category labels and sentences corresponding to what you are investigating (Sartori et al., 2008). For example, it can be used for examining drugs use (Vargo & Petróczi, 2013), assessing anxiety disorders (Teachman, Gregg, & Woody, 2001), and mock crimes. Second, it can be administered rapidly as it only takes approximately 10 minutes to complete the test (Agosta & Sartori, 2013). Third, it is cheap and only requires a computer with the aIAT software and no need for special training to administer the test (Agosta & Sartori, 2013). Finally, it can be administered remotely to participants (Agosta et al., 2013; Agosta & Sartori, 2013).

However, more recent research suggests that the aIAT should be use with caution and that multiple factors can moderate its accuracy. For example, Agosta and colleagues (2011) suggested that the aIAT is less effective when negative sentences (e.g. I have not been to Rome), including counter-affirmative sentences (e.g. I have been to a different place from Rome), were used. They found a reduction about 30% in accuracy in identifying the true autobiographical event and suggested that these kinds of sentences should not be used in the aIAT. Moreover, participants who are told about how the test works can employ countermeasure strategies to distort the test outcome. Studies found that instructing participants to speed up or slow-down responses can alter the test outcome (Hu et al., 2012; Verschuere, Prati, & Houwer, 2009). Vershuere and colleagues (2009) found that participants can alter the aIAT outcome by using simple strategies when they were given information on how the test works, when they had previous experience with the aIAT, and when they tried to speed up their responses. These findings were later confirmed by Hu and colleagues (2012). They found that speeding up in the incongruent block and slowing down in the congruent block can help guilty suspects obtain an innocent test outcome. This reduction in detection was even more pronounced when they asked participant to practice prior to the test (Hu et al., 2012)

Nevertheless, it has been suggested that fakers that intentionally change their response speed can be detected through a faking algorithm (Agosta, Ghirardi, Zogmaister, Castiello, & Sartori, 2011). Agosta and colleagues (2011) found a specific response pattern that could identify successful fakers. They suggested that successful fakers are more likely to respond abnormally slow in the congruent double classification block, compared to the preceding single classification blocks. They developed an algorithm to calculate a “faking index” by using the ratio

between the average reaction time in the fastest double classification block with the average reaction time of the corresponding single classification block. It is expected that this ratio will be larger for fakers than non-fakers due to fakers intentionally slowing in the critical congruent double classification block, compared to the single classification block (see also Cvencek, Greenwald, Brown, Gray, & Snowden, 2010). Thus, although the aIAT results can be faked, there are algorithms that may be possible to use to detect those fakers.

However, there is less research on other strategic countermeasures that individuals may use *prior* to the test, that do not require suspects to try to modulate their response times during the test itself, nor know exactly how the test works. As in real-life, guilty suspects may try their best to find a strategy to hide their guilt. For example, suspects may try to modify their actual memory by suppressing crime-related memories from coming to mind, or by generating a fake alibi to convince others of their innocence. According to Hu, Bergstrom, Bodenhausen, & Rosenfeld (2015), suppressing crime-related memories before the test can reduce aIAT detection accuracy. This finding extended earlier evidence that intentionally suppressing crime-related memories can reduce guilty detection with EEG (Bergstrom et al., 2013) and shows that suppression also impairs implicit influences of memory on the aIAT test (Hu et al., 2015). Thus, guilty suspects can succeed in escaping the test with this strategy without engaging in any intentional strategy during the test.

Another possibility is that suspects may lie during the interrogation and modify their memory prior to the test. They may come up with a counterfactual scenario to use as an alibi in order to attempt to appear innocent. As discussed previously, evidence suggests that people can create a vivid false memory of an event they never

experienced by imaging the event (Loftus & Pickrell, 1995; Schacter et al., 2011), and these false memories tend to have similar characteristics as true memories of events we have experienced (Mitchell & Johnston, 2009). Imagining false events can enhance the implicit association between the truth and the false memory, as measured with the aIAT. Previous research found that imagining simple actions increases the implicit association between the truth and the imagined event when compared with non-imagined events (Shidlovski et al., 2014). However, it is yet to be further investigated if the aIAT is susceptible to distorting effects of counterfactual imagination when applied as a countermeasure involving a false alibi for a mock crime, as addressed in this thesis (Chapters 2-3).

1.3.2. Concealed Information Test (CIT)

The Concealed Information Test (CIT) is another protocol that have been used to assess if a person recognises specific information known only by guilty suspects (Lykken, 1988; Verschuere et al., 2011), and that was also used in this thesis (Chapter 4). For example, in a theft case, this test could be used to investigate if a suspect recognises what has been stolen. Early CIT research were based on physiological response measures, particularly the skin conductance response (SCR; Lykken, 1960). Results from these studies were promising, in that 88% of guilty participants were correctly classified as guilty and none of the innocent participants were classified incorrectly. As a consequence, research in this area is expanding and the CIT have been used with various psychophysiological measures such as electrical brain activity in the form of P300 event-related potentials (Rosenfeld, Hu, Labkovsky, Meixner, & Winograd, 2013), and brain imagining techniques such as fMRI. Strikingly, the CIT is now being regularly used in the field and court in Japan

(Osugi, 2011), but is yet to be validated for the use in the U.S. Among different psychophysiological methods, P300-based ERPs seems to outperform autonomic nervous system measures, namely measured respiration line length and heart rate measures (Ben-Shakhar, 2012; Meijer, Selle, Elber, & Ben-Shakhar, 2014), and is considerably cheaper and easier to implement than fMRI methods, which has led to a surge of research on the P300-based CIT. This measure is also the focus in this thesis.

The CIT is a memory detection test that assesses whether a suspect recognises a critical stimulus (e.g. an item they are suspected of having stolen). The traditional CIT consists of two types of stimuli: relevant and neutral stimuli. It contains a series of multiple choice questions about crime-related detail (Lykken, 1959). For instance, for a theft a suspect may be asked "What did you steal, was it a…?"; and then given six multiple choices: (a) ring, (b) necklace, (c) watch, (d) key, (e) wallet, (f) phone. If there is a consistent enlarged physiological response for the relevant stimulus that reveals recognition, which enables investigators to infer that a suspect is guilty of the crime. On the other hand, for innocent suspects the relevant item will elicit a similar physiological response as any other neutral stimulus, since an innocent person cannot distinguish the relevant item from the other alternatives. It seems that the CIT outperforms other existing methods such as the CQT due to a more solid theoretical framework underlying the CIT. The rationale behind the CIT is based on the orienting response, which is a physiological response that many organisms show in response to a change in their environment, and involves an increase in skin conductance, a decrease in respiration and changes to heart rate (Lykken, 1959; Verschuere et al., 2011). It has been suggested that changes in this physiological response are due to this orienting reflex towards crime-relevant

stimuli, as people tend to show larger response to significant stimuli, compare to non-relevant stimuli. This in turn would indicating that a particular choice was more meaningful to a suspect than another item; thus, revealing that the person is guilty.

Related to the orienting response, the P300-based ERPs utilises P300 amplitude, which is known to be associated with meaningful recognition, as an index of critical crime details or concealed knowledge recognition (Lui & Rosenfeld, 2008). The P300 is a positive peak that occurs around 300-600ms after stimulus onset. It is still unclear exactly what cognitive process the P300 is indexing, and different subcomponents of the P300 have been related to attention and the orienting response (the novelty P300, or P3a component) versus stimulus categorisation, target detection or working memory updating (the P3b, see Polich, 2007). However, despite this uncertainty, practical applications of P300 are still possible since the amplitude of the P300 is very sensitive to recognition of a meaningful stimulus, and it is therefore expected that enlarged P300s will be observed for crime-relevant stimuli when compared to crime irrelevant stimuli for guilty suspects. On the other hand, if the person is innocent, the crime-relevant stimulus would be just another irrelevant stimulus to them, and there should not be any differences in the elicited P300 amplitude among stimulus types. The early P300-CIT, also known as "3-Stimulus Protocol", consists of three types of stimuli: a rare, crime-relevant probe (e.g. a stolen object), frequent irrelevant distractors (e.g. other objects that were not stolen), and a rare target stimuli (e.g. another object that is not relevant for the crime but that requires a special response in a target-detection task). The probe and irrelevant stimuli are mapped to the same key response, while the target stimulus has a unique response key (Allen et al., 1992; Farwell & Donchin, 1991; Rosenfeld et al., 1991, 1988). Incorporating a rare target stimulus that requires a special response

was designed to force respondent's attention to the stimuli sequence and avoid that participants fail to process the alternatives. However, more recent research found that this version of the P300-based CIT is vulnerable to countermeasures. It had been found that these different types of stimulus (probe, relevant, and target) were competing for attention (Rosenfeld et al., 2008). This in turn reduced P300 activation and, therefore, reduced accuracy of the CIT.

To improve the P300-based CIT, Rosenfeld and colleague (2008) introduced a Complex Trial Protocol (CTP) version of the CIT, which is claimed to be more accurate and more resistant to countermeasures. The new CTP-CIT is similar to the original CIT, but it separates the probe/irrelevant and target/non-target discrimination by a time delay. For each trial, respondents are first presented with one of two types of stimuli; either a probe (a crime-related stimulus, for example, the word "ring" if a ring was stolen) or irrelevant stimuli (distractors that are similar to the probe, for example, the words "wallet, key, necklace, watch, and phone"), and then after a delay, either a target (a number string that is assigned to different response button than non-target stimuli, for example, '111111') or non-target stimuli (other string numbers that are assigned to another response button than the target stimulus). The first response is a simple response to the probe/irrelevant stimulus by pressing a button to acknowledge that the stimulus was seen – the "I saw it" response - without any explicit discrimination. Then, participants have to explicitly discriminate between targets and non-targets by pressing different buttons for each in a target detection task, which is designed to force respondents to attend to the stimuli. In addition, respondents are informed at the beginning of the task that the test will be paused every few minutes and they will be asked to recall the most recent word (either probe or irrelevant) they have seen. This instruction was designed to

force respondents to pay more attention to the words, which is expected to enlarge P300 effect for probes if participants are guilty.

Several studies had found that the new P300-based CTP is better than the old three-stimulus version on both sensitivity and specificity (see Rosenfeld et al., 2013). The CTP has been found to be able to identify guilty suspects with more than 90% correct classifications, and it is claimed to be resistant to countermeasures (Winograd & Rosenfeld, 2011). In Winograd and Rosenfeld's study, participants were instructed to execute given countermeasure responses to all irrelevant stimulus. This manipulation was based on the previous finding that the P300 is enhanced by recognisable and meaningful items, and therefore executing responses to irrelevant items can enhance the P300 to those irrelevant, making it more difficult to discriminate between the probe and irrelevant items and impairing memory detection in the three-stimulus protocol (Rosenfeld, Soskins, Bosh, & Ryan, 2004). Therefore, participants were assigned different countermeasures to each irrelevant stimulus (e.g. left index finger pressure on the leg, left thumb pressure on the leg, and left big toe wiggle). However, in the CTP, guilty participants in the countermeasure group still showed enlarged P300 to the probe when compare to irrelevant stimulus and all participants were classified correctly according to their groups.

However, more recent research indicated that P300-based CTP can be less sensitive if suspects try to suppress crime-related memories before the test. According to Hu et al. (Hu et al., 2015), memory suppression reduced P300 activity associated with crime memories, which rendered the P300 effect for guilty and innocent suspects indistinguishable. This finding thus supported evidence from a previous study with the older three-stimulus protocol showing that when guilty participants suppressed their memory, the P300 marker related to memory was

significantly decreased (Bergstrom et al., 2013). As a consequence, “guilty” individuals escaped the detection and were classified as innocent. This is a crucial limitation for real-life applications of the P300 CIT, because simply trying to push the crime-related memory out of mind can help the guilty suspect evade detection.

Similar to the aIAT limitations, CIT also has a generalisability issue as most of the studies were conducted in a laboratory setting. Although CIT is being used regularly in Japan (Osugi, 2011), more field testing is required to validate this test. Also, there might be information leakage about details of the crime to innocent suspects via media. It is thus possible that innocent suspects may be found guilty for a crime that they did not commit simply because crime-related knowledge has been acquired and recognised from the media. Moreover, most of P300-CIT research has been conducted in the same laboratory (Rosenfeld and colleague). Although, the findings that they found in their in studies replicated one another, for validation purposes it would be important for other laboratories using the same method to also obtain similar findings. Therefore, further research is required to fully understand the validity and reliability of the CIT to determine whether it should be used in real-life. Critically as addressed in this thesis (Chapter 4), no one has investigated whether counterfactual imagination of a false alibi might distort suspects true memories of their crime, and thereby reduce memory detection accuracy with the CIT.

1.4. Overview of the thesis

This thesis investigated behavioural and ERP measures of how memory changes as a consequence of counterfactual imagination. Specifically, I investigated how memory changes after people imagine a fabricated version of a past event that involved real actions and interacting with real objects, thus leading to sensorimotor

rich, autobiographical memories. This introductory chapter highlighted relevant research findings related to memory distortion via imagination, forensic application in detecting concealed information, and possible theories that might explain how counterfactual imagination may give rise to memory distortion.

In Chapter 2, I present two experiments that investigated the effects of imagining and rehearsing an alibi as a countermeasure on the aIAT. Experiment 1 investigated whether imagining a false alibi impaired guilty detection with the aIAT when the false alibi and the true mock crime event are directly contrasted. Experiment 2 investigated whether the effect found in Experiment 1 persist regardless of which other event the mock crime is compared to, in order to better understand if the false alibi manipulation affected memory only by enhancing the implicit truth value of the alibi, and/or if it also decreased the implicit truth value of the committed mock crime.

Experiment 3, presented in Chapter 3, extended on Experiments 1 and 2, aiming to replicate findings of the first two studies but also investigating the effect of repeated rehearsal of a false alibi over a longer period of time, on aIAT memory detection. Specifically, I tested whether repeated rehearsal of a false alibi over a week long period might be more effective at impairing the true mock crime memories compared to a single brief alibi intervention just before the aIAT.

In Experiment 4, presented in Chapter 4, I investigated another concealed information detection method— the CIT in combination with ERP P300 measures. This chapter examined the extent to which people can modify crime-related memories through rehearsing an alibi by testing memory with the P300-based CIT. Finally, in Experiment 5, presented in Chapter 5, I addressed the neurocognitive mechanisms by which counterfactual imagination causes memory distortion using

behavioural and ERP measures. This chapter examined how repeated imagination of a counterfactual action might distort true memories of an action, and recorded ERPs to investigate neural activity during counterfactual imagination that predicts later false memories, and during the subsequent test to investigate after-effects of counterfactual imagination on retrieval processes. Finally, in Chapter 6 I discuss the main findings and implications of the empirical chapters. Limitations and future suggestions are also discussed in this chapter.

Chapter 2: The effect of imagining a false alibi on the autobiographical Implicit Association Test

In Chapter 1, I reviewed previous evidence that imagining an alternative version of an event can affect memory, and suggested that memory distortions as a result of counterfactual imagination may have severe consequences in forensic settings when investigators are trying to assess if a suspect is withholding incriminating knowledge. In this chapter, I present two experiments that focused on effects of counterfactual imagination on the autobiographical Implicit Association Test (Sartori et al., 2008) as a method used for detecting guilty memories. The experiments were based on previous findings that suppressing memories in advance of the aIAT can be an effective countermeasure that makes guilty participants appear innocent (Hu et al., 2015) and that learning alternative false details about a crime can interfere with memory detection in a Concealed Information Test (CIT) when used with autonomic measures (Gronau, Elber, Satran, Breska, & Ben-Shakhar, 2015). To my knowledge however, no previous research had investigated whether guilty suspects can intentionally create a counterfactual memory indicative of innocence as a countermeasure strategy for evading guilt detection with the aIAT, as assessed in the current experiments. In the current studies, I extended on previous findings by investigating a novel but ecologically valid strategy that guilty suspects may employ to appear innocent – namely imagining and rehearsing a false alibi before an interrogation.

Experiment 1

The first experiment investigated the effects of rehearsing a false alibi on aIAT memory detection. As previously described, the aIAT contrasts two different versions of autobiographical events in order to determine which event is relatively more strongly associated with the truth (Sartori et al., 2008). In the first study, I examined whether participants who had committed a mock crime (a simulated theft) could make themselves appear innocent by learning and imagining a false alibi scenario when the mock crime and alibi versions of events were directly contrasted in the aIAT. That is, would the aIAT be able to detect that the objectively true mock crime scenario was more associated with the truth than the false alibi scenario, or would aIAT truth detection be biased as a result of participants rehearsing and imagining the false alibi prior to the test?

The study was conducted in three stages. First, "guilty" participants carried out a mock crime which involved stealing a ring from a bag in a University staff office area, whereas "innocent" participants carried out an innocent act that involved going to the same office area but instead writing their email address on a paper slip on a staff member's door. Next, half of the guilty participants were instructed to imagine performing the innocent act with the explicit intention of using this as a false alibi in order to appear innocent. The other half of guilty participants and the innocent group performed an unrelated filler task. Finally, all three groups undertook an aIAT where the relative truth value of the mock crime and innocent/false alibi events were compared in all three groups.

I hypothesised that imagining a false alibi would create a memory for the innocent act, which may have some implicit associations with the truth even though participants knew their alibi was fake at an explicit level (Shidlovski et al., 2014).

Imagining a fake alibi would thus lead to lower aIAT discrimination between the objectively true mock crime and the objectively false innocent act when this group was compared to the guilty group who did not imagine the alibi. If imagining an alibi as a countermeasure was completely successful at making guilty suspects appear innocent, aIAT performance for these guilty participants would be indistinguishable from the innocent group who actually conducted the innocent act in real life.

Methods

Participants

Undergraduate students (N = 108) at the University of Kent were recruited via a research participation scheme in return for course credits. Participants were randomly assigned to three experimental groups (N = 36 in each); the Guilty-Alibi group (30 female and 6 male), the Guilty-Standard group (29 female and 7 male), and the Innocent group (28 female and 8 male). Twenty participants were excluded due to technical problems or for not followings the instruction during the experiment. Participants' age ranged from 18-28 (M_{age} = 19.83, SD = 1.62). The groups did not significantly differ in terms of age ($F(2,104) = .80$, $p = .451$, $\eta_p^2 = .02$) nor gender ($\chi^2(2) = .36$, $p = .837$, $\varphi = .84$). All participants had English as their first language, had normal or corrected-to-normal vision, and had no diagnosis of dyslexia. The study was approved by the University of Kent Psychology Ethic committee.

Materials, design and procedures

First, participants in the two Guilty groups were required to go to a kitchen adjacent to staff offices in a university building, find a bag, and steal a box from

inside the bag. They were explicitly asked to look and take note of what was inside the box (a ring), and then return with the box to the experimental room. The word ring was not mentioned in the instructions so that the memory of the ring was gained solely from enacting the crime. Innocent participants were required to go to the same area in the building, but instead they were told to write their email address on an appointment sign-up sheet on the door of a lecturer's office. Thus, Innocent participants were unaware of the mock crime.

Next, participants in the Guilty-Alibi group were provided with a fake alibi scenario, which was designed to help them appear innocent on the aIAT. Participants were told that they would soon take part in a test designed to detect their guilt, however they should aim to appear innocent by adopting the alibi. Participants were instructed that it was essential that they try to imagine the scenario as if it were true and that their memory for scenario details would later be tested. The alibi scenario was a short verbal description of the innocent act: "You were on your way to find your lecturer. On their door, there was a sheet of paper specifying that you could leave your email address for the lecturer to get back to you. So you tore off a bit of paper and wrote your email address and left it in the envelope provided and came back here. The envelope has since been destroyed so there is no evidence that your alibi is false". Participants were told to close their eyes and vividly imagine the alibi for two minutes. Next, they were asked to describe the scenario in detail and answer a few questions about it. If they gave incorrect answers, the alibi story was repeated and the questions asked again until the correct answers were given. Participants in the Guilty-Standard and Innocent groups were instead required to carry out a filler task of solving Sudoku puzzles.

They were given two puzzles as well as written instructions and told to do the best they could while they were timed for 5 minutes.

In the final stage, all participants took part in a seven-block computerised aIAT (Hu et al., 2015; Sartori et al., 2008). Participants were instructed that multiple sentences would appear on the screen and they would need to classify them as either logically true or false, or ring-related or email-related by pressing buttons on the keyboard. To avoid on-line attempts to modify the test result, they were not informed regarding how the test worked or how to alter their responses to appear innocent (cf. Agosta, Ghirardi, et al., 2011; Hu et al., 2012; Verschuere et al., 2009). The first block (20 trials) was a simple classification block that required participants to classify 5 true and 5 false sentences, with each sentence repeated twice in random order. Participants were instructed to press key 'Z' for logically true sentences (e.g., "I am a research participant") and key 'M' for logically false sentences (e.g., "I am playing football"), based on what they were doing at that time. The second block (20 trials) was a simple classification block that required participants to classify 5 sentences related to the guilty act (e.g., "I took a ring") and 5 sentences related to the innocent act/alibi scenario (e.g., "I wrote my email"). Participants were asked to press key 'Z' for ring-related sentences and 'M' for email-related sentences. Blocks three (20 trials) and four (40 trials) were critical double classification blocks which tested participants' responses to guilt congruent sentence pairings, because logically true and autobiographically true sentences for the Guilty groups were paired to the same response button. Participants were instructed to press 'Z' if the sentence was logically true or ring-related and 'M' if the sentence was logically false or email-related. Block five (20 trials) was a practice reverse simple classification block, which reversed the key assignments for ring and email-related sentences ('Z' for

email-related and ‘M’ for ring-related sentences). The final blocks six (20 trials) and seven (40 trials) were also critical double classification blocks but with the reversed keys, thus testing participants’ responses to guilt incongruent sentence pairings, because logically false and autobiographically true sentences for the Guilty groups were paired to the same response button. Participants were instructed to press ‘Z’ if the sentence was logically true or email-related and ‘M’ if the sentence was logically false or ring-related. Faster RT and higher accuracy for guilt congruent blocks than guilt incongruent blocks indicate an association between the crime and the truth, whereas the reverse pattern indicate an association between the innocent act and the truth.

Half of the participants conducted the blocks in the order described above, while blocks 2-4 and 5-7 were swapped for the other half of participants in order to counterbalance the order of guilt congruent and guilt incongruent blocks. For all blocks, sentences were presented on the screen in random order, and stayed on the screen until participants pressed a button. Participants were instructed to respond as quickly and accurately as possible, and if they pressed the incorrect button a red ‘X’ appeared on the screen until the pressed the correct button.

Results

D-scores

The main measure of guilt in the aIAT is the D-score, which combines accuracy and RT into a single, standardized measure (Greenwald et al., 2003; Hu et al., 2015; Sartori et al., 2008). To calculate this score, first, extreme RTs (<100ms or >10,000ms) were deleted. Incorrect responses were given a 600ms

penalty, and the mean RTs were calculated for the guilt congruent and guilt incongruent blocks separately, including the incorrect responses with the applied penalties. Finally, the mean RT difference between guilt congruent and guilt incongruent blocks was divided by the standard deviation of the RT distribution for correct trials only, from both blocks combined, in order to obtain the D-score. A positive D-score indicates guilt because it suggests that participants associated sentences representing the mock crime with the truth, whereas a negative D-score indicates innocence because it suggests that participants associated sentences representing the innocent act with the truth.

Mean D-scores were in the expected direction (Figure 2.1.) and were significantly different between the groups ($F(2, 105) = 9.46$, $p < .001$, $\eta p2 = 0.15$). The innocent participants, who undertook the innocent act but did not have any knowledge of the mock crime, elicited D-scores significantly below zero ($t(35) = -2.48$, $p = .018$, $d = 0.41$; calculated here and subsequently as the difference between means divided by the pooled standard deviation to ensure unbiased effect size estimates; Dunlap, Cortina, Vaslow, & Burke, 1996). Guilty-Standard participants, who committed the mock crime but did not have any knowledge of the innocent act, elicited D-scores significantly above zero ($t(35) = 3.25$, $p = .003$, $d = 0.54$). The Guilty-Alibi participants, who committed the mock crime and were also provided with an alibi scenario consistent with the innocent act, elicited D-scores non-distinguishable from zero ($t(35) = 0.18$, $p = .86$, $d = 0.03$). D-scores were significantly higher in the Guilty-Standard group than both Innocent ($t(70) = 4.06$, $p < .001$, $d = 0.96$) and Guilty-Alibi groups ($t(70) = 2.66$, $p = .010$, $d = 0.63$). However, there was only a non-significant trend for lower D-scores in the Innocent compared to the Guilty-Alibi group ($t(70) = 1.80$, $p = .076$,

d = 0.42). These results indicate that, as expected, imagining a fake alibi consistent with innocence impaired memory detection with the aIAT.

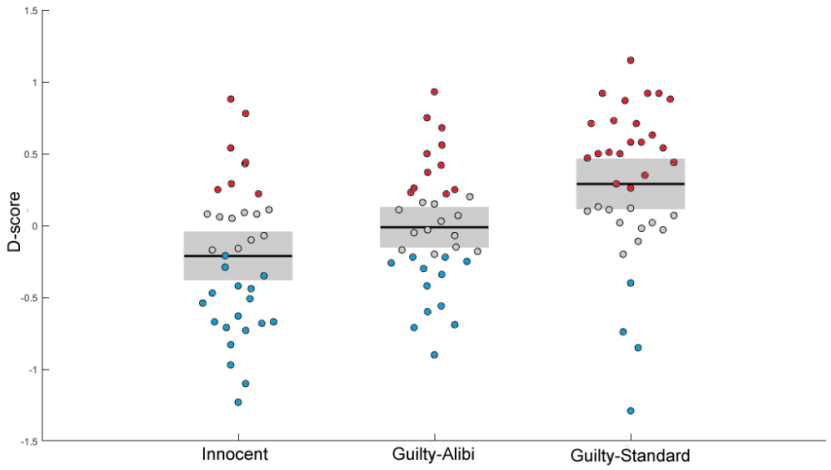

Figure 2.1. D-scores for the three groups from the Mock Crime/Innocent event aIAT in Experiment 1. The black lines shows the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the mock crime-related sentences are associated with the truth) and D-scores below zero suggest innocence (that the innocent-related sentences are associated with the truth). Scores consistent with guilt (>0.2) are marked with red dots, and scores consistent with innocence (<-0.2) are marked with blue dots. Grey dots indicate inconclusive scores. Scores are jittered along the x-direction for display purposes.

The aIAT was developed to diagnose guilt or innocence at the individual level, which is typically done by classifying individuals with positive D-scores as "guilty" and individuals with negative D-scores as "innocent" when contrasting a guilty vs. innocent event in this way (Sartori et al., 2008). We followed this previous research and compared classification rates across the different groups, after excluding participants scoring too close to zero (absolute D-scores between 0-0.2) as inconclusive (Agosta & Sartori, 2013). In the Guilty-Standard Group, 84% of participants were correctly classified as guilty, whereas in the Innocent group, guilt was classified significantly less frequent at 31% of the time ($\chi^2(1)$ =

14.72, $p < .001$, $\varphi = 0.54$). In the Guilty-Alibi group, guilt/innocence classification was around equal (48% guilty) which was significantly lower than in the Guilty-Standard group ($\chi^2(1) = 7.05$, $p = .008$, $\varphi = 0.38$) but not significantly different from the Innocent group ($\chi^2(1) = 1.50$, $p = .22$, $\varphi = 0.18$).

However, because the above classification rates are dependent on choosing specific cut-offs and the optimal cut-off may vary across samples, we also conducted a threshold-independent ROC analysis to evaluate classification performance using Areas Under the Curve (AUCs). The AUCs reflect the accuracy with which a randomly chosen participant can be classified into the correct group (Guilty or Innocent), where .5 reflects chance classification and 1.0 reflects perfect classification. This analysis showed that when comparing Guilty-Standard and Innocent groups, D-score classification was significantly better than chance (*AUC* = .70, *SE* = 0.06, $p = .004$), but comparing Guilty-Alibi and Innocent groups, D-score classification was less accurate and not significantly different than chance (*AUC* = .62, *SE* = .07, $p = .093$). Thus, individual classification rates also supported our prediction that imagining a false alibi would impair memory detection.

Reaction times and Accuracy

For RT (Figure. 2.2A), a 3 (Group) x 2 (Block) mixed ANOVA showed no main effects of neither Block ($F(1, 105) = 0.01$, $p = .932$, $\eta_p^2 < 0.001$), nor Group ($F(2, 105) = 1.47$, $p = .234$, $\eta_p^2 = .03$), but a significant interaction between Group and Block ($F(2, 105) = 5.46$, $p = .006$, $\eta_p^2 = 0.09$). Follow-up paired t-tests showed no significant RT difference between guilt congruent and guilt incongruent blocks in the Guilty-Alibi group ($t(35) = 0.47$, $p = .639$, $d =$

0.08). The Innocent group had significant slower RTs in the guilt congruent than the guilt incongruent block ($t(35) = 2.13$, $p = .040$, $d = 0.40$), whereas the Guilty-Standard group showed the opposite pattern ($t(35) = 2.27$, $p = .029$, $d = 0.47$).

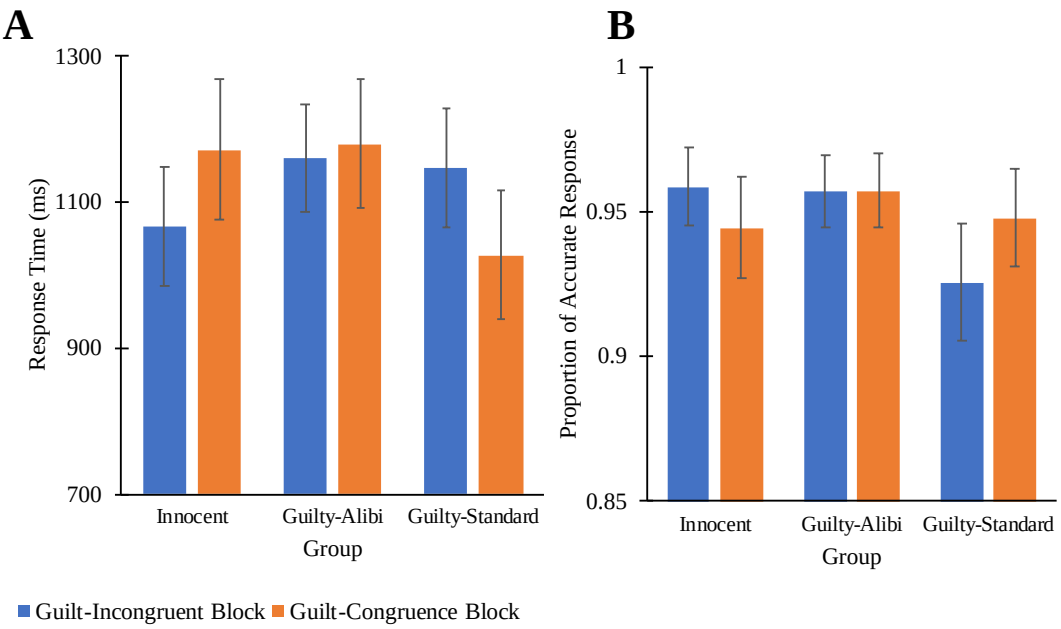

Figure 2.2. Proportion accurate responses (A) and mean reaction time (B) from the Guilty-Incongruent (True+Email/False+Ring) and Guilty-Congruent (True+Ring/False+Email) blocks of the Mock Crime/Innocent event aIAT in Experiment 1. Error bars denote 95% confidence intervals.

Comparing the groups directly within each block separately revealed that the Guilty-Standard group responded significantly faster than the Innocent group in the guilt congruent block ($t(70) = 2.24$, $p = .028$, $d = 0.53$). The Guilty-Standard group also responded significantly faster than the Guilty-Alibi group in the guilt congruent block ($t(70) = 2.46$, $p = .016$, $d = 0.58$) as predicted. However, there was no reaction time difference between Innocent and Guilty-Alibi group ($t(70) = 0.12$, $p = .905$, $d = 0.03$). There were no significant RT differences between the groups during the guilt incongruent block (Innocent vs. Guilty-Alibi: $t(70) = 1.72$, $p = .09$, $d = 0.43$; Innocent vs. Guilty-Standard: $t(70)$

= 1.40, p = .16, d = 0.33; Guilty-Standard vs. Guilty-Alibi: $t(70)$ = 0.25, p = .81, d = 0.06).

For accuracy (Figure. 2.2B), a 3 (Group) x 2 (Block) mixed ANOVA showed no main effect of neither Block ($F(1, 105)$ = .252, p = .617, $\eta_p^2 < 0.001$) nor Group ($F(2, 105)$ = 3.02, p = .053, η_p^2 = 0.05), but a significant interaction between Group and Block ($F(2, 105)$ = 3.65, p = .029, η_p^2 = 0.07). Paired t-tests revealed no significant difference in accuracy between guilt congruent and guilt incongruent blocks in the Innocent group ($t(35)$ = 1.38, p = .176, d = 0.30), and Guilty-Alibi group ($t(35)$ = .04, p = .971, d = 0.01). However, the Guilty-Standard group were more accurate in the guilt congruent block than guilt incongruent block ($t(35)$ = 2.09, p = .044, d = 0.41).

Comparing the groups directly within each block separately revealed that the Innocent group was significantly more accurate than the Guilty-Standard group in the guilt incongruent block, ($t(70)$ = 2.77, p = .007, d = 0.67). The Guilty-Alibi group was also significantly more accurate than the Guilty-Standard group in the guilty-incongruent block ($t(70)$ = 2.69, p =.009, d = 0.65). However, there was no difference between Innocent and Guilty-Alibi groups in the guilt incongruent block ($t(70)$ = 0.19, p = .853, d = 0.04). There were no significant Accuracy differences between groups during the guilt congruent block (Innocent vs. Guilty-Alibi: $t(70)$ = 1.20, p = .23, d = 0.28; Innocent vs. Guilty-Standard: $t(70)$ = 0.28, p = .780, d = 0.07; Guilty-Standard vs. Guilty-Alibi: $t(70)$ = 0.92, p = .360, d = 0.27).

Thus, this analysis showed that raw reaction times and accuracy on the critical guilt congruent and incongruent blocks only distinguished between the Guilty-Standard and the other two groups, whereas there were no significant

differences between the Guilty-Alibi and Innocent groups on either measure in either block. Therefore, Guilty-Alibi participants managed to appear innocent also when analysing raw RTs and Accuracy separately.

Faking analysis

In a final analysis, I calculated a "faking index" (Agosta, Ghirardi, et al., 2011) to assess whether rehearsing a false alibi would result in unusual reaction time patterns across aIAT blocks, since such patterns may function as signals of guilt even when the main guilt measure (i.e. D-score) is disrupted by countermeasures. The faking index is based on calculating the ratio between the mean RT in whichever double classification block is fastest for a particular person (which presumably reflects the truth congruent block for that person) with the mean RT in the corresponding single classification blocks, based on the logic that suspects who are trying to beat the test may be slowing down more in the critical double classification blocks than in the non-critical single classification blocks. Thus, the higher this index, the more that person is slowing down in the critical compared to non-critical blocks. To calculate the index, first all RTs below 150ms and above 10000ms were excluded. Next, any errors were replaced with the average RT of the block plus a penalty of 600ms. Finally, the ratio between the average RT of the fastest block (between 3 or 5) and single tasks that are directly connected to the fastest block in terms of motor response (1 and 2 or 1 and 4, respectively) was calculated (see Agosta et al., 2011, for more information).

In Experiment 1, the average faking index was higher in the Guilty-Alibi group (M = 1.05, SD = 0.20) than the Guilty-Standard (M = 0.97, SD = 0.15; $t(70)$ = 2.06, p =.040, d = 0.49) and Innocent groups (M = 0.95, SD = 0.15; $t(70)$ = 2.43, p

=.02, d = 0.58), who did not differ from each other ($t(70) = 0.40$, $p = .69$, $d = 0.10$). Using a cut-off value of 1.08 on the index (as suggested by Agosta et al., 2011), around 47% of the Guilty-Alibi group but only 19% of the Guilty-Standard group were classified as faking, and these rates were significantly different ($\chi^2(1) = 6.25$, $p = .012$, $\varphi = .30$). Faking classification was also higher in the Guilty-Alibi group than in the Innocent group (25%; $\chi^2(1) = 3.85$, $p = .050$, $\varphi = 0.23$), however classification rates did not differ between Guilty-Standard and Innocent groups ($\chi^2(1) = 0.32$, $p = .570$, $\varphi = 0.07$).

Similar to the D-score analysis, we also conducted a threshold-independent ROC analysis to evaluate faking classification performance. This analysis is appropriate because the most suitable threshold to use for detecting faking may differ across studies. The ROC analyses showed that when comparing Guilty-Alibi and Innocent groups, faking classification was significantly better than chance ($AUC = .65$, $SE = .07$, $p = .027$). When comparing Guilty-Standard and Innocent groups, faking classification was not different from chance ($AUC = .55$, $SE = .07$, $p = .480$). Thus, the faking analysis showed that guilty suspects who rehearsed a false alibi may reveal themselves by unusual reaction time patterns across aIAT blocks, although classification performance based on the faking-index was fairly poor. With only a 65% probability of classifying an individual correctly, this index would not be suitable to apply in practice.

Experiment 1 Discussion

In Experiment 1, the aIAT showed relatively good discrimination between guilt and innocence in participants who did not employ countermeasures, consistent

with previous findings (Agosta & Sartori, 2013; Sartori et al., 2008). However, the false alibi countermeasure significantly reduced memory detection with the key aIAT measure of guilt, the D-score, when compared to a standard guilty group who were not trying to evade the test. The false alibi countermeasure also significantly reduced markers of memory in raw reaction times and accuracy, suggesting that the success of this countermeasure was not dependent on the specific D-score algorithm. In fact, I did not find any significant differences in D-scores, reaction times, or accuracy between the Guilty-Alibi group versus a truly Innocent group, suggesting that the false alibi countermeasure was overall very successful. However, there were some trend-level differences between the groups, indicating that rehearsing a false alibi may not always be effective at making guilty suspects appear innocent in all cases.

Performance in the Innocent group showed a stronger relative association between the innocent act and the truth than the mock crime and the truth, whereas performance in the Guilty-Standard group indicated the opposite relative association. Performance in the Guilty-Alibi group however was equivocal as to which scenario was truthful. This pattern indicates that imagining a fake alibi created a memory for the innocent act that had some implicit associations with the truth, even though participants knew their alibi was fake at an explicit level (Shidlovski et al., 2014; Takarangi et al., 2015, 2013) This account is consistent with more general findings that imagining an event can create a memory for that event that has similar perceptual and behavioural characteristics as memories based on true experiences (Loftus, 2003; Loftus & Pickrell, 1995; Mitchell & Johnson, 2009; Schacter et al., 2011). Presumably, because both the mock crime and the innocent act had some

associations with the truth, neither of the critical aIAT blocks were truly congruent or incongruent with their memories, leading to similar performance in both blocks.

Since our false alibi participants were not instructed to intentionally alter their reaction times and were not informed regarding how the aIAT works, I predicted that a measure of faking that works through detecting abnormal response slowing during critical test blocks (Agosta, Ghirardi, et al., 2011) would not be particularly effective against this countermeasure. Interestingly however, detection rates with the faking index for the Guilty-Alibi group were above chance, showing that this group did indeed slow down responses in the critical blocks. This slowing was however not intentional as in previous research (Agosta, Ghirardi, et al., 2011; Verschuere et al., 2009), but rather more likely due to the previous creation of a false memory for the alibi and the resulting lack of a truly congruent block, which presumably led to some degree of response conflict (Marini, Agosta, & Sartori, 2016) in both critical blocks and thereby slower overall reaction times. This pattern contrasts with the Standard-Guilty and Innocent groups who both had relatively fast reaction times in one of the two blocks (although the faster block was the opposite across groups), which lowered their faking indexes. Thus, some participants in the Guilty-Alibi group could be detected with the faking index, although since classification was relatively poor it is questionable whether this index could be applied in practice to diagnose individual cases.

The results are consistent with the explanation that imagining a false alibi increased the truth value of that scenario, which thereby disrupted aIAT discrimination between the alibi and the mock crime. However, learning a counterfactual version of an event may also interfere with the veridical memory of the event and decrease its implicit truth value (Otgaar & Baker, 2018). Gronau et al.

(Gronau et al., 2015) asked participants to learn a hypothetical crime scenario with various details that were different from a mock crime they had actually conduced. Results showed that learning a false version of the mock crime impaired explicit recall of true crime details, and furthermore, reduced skin-conductance markers of true crime memories. They argued that true crime memories may have become inhibited as a result of retrieval competition between true and false crime details, similarly to the retrieval-induced forgetting phenomenon (Anderson, Bjork, & Bjork, 2000; Anderson et al., 1994; Anderson & Levy, 2007). Because the aIAT in Experiment 1 measured the relative truth of the false alibi versus mock crime scenarios, we can conclude that these scenarios had similar implicit truth values in the countermeasure group. However, we cannot determine whether the lack of a difference was due to increased implicit truth value of the false alibi, or reduced implicit truth value of the mock crime, or a combination of both. This issue was addressed in the next experiment.

Experiment 2

Experiment 2 used exactly the same false alibi manipulation, materials and procedure as in Experiment 1, with the only change being that the final test involved a different aIAT design that contrasted the mock crime with a non-experienced event that was clearly different from the learned false alibi. Thus, this study investigated whether imagining a false alibi would still impair detection of the mock crime regardless of which other scenario it is compared to. If such a pattern was found, it would indicate that the implicit truth value of the original crime-related memory was weakened by rehearsing an alibi, since any reduction in mock crime detection in this aIAT could not be due to inflated implicit truth value of the imagined alibi event as this scenario was not used as a contrast in the test.

I hypothesised that if the alibi manipulation was successful at reducing the implicit truth value of the true mock crime memory through an inhibition or interference mechanism (Anderson et al., 1994; Anderson & Levy, 2007; Gronau et al., 2015), rehearsing an alibi should reduce detection of guilty suspects on the aIAT when compared to guilty suspects who did not rehearsed an alibi after committing a mock crime. As a consequence, guilty suspects who rehearsed an alibi may not be distinguishable from the innocent group, who did not have any knowledge about the mock crime. Alternatively, if our previous finding was caused only by an increase in implicit truth value of the alibi scenario due to an imagination inflation-related process (Loftus & Pickrell, 1995; Shidlovski et al., 2014), then there should be no difference in aIAT performance between the guilty-alibi and guilty–standard groups as guilt detection rates in both groups should be equal, but both groups should be more likely to be detected as guilty than the innocent group.

Method

Participants

One-hundred and twenty participants took part in this study at University of Kent via a research participation scheme and received course credits for their participation. Twelve participants were excluded from the final results due to technical issues or failure to follow instructions during data collection. Thus, the final sample consisted of 108 participants in total (M_{age} = 18.94 year, *SD* = 1.98, age range = 18-36 years). Participants were randomly assigned to three experimental groups (N = 36 in each group): the Guilty-Alibi group (31 female and 5 male), the Guilty-Standard group (33 female and 3 male), and the Innocent group (30 female and 6 male). The groups did not differ in age ($F(2, 105) = .78$, $p = .461$, $\eta_p^2 = 0.02$), nor gender ($\chi^2(2) = 1.15$, $p = .563$, $\varphi = 0.10$). All participants had English as their

first language, had normal or corrected-to-normal vision, and had no diagnosis of dyslexia. The study was approved by the University of Kent Psychology Ethic committee.

Materials, design and procedure

The materials, design and procedure were identical to Experiment 1 with one exception; the aIAT version was different. As in Experiment 1, the study was conducted in three stages. First, participants in the two guilty groups carried out a mock crime in which they required to go to an office block and steal a ring from a bag, whilst innocent participants carried out an innocent act, involving writing their email address on a paper in the same area as the guilty participants. Next, half of the guilty participants were instructed to imagine performing the innocent act as a fake alibi with the explicit intention to use it as a strategy to appear innocent. The rest of participants performed a filler task. Finally, all three groups took an aIAT, which assessed which of two events had a stronger relative association with the truth. Importantly, instead of contrasting the mock crime and innocent act/false alibi directly, the aIAT in Experiment 2 contrasted the mock crime with a completely novel unexperienced event involving entering a lecturer's office and stealing an USB stick with exam questions on (henceforth referred to as the "exam" event, adapted from Sartori et al., 2008) that should not be associated with any truth value. Another addition in this study was that after the main experiment, all participants were asked to complete a questionnaire (see Appendix A-C). They were asked to give various ratings on a 0-6 scale regarding how they experienced and conducted the different tasks. They rated their nervousness during the mock crime (where 0 indicating not nervous at all; 6 extremely nervous), and how often they were thinking

about the mock crime during the aIAT (where 0 indicating not at all; 6 indicating all the time), their motivation to beat the aIAT (where 0 indicating not motivated at all; 6 indicating extremely motivated), and open-ended questions on whether they used any strategy to intentionally distort the test. There were also two additional questions for guilty-alibi participants: how vividly they had been able to imagine the alibi (where 0 indicating not vivid at all; 6 indicating extremely vivid) and how often they were thinking about the alibi during the aIAT (where 0 indicating not at all; 6 indicating all the time).

Results

D-scores

Mean standardized D-score indices of guilt (Greenwald, Nosek, & Banaji, 2003; Hu et al., 2015) were significantly different between the groups ($F(2, 105) = 6.73$, $p = .002$, $\eta_p^2 = 0.11$; see Figure 2.3.). Innocent participants, who had no knowledge of neither the mock crime nor the novel "exam" event, obtained a D-score that was not significantly different from zero as expected *(*$t(35) = .57$, $p = .569$, $d = .10$). Guilty-Standard participants, who committed the mock crime and did not have any knowledge of the exam event, elicited D-scores significantly above zero ($t(35) = 4.10$, $p < .001$, $d = .68$). The Guilty-Alibi participants, who committed the mock crime, were provided with an alibi scenario, and did not have any knowledge about the exam event, also elicited D-scores significantly above zero ($t(35) = 2.28$, $p = .029$, $d = .38$). D-scores were significantly lower in the Innocent group than Guilty-Standard ($t(70) = 3.59$, $p < .001$, $d = 0.86$) and Guilty-Alibi groups ($t(70) = 2.06$, $p = .043$, $d = 0.49$). However, there was also a trend towards lower D-scores in the Guilty-Alibi than Guilty-Standard group

($t(70) = 1.68$, $p = .097$, $d = 0.40$). These results indicate that imagining a false alibi does not abolish the implicit truth value of the true crime memory since the mock crime could still be significantly detected in the Guilty-Alibi group, but aIAT memory detection was somewhat less accurate than in a standard guilty condition.

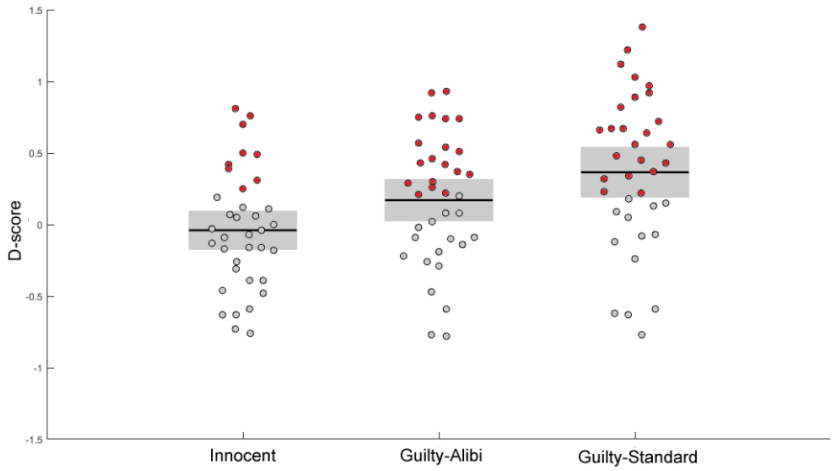

Figure 2.3. D-scores for the three groups from the Mock Crime/Unexperienced event aIAT in Experiment 2. The black lines shows the mean scores and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the ring-related sentences are associated with the truth). D-scores close to zero suggest that the events were equally associated with the truth, but because the test did not include a truly "innocent" event, innocence cannot be classified in this aIAT version. Scores consistent with guilt (>0.2) are marked with red dots, and inconclusive scores (<0.2) are marked in grey.

Similar to previous research (Agosta & Sartori, 2013; Sartori et al., 2008) and Experiment 1, I also classified individuals who elicited a positive D-score as "guilty" and compared classification rates between groups, after first excluding participants who scored too close to zero (absolute D-scores between 0-0.2). Since this version of the aIAT predicts scores close to zero for innocent participants, this criterion led to a high number of exclusions for innocent participants (excluded N for Guilty-Standard = 8, Guilty-Alibi = 9, Innocent =

16). In the Guilty-Standard Group, 82% of the remaining participants were correctly classified as guilty, whereas in the Innocent group, guilt was classified significantly less frequent at 50% of the time, as would be expected since neither event was true for this group, and these proportions were significantly different ($\chi^2(1) = 7.24$, $p = .007$, $\varphi = 0.39$). In the Guilty-Alibi group, guilt classification was 74% which was not significantly lower than in the Guilty-Standard group ($\chi^2(1) = .53$, $p = .469$, $\varphi = 0.10$), but significantly higher than in the Innocent group ($\chi^2(1) = 4.11$, $p = .043$, $\varphi = 0.30$).

Similarly to Experiment 1, I also conducted a threshold-independent ROC analysis to evaluate classification performance using Areas Under the Curve (AUCs). This analysis showed that when comparing Guilty-Standard and Innocent groups, D-score classification was significantly better than chance (*AUC* = .73, *SE* = 0.06, *p* = .001). Comparing Guilty-Alibi and Innocent groups, D-score classification was lower, but also better at chance (*AUC* = .64, *SE* = 0.07, *p* = .043). The D-score results thus indicated that rehearsing an alibi did not fully impair the original memory of the mock crime because these participants could still be detected as guilty, yet there was a numerical reduction in guilt classification for Guilty-Alibi participants.

Reaction Times and Accuracy

Next, raw RTs and accuracy (Figure 2.4.) were analysed separately to gain further insight into exactly how the Alibi manipulation affected performance. For RT, a 3 (group: Innocent vs. Guilty-Standard vs. Guilty-Alibi; between subjects) x 2 (block: congruent vs. incongruent; within subjects) mixed ANOVA showed a significant main effect of Block ($F(1, 105) = 18.30$, $p < .001$,

η_p^2 = 0.15) no main effect of Group (F(2, 105) = .82, p = .424, η_p^2 = 0.02), but an interaction between Group and Block (F(2, 105) = 6.98, p = .001, η_p^2 = 0.12). Follow-up paired t-tests showed significantly faster RTs in the guilt congruent than incongruent blocks for both Guilty-Alibi (t(35) = 2.48, p = .018, d = 0.38) and Guilty-Standard groups (t(35) = 4.76, p < .001, d = 0.70), but no significant RT differences between blocks in the Innocent group (t(35) = .39, p = .699, d = 0.05).

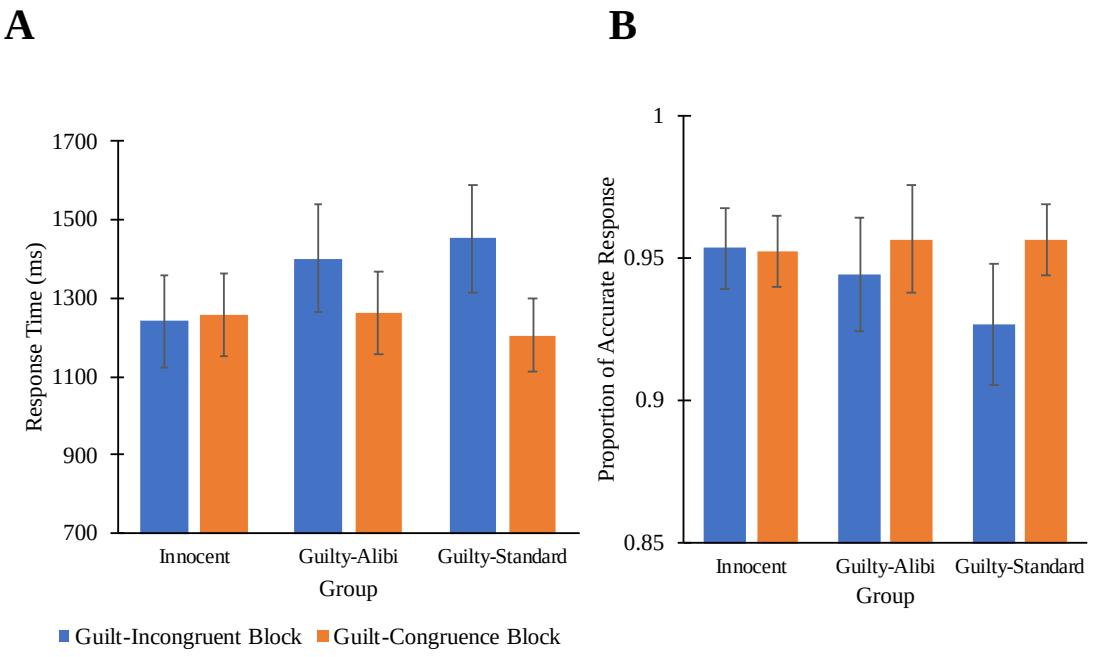

Figure 2.4. Mean response times (A) and proportion accurate responses (B) from the guilt-incongruent (True+Exam/False+Ring) and guilt-congruent (True+Ring/False+Exam) blocks of the Mock Crime/Unexperienced event aIAT in Experiment 2. Error bars denote 95% confidence intervals.

Comparing the groups directly within each block separately revealed that the Guilty-Standard group responded significant slower than the Innocent group in the guilt incongruent block (t(70) = 2.49, p = .015, d = 0.60), and there was a trend in the same direction for the Guilty-Alibi group compared to the Innocent group (t(70) = -1.81, p = .074, d = 0.43), but no significant RT differences between Guilty-Alibi vs. Guilty-Standard groups in the guilt incongruent

block($t(70) = 0.60$, $p = .552$, $d = 0.15$). There were no significant RT differences between any groups during the guilt congruent block (Innocent vs. Guilty-Alibi: $t(70) = 0.90$, $p = .929$, $d = 0.02$; Innocent vs. Guilty-Standard: $t(70) = 0.73$, $p = .468$, $d = 0.17$; Guilty-Standard vs. Guilty-Alibi: $t(70) = 0.82$, $p = .415$, $d = 0.20$).

For accuracy, a 3 (group: Innocent vs. Guilty-Standard vs. Guilty-Alibi; between subjects) x 2 (block: congruent vs. incongruent; within subjects) mixed ANOVA showed a significant main effect of Block ($F(1, 105) = 5.50$, $p = .021$, $\eta_p^2 = .05$). However, there was no main effect of Group ($F(2, 105) = .812$, $p = .447$, $\eta_p^2 = .02$) and the interaction was at trend-level ($F(2, 105) = 2.32$, $p = .104$, $\eta_p^2 = .042$). Paired t-tests revealed no significant difference in accuracy between guilt congruent and guilt incongruent blocks in the Innocent group ($t(35) = .14$, $p = .890$, $d = 0.07$), nor the Guilty-Alibi group ($t(35) = 1.27$, $p = .211$, $d = 0.22$). However, the Guilty-Standard group were more accurate in the guilt congruent block than guilt incongruent block ($t(35) = 2.46$, $p = .019$, $d = 0.57$).

Comparing the groups directly within each block separately revealed that there were significantly lower accuracy in the Guilty-Standard than Innocent group in the guilty incongruent block, but no other group differences in that block (Innocent group vs. Guilty-Standard group: $t(70) = 2.13$, $p = .037$, $d = 0.51$; Guilty-Alibi vs. Guilty-Standard: $t(70) = 1.21$, $p = .229$, $d = 0.29$; Innocent vs. Guilty-Alibi: $t(70) = .76$, $p = .450$, $d = 0.18$). There were also no significant accuracy differences between groups during the guilt congruent block (Innocent vs. Guilty-Alibi: $t(70) = 3.96$, $p = .693$, $d = 0.09$; Innocent vs. Guilty-Standard: $t(70) = .60$, $p = .548$, $d = 0.14$; Guilty-Standard vs. Guilty-Alibi: $t(70) = .08$, $p = .941$, $d = 0.02$). Thus, these results suggest suggests that manipulation effects on accuracy were rather limited and the main D-score findings were mostly driven

by group differences in speed at responding during the guilt incongruent block where the guilty groups were slower than the innocent group, presumably due to increased response conflict.

Faking Analysis

As in Experiment 1, the faking index was also calculated to investigate unusual response patterns. However, there were no difference between Innocent (M = 1.01, SD = .18) and Guilty-Standard groups (M = .94, SD = 0.16; $t(70)$ = 1.78, p = .080, d = 0.42), Innocent and Guilty-Alibi groups (M = .97, SD = 0.16; $t(70)$ = 1.00, p = .320, d = 0.24), nor Guilty-Standard and Guilty-Alibi groups ($t(70)$ = .80, p = .427, d = 0.19) in the average faking index. As suggested by Agosta and colleagues (2011), suspects who scored higher than 1.08 could be classified as a faker. Using this cut-off classification, 25% of the Guilty-Alibi group and 17% of the Guilty-Standard group were classified as faking, which was not significantly different ($\chi^2(1) = 0.76$, p = .384, φ = 0.103). There was also no difference between Innocent (31%) and Guilty-Alibi group ($\chi^2(1) = .28$, p = .599, φ = 0.06) nor between Innocent and Guilty-Standard group ($\chi^2(1) = 1.93$, p = .165, φ = 0.16) in faking classification proportions.

The ROC analysis showed that faking classification was not different from chance when comparing Innocent and Guilty-Alibi groups (AUC = .56, SE = .07, p = .368), nor when comparing Innocent and Guilty-Standard groups (AUC = .60, SE = .07, p = .128), nor when comparing Guilty-Alibi groups and Guilty-Standard groups (AUC = .55, SE = .07, p = 454). Thus, the faking analysis in Experiment 2 showed that rehearsing an alibi did not cause any unusual reaction time patterns across aIAT

blocks when the aIAT contrasted the mock crime with an unexperienced event, because faking classification was relatively low and similar across all groups.

Post-Experiment Questionnaire Analysis

Ten participants (4 Innocent, 3 Guilty-standard and 3 Guilty-Alibi) were excluded from the questionnaire analysis due to missing responses. The results revealed no differences between Guilty-Standard (M = 2.76, SD = 1.60) and Guilty-Alibi (M = 2.60, SD = 1.46) groups in nervousness during the mock crime ($t(64)$ = 0.40, p = .689, d = 0.10) and the extent to which they thought about the mock crime during the aIAT (M = 3.21, SD = 1.53; M = 3.52, SD = 1.17, respectively; $t(64)$ = 0.90, p = .372, d = 0.23). However, there was a significant difference between guilty groups in their motivation to beat the test: the Guilty-Alibi (M = 4.15, SD = 1.18) group was more motivated to appear innocent than the Guilty-Standard group (M = 3.45, SD = 1.35; $t(62)$ = 2.24, p = .029, d = 0.56). The Innocent group reported being significantly less nervous while conducting the innocent task than the Guilty groups were while conducting the mock crime (Innocent M = 1.78, SD = 1.60; Innocent vs. Guilty-Alibi: ($t(63)$ = 2.17, p = .033, d = 0.55; Innocent vs. Guilty-Standard: ($t(63)$ = 2.46, p = .017, d = 0.62). They also thought less about the innocent act during the aIAT than the two Guilty groups thought about the mock crime during the aIAT (Innocent M = 1.00, SD = 1.50; Innocent vs. Guilty-Alibi: ($t(63)$ = 7.53, $p < .001$, d = 1.90; Innocent vs. Guilty-Standard: ($t(63)$ = 5.87, $p < .001$, d = 1.48), as would be expected since there were no sentences related to the innocent act in this aIAT version.

Exploratory correlation analyses were also conducted to investigate factors that possibly related to performance in the aIAT. In the Guilty-Standard group, there were no significant correlations between D-score and either nervousness ($r(33)$ = -

.19, p = .289), motivation to beat the test (r(33) = -.32, p = .073), or thinking about the mock crime during the task (r(33) = .17, p = .352). In the Guilty-Alibi group, there were no significant correlations between the D-score and any factors: the extent to which participants were thinking about the mock crime during the aIAT (r(33) = .26, p = .139), nervousness (r(33) = -.08, p = .661), motivation to beat the best (r(33) = -.20, p = .266), the extent to which participants were thinking about the alibi scenario during the aIAT (r(33) = -.11, p = .532) or how vivid they imagined the alibi during the preceding imagination task (r(33) = .23, p = .195).

Experiment 2 Discussion

Experiment 2 assessed whether imagining a false alibi reduces the implicit truth value of the true crime memory, in line with previous findings that have shown that learning counterfactual details after a mock crime can interfere with true memories of the crime (Gronau et al., 2015). In Experiment 1, the results showed that the aIAT was unable to determine whether an experienced mock crime or an imagined false alibi was true. However, because of the aIAT design, I was unable to test whether this lack of discrimination was caused by increased truth value of the imagined alibi or decreased truth value of the mock crime, or a combination of both. In Experiment 2, I therefore contrasted the mock crime with a novel event that had been neither experienced nor imagined in an aIAT, order to assess the implicit truth value of the mock crime memory independent of the alibi memory. In this study, the mock crime was still detected despite participants previously imagining a false alibi, suggesting that the alibi had not impaired the true memory of the crime to a substantial extent.

As expected in Experiment 2, the mean D-score of innocent participants was close to zero, suggesting that they associated both events equally with the truth. Both guilty groups scored above zero, indicating that they associated the mock crime with the truth more than the unexperienced event. However, there was a tendency for a reduction in detecting that a suspect was guilty of the crime when they adopted an alibi. Even though the D-score suggested that guilty suspects who imagined a false alibi still associated the mock crime with truth there was a non-significant trend towards a lower score in the alibi than the standard group, and there were no differences between innocent and alibi groups in raw reaction times or accuracy, even though such differences were found between the standard guilty and innocent groups. Furthermore, in contrast with Experiment 1, the "faking index" (Agosta, Ghirardi, et al., 2011) was not able to detect that participants had employed a countermeasure in the alibi group, indicating that the usefulness of this index is questionable.

Thus, the results in Experiment 2 showed that with this aIAT design, the mock crime could still be significantly detected after imagining an alibi, although there was a numerical tendency towards lower detection than in a standard guilty group who received no countermeasure instructions. Therefore, it appears that the low discrimination between the experienced mock crime and imagined alibi in Experiment 1 was mainly driven by the alibi manipulation increasing the implicit truth value of the imagined scenario, and only a subtle reduction (if any) of implicit truth value of the mock crime memory.

General discussion

The aIAT has been promoted as an accurate tool for determining which of two autobiographical events are true, with promising applications in forensic memory detection (Agosta et al., 2013; Sartori et al., 2008). However, recent research has revealed potential countermeasures that guilty suspects can adopt to make themselves appear innocent, such as intentionally altering their responses during the test itself (Agosta, Castiello, et al., 2011; Hu et al., 2012; Verschuere et al., 2009), or suppressing their incriminating memories in advance of the test (Hu et al., 2015). I tested whether a novel countermeasure that has recently been applied in physiological memory detection (Gronau et al., 2015) would also be effective at reducing detection using the aIAT. Specifically, I assessed whether instructing guilty suspects to intentionally create a memory for a false alibi would affect aIAT performance. The results suggest that rehearsing a false alibi could affect mock crime detection with the aIAT, depending on how the aIAT is constructed. This finding is consistent with previous evidence that imagining an alternative version of an event can reduce physiological memory detection (Gronau et al., 2015) and affect implicit truth value as measured with an aIAT (Shidlovski et al., 2014; Takarangi et al., 2015, 2013). However, Experiment 2 showed that there was only a subtle reduction in aIAT detection after rehearsing a false alibi when contrasting the mock crime to a non-experienced event. This pattern suggests that imagining a false alibi might somehow create an implicit association between the imagined event and the truth, even though participants were aware that the event had never occurred.

The results are predicted by the literature on counterfactual thinking, which has shown that repeatedly mental stimulating an event may cause the imagined event to become more salient and memories for the imagined details may be as vivid as the

actual event memory (Gronau et al., 2015). Similarly, Takarangi et al. (2013) found that aIAT was less effective in distinguishing between two autobiographical events when participants had imagined themselves performing an action that they had actually not performed (Takarangi et al., 2015). Furthermore, Foerster and colleagues (2017) suggested that rehearsing a false alibi can cause it to become a default response such that when a cue triggered a memory about a mock crime, that memory is automatically inhibited to facilitate a false alibi response (Foerster et al., 2017). However, this literature has also found that repeatedly thinking counterfactually can impair memories for the event that actually occurred (Petrocelli & Crysel, 2009), but my Experiment 2 did not find a reliable impairment of the true mock crime memory.

To conclude, results from my first two Experiments revealed that imagining a false alibi can disrupt aIAT memory detection if the test directly contrasts the true event with the false alibi. This is problematic for forensic applications of the aIAT because these may set up the test to contrast a crime event with the suspect's version of what happened (which if they are guilty might be their false alibi). My findings suggest that the aIAT is very vulnerable to countermeasures and that adopting the aIAT in real life cases is premature, because the test does not seem to accurately measure the actual truth value of an event.

Chapter 3: The effect of repeatedly rehearsing an alibi on aIAT memory detection

In the previous two experiments, I investigated whether participants who imagined a false alibi after committing a mock crime could evade subsequent memory detection with the autobiographical Implicit Association Test (Sartori et al., 2008). I found that aIAT memory detection was substantially impaired when contrasting a true mock crime with the false alibi directly, but the alibi did not seem to impair the original memory of the mock crime. There was only a subtle reduction in detecting the mock crime memory in these participants when the aIAT contrasted the mock crime with an unexperienced event, showing that the alibi countermeasure was not completely effective at hiding people's true incriminating memories. In the current experiment, I therefore extended on the previous studies to investigate whether a potentially more powerful version of the alibi manipulation might impair the true mock crime memory, which would have interesting theoretical and practical implications.

One possible reason why the true mock crime memory was unimpaired in Experiment 2 might be that the alibi manipulation was only implemented through one brief rehearsal and imagination phase. Thus, the effect of the alibi manipulation may not have been as strong as in real life situations where suspects may prepare and imagine an alibi repeatedly and over a long-time period before the interrogation. If participants were able to rehearse/imagine the alibi in this way, it may be more likely to impair the true memory of the mock crime, either by increased retroactive interference or by inhibition of the crime memory representation itself (Gronau et al., 2015). Previous research has shown that when multiple memories are associated to the same cue, repeatedly retrieving one memory in the face of competitive activation

of another memory can cause the non-selected memory to become inhibited (Anderson et al., 1994). Likewise, repeatedly pushing an unwanted memory out of mind by thinking of a substitute thought may interfere with (Bergstrom, de Fockert, & Richardson-Klavehn, 2009) retrieval of the original memory, or even inhibit it (Benoit & Anderson, 2012). The literature on motivated forgetting suggests that such impairments of unwanted memories are gradual and increase with repetition (e.g. Anderson & Green, 2001), predicting that a true crime memory might only become impaired if a false alibi is *repeatedly* retrieved. Thus, the next experiment assessed whether repeated and temporally extended imagination of an alibi impairs the original crime memory.

Experiment 3

Experiment 3 was designed to replicate and extend on findings from the previous studies, with particular focus on whether repeated rehearsal of a false alibi over an extended time period might be more effective at impairing the true memories compared to a single brief alibi intervention just before the aIAT. In the previous two experiments all experimental phases were conducted in the same session; participants first conducted a mock crime, then immediately learned and imagined the false alibi, which was followed by the aIAT. The current study therefore added a time delay of one week between the mock crime and test, which made the design more realistic and enabled us to investigate the effect of repeated and distributed false alibi rehearsal on aIAT memory detection.

The experimental design was similar to the previous studies, except that it was conducted in two sessions one week apart, and included an additional experimental group. Furthermore, in the second session, all participants completed

three versions of the aIAT that contrasted the mock crime vs. the innocent/alibi event (same aIAT as in Experiment 1), the mock crime vs. an unexperienced event (same aIAT as in Experiment 2), and the alibi vs. the unexperienced event (a new aIAT version to assess the implicit truth value of the innocent act/alibi independently of the mock crime). Similarly to previous experiments, participants first conducted either an innocent act or a mock crime, depending on which group they were assigned to. All participants then came back for the aIAT session a week later. In one countermeasure group ("Guilty-Alibi"), participants conducted a mock crime during the first session, then left and returned a week later at which point they learned and imagined the false alibi immediately before the aIATs. In the other countermeasure group ("Guilty-Alibi with home training"), participants learned and imagined the false alibi during the first session immediately after conducting the mock crime, and were also required to repeat this imagination task at home once a day for a week before returning to complete the aIATs. These two countermeasure groups were compared against Innocent and Guilty-Standard groups, as in the previous two studies.

I expected that participants who carried out an innocent act should be detected as innocent and participants who committed a mock crime without learning an alibi should be detected as a guilty across the relevant aIAT versions. However, participants who learned the false alibi would be less likely to be detected as guilty than the standard guilty group. If imagining a false alibi leads to gradual strengthening of the false memory and/or gradual impairment of the true memory with repetition, then extended rehearsal of a false alibi for a week before the test should be particularly effective at making guilty suspects appear innocent.

Methods

Participants

The final sample consisted of 144 undergraduate students from the University of Kent who took part via a research participation scheme in return for course credits (M_{age} = 19.13 year, SD = 1.57, age range = 18-34 years). Twenty-eight additional participants were excluded due to technical errors, failures to follow instructions, or failure to attend both sessions. Participants were randomly assigned to one of the four groups (N = 36 in each group): Innocent (30 female, 6 male), Guilty-Standard (30 female, 6 male), Guilty-Alibi (27 female, 9 male), and Guilty-Alibi with Home Training (HT; 31 female, 5 male). The groups did not differ in terms of age ($F(3,140) = 0.74$, $p = .531$, $\eta_p^2 = .02$) nor gender ($\chi^2(3) = 1.69$, $p = .639$, $\varphi = 0.11$). All participants had English as their first language, had normal or corrected-to-normal vision, and had no diagnosis of dyslexia. The study was approved by the University of Kent Psychology Ethic committee.

Materials, Design, and Procedures

The design of this study was closely based on the experiments in the previous chapter with three exceptions: 1) there was an additional experimental group (Guilty Alibi with HT), 2) there was a week delay between the mock crime and the aIAT session for all groups, and 3) all groups took three aIAT versions, contrasting the mock crime with the innocent/alibi event, the mock crime with an unexperienced event, and the innocent/alibi event with the unexperienced event.

To begin with, participants in all three Guilty groups committed a mock crime involving going to a staff office area and stealing a ring whereas participants in the Innocent group completed an innocent task involving writing their email

address on a note in the same area (both these tasks were kept identical to Experiments 1 and 2). Next, all participants were dismissed and asked to come back the laboratory after a week, except the Guilty-Alibi with HT group. The latter group were given instructions to perform an extra task after completing the mock crime. They first learned and rehearsed a false alibi which described the innocent act, using the same materials and procedure as in Experiments 1 and 2. Next, they were given a home training task, which required them to access an internet link in order to rehearse the false alibi once every day in the intervening six days until the test day. When they accessed the link, they were asked to read a description of the alibi (same text as used on the first day) and imagine themselves doing the described actions as vividly and accurately as possible. After that, they asked to write down a detailed description of the scenario they had imagined and rate how vivid their alibi imagination had been.

After a week, all participants came back to the lab to complete the rest of the study. Participants in Innocent and Guilty-Standard group were asked to complete a filler task (solving Sudoku puzzles), while the two Alibi groups rehearsed the same alibi (describing the innocent act). For the Guilty-Alibi group, this was the first time they learned that they needed to use an alibi to appear innocent and found out the details of the alibi/innocent act, whereas for the Guilty-Alibi with HT group it was another chance to rehearse the alibi they had learned and repeated during the preceding week. Finally, all participants completed three versions of the aIAT: 1) contrasting the mock crime vs. innocent/alibi events (same aIAT as in Experiment 1); 2) contrasting the mock crime vs. non-experienced (stealing exam) events (same aIAT as in Experiment 2); and 3) contrasting the innocent/alibi vs. non-experienced (exam) events (a novel aIAT version used to assess whether the innocent event

would be detected as true after rehearsing a false alibi). The aIAT task design, sentences and instructions were identical to those used in the previous chapter, with the only changes being the new version 3, and that all participants undertook all three versions. The order of aIAT congruent/incongruent blocks and versions were counterbalanced across participants to prevent order effects.

After the experiment, participants were asked to complete a questionnaire (see appendix D-F), which was similar to the one used in Experiment 2 with a few additional questions about details of the innocent act or mock crime. For the Innocent group, participants were required to give answers relating to details of the innocent act and give ratings on a scale from 0 to 6 regarding their behaviour and experience during the initial act and the aIAT (e.g. in how much detail they could remember the act, their motivation to beat the aIAT, and the extent to which they thought about the act during the aIAT). The Guilty groups were asked to provide answers regarding details of the mock crime and provide various ratings on a 0-6 scale regarding their nervousness during the mock crime, their motivation to beat the aIAT, the extent to which they thought about the mock crime during the aIAT, and whether they had intentionally used any strategy to distort the test, including the extent to which they thought about the alibi scenario during the aIAT and how vividly they had imagined an alibi (for the Guilty Alibi groups only).

Results

Ring/Email aIAT

D-score

The Ring/Email version of the aIAT directly contrasted the mock crime (ring) with the innocent/alibi (email) event, and was identical to the aIAT used in Experiment 1. In this test, positive D-scores (Greenwald et al., 2003; Hu et al., 2015)

are indicative of guilt because they suggest participants associate the mock crime with the truth whereas negative D-scores are indicative of innocence because they suggest participants associate the innocent event with the truth. Means and standard deviations of the D-scores are shown in Table 3.1. Results revealed that the mean D-score of the Innocent group was not significantly different from zero ($t(35) = -1.312$, $p = .198$, $d = .22$; see Figure 3.1), inconsistent with the predictions and suggesting that the innocent event was not detected as true in this group on average. The Guilty-Standard group however did obtain a D-score that was significantly above zero ($t(35) = 3.749$, $p = .001$, $d = .62$) indicating successful guilt detection in this group. The Guilty-Alibi group who committed a mock crime and learned a false alibi just prior to the test however had a mean score significantly *below* zero ($t(35) = 2.056$, $p = .049$, $d = .34$), thus looking more innocent than guilty. In contrast, the Guilty-Alibi with HT group, who committed a mock crime and then repeatedly rehearsed a false alibi for a week before the test, did not have a mean D-score that differed from zero $t(35) = 1.014$, $p = .317$, $d = .17$). The mean D-scores were significantly different between the groups ($F(3, 140) = 6.78$, $p < .001$, $\eta_p^2 = 0.13$; see Figure 3.1). Independent t-tests revealed that the mean D-score of the Innocent group was significantly lower than in the Guilty-Standard group, while there were no differences between the Innocent and either of the Alibi groups (see Table 3.2 for t-tests). However, the mean D-score of the Guilty-Standard group was significantly higher than the Guilty-Alibi group, but not different from the Guilty-Alibi with HT group. Likewise, the mean D-score of the Guilty-Alibi with HT group was significantly higher than the Guilty-Alibi group, suggesting that home training with the alibi actually made it a *less* effective strategy for appearing innocent on this aIAT version.

Table 3.1. Means and standard deviations of d-scores, accuracy, and reaction times for all aIAT versions in Experiment 3.

Group	D-Score		Accuracy				RT			
			Guilt-Incongruent		Guilt-Congruent		Guilt-Incongruent		Guilt-Congruent	
Ring/Email aIAT	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Innocent	-0.08	0.34	0.95	0.04	0.95	0.04	1020.82	239.41	1051.57	197.84
Guilty-Standard	0.21	0.33	0.94	0.04	0.97	0.03	1063.87	218.58	978.4	142.75
Guilty-Alibi	-0.12	0.35	0.96	0.04	0.95	0.05	1032.56	239.2	1119.5	332.02
Guilty-Alibi with HT	0.06	0.33	0.96	0.05	0.96	0.04	1143.21	323.3	1121.58	278.38
Ring/Exam aIAT										
Innocent	-0.01	0.43	0.96	0.04	0.95	0.05	1069.68	282.87	1090.88	293.83
Guilty-Standard	0.22	0.33	0.92	0.12	0.96	0.04	1159.61	229.08	1042.81	203.26
Guilty-Alibi	0.14	0.44	0.94	0.05	0.96	0.03	1102.35	299.38	1044.84	279.20
Guilty-Alibi with HT	0.2	0.39	0.96	0.05	0.97	0.04	1192.1	348.25	1100.51	312.96
			Innocence-Incongruent		Innocence-Congruent		Innocence-Incongruent		Innocence-Congruent	
Email/Exam aIAT	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Innocent	0.02	0.34	0.95	0.04	0.96	0.04	1085.68	280.36	1087.08	285.14
Guilty-Standard	-0.01	0.29	0.95	0.04	0.95	0.04	1057.25	161.3	1074.22	213.84
Guilty-Alibi	0.13	0.35	0.94	0.07	0.96	0.03	1144.41	340.97	1084.02	298.95
Guilty-Alibi with HT	0.13	0.36	0.96	0.04	0.96	0.05	1192.81	297.7	1138.42	282.16

Note. N = 36 for all the groups

Table 3.2 Independent t-test results comparing group performance during the Ring/Email aIAT

Variable	D-score			RT						ACC					
				Guilt-congruent			Guilt-incongruent			Guilt-congruent			Guilt-incongruent		
	t	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
Innocent x Guilty-Standard	**3.54***	**<.001**	**-0.84**	1.8	0.08	0.42	0.8	0.43	-0.19	1.58	0.12	-0.37	1.30	0.2	0.31
Innocent x Guilty-Alibi	0.53	0.6	0.13	1.06	0.3	-0.25	0.21	0.84	-0.05	0.05	0.96	-0.01	0.49	0.62	-0.12
Innocent x Guilty-Alibi with HT	1.64	0.1	-0.39	1.23	0.22	-0.29	1.83	0.07	-0.43	0.86	0.39	-0.2	0.38	0.7	-0.09
Guilty-Standard x guilty- alibi	**4.08***	**<.001**	**-0.96**	**2.34***	**0.02**	**0.55**	0.58	0.56	-0.14	1.46	0.15	0.34	1.77	0.08	0.42
Guilty-Standard x Guilty-Alibi with HT	1.94	0.56	-0.46	**2.75***	**0.01**	**0.65**	1.22	0.23	0.29	0.72	0.15	-0.17	1.46	0.15	0.35
Guilty-Alibi x Guilty-Alibi with HT	**2.19***	**0.03**	**-0.52**	0.03	0.98	-0.01	1.65	0.1	-0.39	0.78	0.44	0.18	0.03	0.98	0.00

Note. Significance values below .05 are shown in bold.

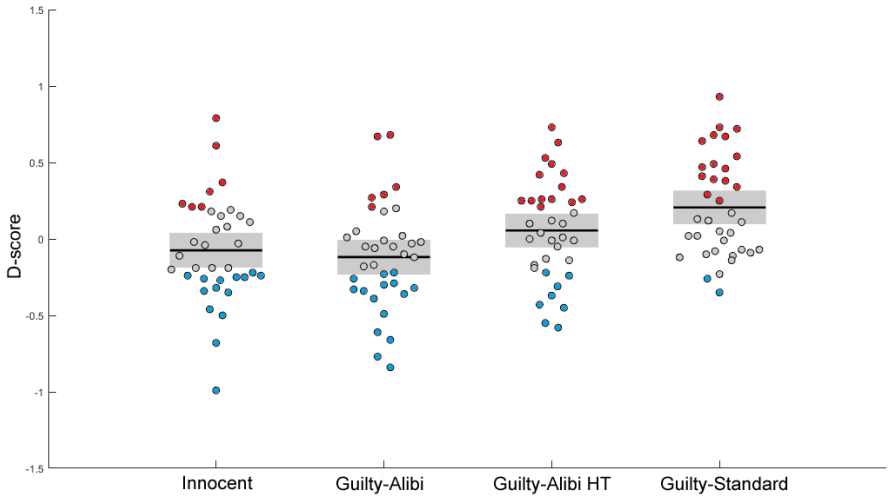

Figure 3.1. D-scores for the four groups on the aIAT contrasting the mock crime with the alibi/innocent act in Experiment 3. The black lines show the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the ring-related sentences are associated with the truth) and D-scores below zero suggest innocence (that the email-related sentences are associated with the truth). Scores consistent with guilt (>0.2) are marked with red dots, and scores consistent with innocence (<-0.2) are marked with blue dots. Grey dots indicate inconclusive scores.

As in the previous experiments and prior research, I also classified individuals with a positive score as "guilty" and individuals with a negative score as "innocent", after first excluding participants who scored too close to zero (absolute d-score between 0 – 0.2; excluded *N* for Innocent = 14, Guilty-Standard = 17, Guilty-Alibi = 15; and Guilty-Alibi with HT = 14). Then, I compared classification rates across the different groups. The results revealed that 84.2% of the Guilty-Standard group was correctly classified as guilty, whereas only 31.8% was classified as guilty in the Innocent group, which was significantly lower ($\chi^2(1) = 11.36$, $p = .001$, $\varphi = .53$). In Guilty-Alibi with HT group, guilty classification was 63.6% which was not significantly different from the Guilty-Standard group ($\chi^2(1) = 2.20$, $p = .138$, $\varphi = 2.32$), but it was significantly higher than in the Innocent group ($\chi^2(1) =$

4.46, p = .035, φ = .319). Guilt classification for the Guilty-Alibi group was 28.6% and was significantly lower than both Guilty-Standard ($\chi^2(1) = 12.48$, $p < .001$, φ = .558) and Guilty-Alibi with HT groups ($\chi^2(1) = 5.31$, p = .021, φ = -.351), but not different from the Innocent group ($\chi^2(1) = .054$, p = .817, φ = -.035).

However, the above classification is depending on a specific cut-off score, and it is difficult to know what is the most appropriate cut-off to use. Therefore, I also conducted ROC analysis to investigate classification performance using Area Under Curve (AUC). AUCs reflect the accuracy with which a randomly chosen participant can be classified into the correct group, where .5 reflects chance classification and 1.0 reflects perfect classification. This analysis showed that when comparing Innocent group to Guilty-Standard group, d-score classification was significantly better than chance (*AUC* = .72, *SE* = .060, p = .001). However, D-score classification was not accurate and not significant when compared Innocent to Guilty-Alibi group (*AUC* = .54, *SE* = .69, p .581), as well as when compared to Guilty-Alibi with HT group (*AUC* = .62, *SE* = .067, p = .073).

Reaction times and accuracy

Raw RT and accuracy (see Figure 3.2 and Table 3.2) were analysed separately to gain more insight on what aspects of behaviour contributed to the D-score differences. For RT, a 4 (Groups: Innocent, Guilty-Standard, Guilty-Alibi, and Guilty-Alibi with HT; between group) x 2 (Block: congruent and incongruent; within subject) mixed ANOVA showed that there were no main effect of neither group ($F(3,140) = 1.587$, p = .195, η_p^2 = .033) nor block ($F(1, 140) = .031$, p = .861, η_p^2 = .000). However, there was a significant group x block interaction ($F(3, 140) = 5.91$, p = .001, η_p^2 = .112). Follow-up paired t-tests showed significantly faster RTs in the

guilt congruent than the guilt incongruent block for the Guilty-Standard group ($t(35) = 3.06$, $p = .004$, $d = 0.46$), whereas the Guilty-Alibi group showed the reverse pattern with significant faster RTs in the guilt-incongruent compared to the guilt-congruent block ($t(35) = 2.52$, $p = .016$, $d = 0.30$). There were no RT differences between the two blocks in Innocent ($t(35) = 1.22$, $p = .231$, $d = 0.14$), and Guilty-Alibi with HT groups ($t(35) = 0.67$, $p = .510$, $d = 0.11$). Independent t-tests (Table 3.2) were conducted to compare the groups within each block. These results showed that the Guilty-Standard group was significantly faster in the guilt congruent block compared to Guilty-Alibi and Guilty-Alibi with HT group, whilst the other comparisons were not significant. There were also no significant RT differences in any groups for the guilt incongruent block.

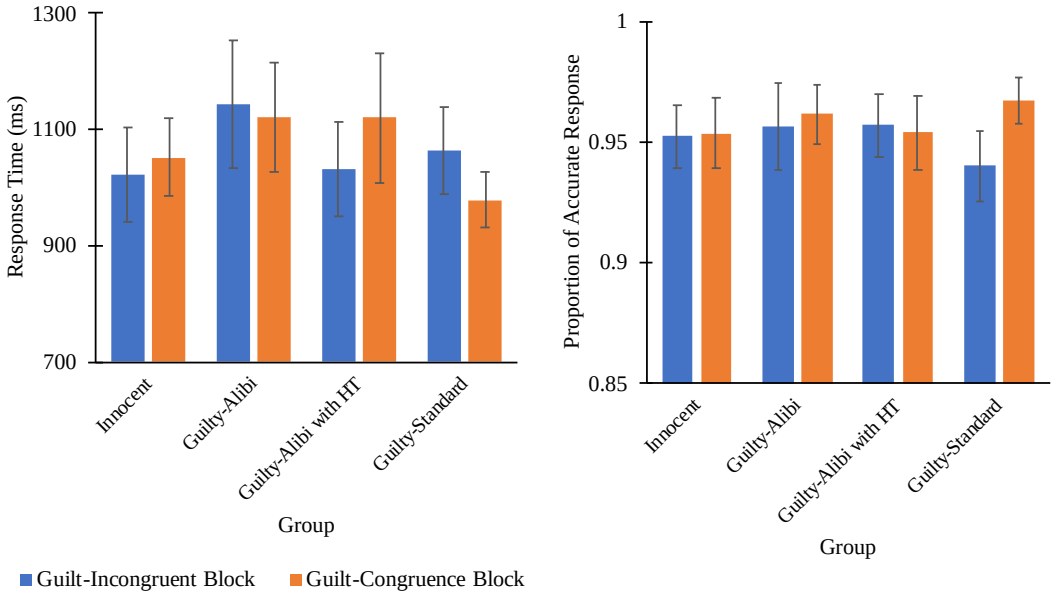

Figure 3.2. Proportion accurate responses and mean response times for guilt-incongruent (True+Email/False+Ring) and guilt-congruent (True+Ring/False+Email) blocks of the Mock Crime/Innocent event aIAT in Experiment 3. Error bars denote 95% confidence intervals.

For accuracy, a 4 (Groups: Innocent, Guilty-Standard, Guilty-Alibi, and Guilty-Alibi with HT; between groups) x 2 (Block: guilt congruent and guilt

incongruent; within subject) mixed ANOVA revealed a significant main effect of block ($F(1, 140) = 4.15$, $p = .044$, $\eta_p^2 = .029$), but no main effect of group ($F(3, 140) = .22$, $p = .880$, $\eta_p^2 = .005$). There was however an interaction between group and block ($F(3, 140) = 3.46$, $p = .018$, $\eta_p^2 = .069$). Follow-up paired t-tests indicated that Guilty-Standard group was more accurate in the guilt congruent block compared to the guilt incongruent block ($t(35) = 3.78$, $p = .001$, $d = .73$), whist there were no differences between blocks in Innocent ($t(35) = .108$, $p = .915$, $d = .02$), Guilty-Alibi ($t(35) = .38$, $p = .704$, $d = .07$), or Guilty-Alibi with HT group ($t(35) = .804$, $p = .427$, $d = .10$). There were no significant accuracy differences between groups in either guilt congruent or guilt incongruent blocks (see Table 3.2). Thus, similar to previous experiments, the strongest effects of the manipulation were on reaction times rather than accuracy, and the Standard Guilty group showed the expected effects on both measures most clearly (slower RT and lower accuracy in guilt incongruent than congruent blocks).

Faking analyses

As in the prior experiments, the faking index was also calculated to investigate unusual response pattern in the aIAT blocks, which could indicate guilt. I used the same faking-detection algorithm as in the previous chapter (pp. 39) to detect if response times in the single blocks were faster than in the critical double classification block, because fakers may try to intentionally slow down their responses in the double classification blocks.

In the aIAT version that contrasted the mock crime and alibi/innocent act directly in Experiment 3, there were no significant differences between Innocent ($M = 1.14$, $SD = 0.16$) and Guilty-Standard groups ($M = 1.10$, $SD = 0.19$; $t(70) = 0.93$, p

= .354, d = 0.22), Innocent and Guilty-Alibi groups (M = 1.15, SD = 0.19; t(70) = 0.41, p = .683, d = 0.10), Innocent and Guilty-Alibi with HT groups (M = 1.21, SD = 0.20; t(70) = 1.57, p = .121, d = 0.37), Guilty-Standard and Guilty-Alibi groups (t(70) = 1.25, p = .215, d = 0.30), or Guilty-Alibi and Guilty-Alibi with HT groups (t(70) = 1.11, p = .272, d = 0.26) in the average faking index. However, the average faking index was lower in the Guilty-Standard than Guilty-Alibi with HT group (t(70) = 2.31, p = .024, d = 0.54). Using the 1.08 cut-off as suggested by Agosta and colleagues (2011), 53% of Guilty-Standard, 67% of Guilty-Alibi, 69% of Innocent and 72% of Guilty-Alibi with HT group were classified as faking in the ring/email classification aIAT. These classification rates were not different (Guilty-Standard vs. Guilty-Alibi group: $\chi^2(1) = 1.44$, p = .230, φ = 0.142; Guilty-Standard vs. Guilty-Alibi with HT group: $\chi^2(1) = 2.90$, p = .088, φ = 0.201; Guilty-Alibi vs. Guilty-Alibi with HT: $\chi^2(1) = 0.262$, p = .61, φ = .060; Innocent vs. Guilty-Standard group: $\chi^2(1) = 2.10$, p = .147, φ = 0.171; Innocent vs. Guilty-Alibi group: $\chi^2(1) = 0.064$, p = .800, φ = .030; Innocent vs. Guilty-Alibi with HT group: $\chi^2(1) = .067$, p = .795, φ = 0.031)

Further, a threshold-independent ROC analysis was also conducted to evaluate faking classification performance. The ROC analyses showed that the classification was not different from chance when comparing Innocent with Guilty-Standard group (AUC = .54, SE = .07, p = .612), when comparing Innocent with Guilty-Alibi group (AUC = .56, SE = .07, p = .386), or when comparing Innocent with Guilty-Alibi with HT group (AUC = .60, SE = .068, p = .128). There were also no differences in classification performance between Guilty-Standard and Guilty-Alibi group (AUC = .57, SE = .068, p = .290) or between Guilty-Alibi and Guilty-Alibi with HT (AUC = .56, SE = .07, p = .356). However, the classification performance was just significantly better than chance when comparing Guilty-

Standard with Guilty-Alibi with HT group (*AUC* = .63, *SE* = .065, *p* = .050). Thus, faking analyses showed that when the aIAT contrasted the mock crime to the innocent/alibi event, rehearsing an alibi repeatedly over a week may cause unusual response patterns in the aIAT blocks, but this effect was rather weak and only significant when compared to a guilty standard group, and not compared to the other groups.

Ring/Exam aIAT

D-Score

Next, I analysed the Ring/Exam version of the aIAT which contrasted the mock crime (ring) with an event that none of the groups had experience nor knowledge of (exam), and was identical to the aIAT version used in Experiment 2. In this test, positive D-scores are indicative of guilt because they suggest that participants associate the mock crime with the truth, whereas D-scores around zero suggest that participants associate both events equally strongly with the truth (i.e. they associate either both, or neither event with the truth). Because none of the two events is indicative of innocence there is no result that would be diagnostic of innocence in this aIAT version, and no groups were predicted to show negative D-scores. In this test, there was only a trend towards differences between the groups in mean D-scores ($F(3, 140) = 2.50$, $p = .062$, $\eta_p^2 = 0.05$; see Figure 3.3), suggesting that this aIAT version did not discriminate between the groups as well as the Ring/Email aIAT (as would be expected since there should be less variability between groups when the test is designed to only produce scores either around zero or above, and no negative scores). The mean D-scores of Guilty-Standard ($t(35) = 3.98$, $p < .001$, $d = .66$) and Guilty-Alibi with HT group ($t(35) = 3.05$, $p = .004$, $d =$

.51) were significantly above zero (see Table 3.1 and Figure 3.1). However, the mean D-score for Innocent ($t(35) = -.17$, $p = .86$, $d = .03$) and Guilty-Alibi groups ($t(35) = 1.90$, $p = .066$, $d = .32$) were not significantly different from zero. Independent t-tests were conducted to investigate potential differences in mean D-scores between groups (see Table 3.3). Results revealed that the Innocent group scored significantly lower than both the Guilty-Standard group and the Guilty-Alibi with HT group. However, there were no differences in mean D-scores between Innocent and Guilty-Alibi groups, Guilty-Standard and Guilty-Alibi groups, and Guilty-Alibi and Guilty-Alibi with HT groups.

As previously, I conducted individual classification after excluding participants who obtained d-scores too close to zero (excluded: Innocent = 15, Guilty-Standard = 15, Guilty-Alibi = 19, Guilty-Alibi with HT = 10). This revealed that 81% of the remaining Guilty-Standard group were correctly identified as guilty whereas less than half (47.6%) of the Innocent group were identified as guilty, as would be expected. On the other hand, 58.8% of Guilty-Alibi and 65.4% of Guilty-Alibi with HT were correctly classified as guilty. The classification rate of Guilty-Standard group was significantly higher than Innocent group ($\chi^2(2) = 5.08$, $p = .024$, $\varphi = .348$). However, there were no significant difference in any other comparison (Innocent vs. Guilty-Alibi ($\chi^2(1) = .473$, $p = .492$, $\varphi = .112$) ; Innocent vs. Guilty-Alibi with HT ($\chi^2(1) = 1.50$, $p = .221$, $\varphi = .179$); Guilty-Standard vs. Guilty-Alibi ($\chi^2(1) = 2.24$, $p = .135$, $\varphi = .243$); Guilty-Standard vs. Guilty-Alibi with HT $\chi^2(1) = 1.41$, $p = .236$, $\varphi = .173$); Guilty-Alibi vs. Guilty-Alibi with HT ($\chi^2(1) = .189$, $p = .662$, $\varphi = .066$).

Table 3.3. Independent t-test results comparing group performance on the Ring/Exam aIAT version

Variable	D-score			RT						ACC					
				Guilt-congruent			Guilt-incongruent			Guilt-congruent			Guilt-incongruent		
	t	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
Innocent x Guilty-Standard	**2.59***	**0.01**	**-0.61**	0.81	0.42	0.19	1.48	0.14	-0.35	0.19	0.85	-0.05	**1.96***	**0.05**	**0.46**
Innocent x Guilty-Alibi	1.48	0.14	-0.35	0.68	0.5	0.16	0.48	0.64	-0.11	0.27	0.79	-0.06	1.3	0.2	0.31
Innocent x Guilty-Alibi with HT	**2.19***	**0.03**	**-0.52**	0.14	0.89	-0.03	1.64	0.11	-0.39	1.61	0.11	-0.38	0	1	0
Guilty-Standard x guilty- alibi	0.9	0.37	-0.21	0.04	0.97	0.22	0.91	0.37	0.11	0.07	0.95	0.36	0.07	0.95	0.44
Guilty-Standard x Guilty-Alibi with HT	0.27	0.79	-0.06	0.93	0.36	0.01	0.47	0.64	-0.22	1.55	0.13	0.01	1.87	0.07	0.28
Guilty-Alibi x Guilty-Alibi with HT	0.61	0.54	-0.14	0.8	0.43	-0.19	1.17	0.25	-0.28	1.61	0.11	-0.38	1.12	0.27	-0.26

Note. Significance values below .05 are shown in bold.

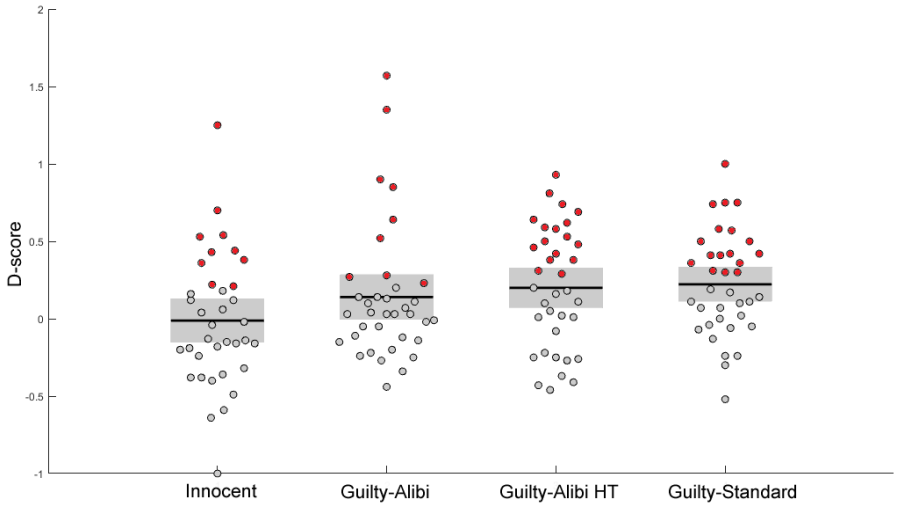

Figure 3.2. D-scores for the four groups on the aIAT contrasting the mock crime with the unexperienced exam event in Experiment 3. The black lines shows the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest guilt (that the ring-related sentences are associated with the truth). D-scores close to zero suggest that the events were equally associated with the truth, but because the test did not include a truly "innocent" event, innocence cannot be classified in this aIAT version. Scores consistent with guilt (>0.2) are marked with red dots, and inconclusive scores (<0.2) are marked in grey.

Next, ROC analyses were conducted to further investigate guilt classification performance with AUC independent of the cut-off point. The analysis showed that D-score classification performance was above chance when comparing the Innocent and Guilty-Standard groups (*AUC* = .68, *SE* = .064, *p* = .009) and when comparing the Innocent and Guilty-Alibi with HT groups (*AUC* = .62, *SE* = .065, *p* = .037). However, classification performance was less accurate when comparing the Innocent and Guilty-Alibi groups (*AUC* = .59, *SE* = .068, *p* = .207). Thus, the D-score analysis in the ring/exam aIAT version showed a similar result to the ring/email version – both Guilty-Standard and Guilty-Alibi with HT groups were detected as guilty compared to the Innocent group, whereas the Guilty-Alibi group who did not receive home training appeared less guilty.

Reaction times and accuracy

Please refer to Table 3.1 for mean scores and standard deviations of RT and accuracy. For RTs, a 4 (group: innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with HT; between groups) x 2 (block: congruent, incongruent; within subject) mixed ANOVA showed significant main effect of block ($F(1, 140) = 13.70$, $p < .001$, $\eta_p^2 = .089$), but no main effect of group ($F(3,140) = 0.54$, $p = .652$, $\eta_p^2 = .012$). However the group x block interaction was significant ($F(3, 140) = 3.30$, $p = .022$, $\eta_p^2 = .066$; see Figure 3.4A). Follow up paired t-tests revealed significant reaction time difference between congruent and incongruent block in Guilty-Standard ($t(35) = 4.09$, $p < .001$, $d = .539$) and Guilty-Alibi with HT group ($t(35) = 2.53$, $p = .016$, $d = .277$). However, there were no significant reaction time difference between blocks in Innocent ($t(35) = .635$, $p = .530$, $d = .074$) and Guilty-Alibi group ($t(35) = 1.71$, $p = .096$, $d = .20$). However, when comparing the groups directly, results showed no differences between groups in either congruent nor incongruent blocks (Table 3.3).

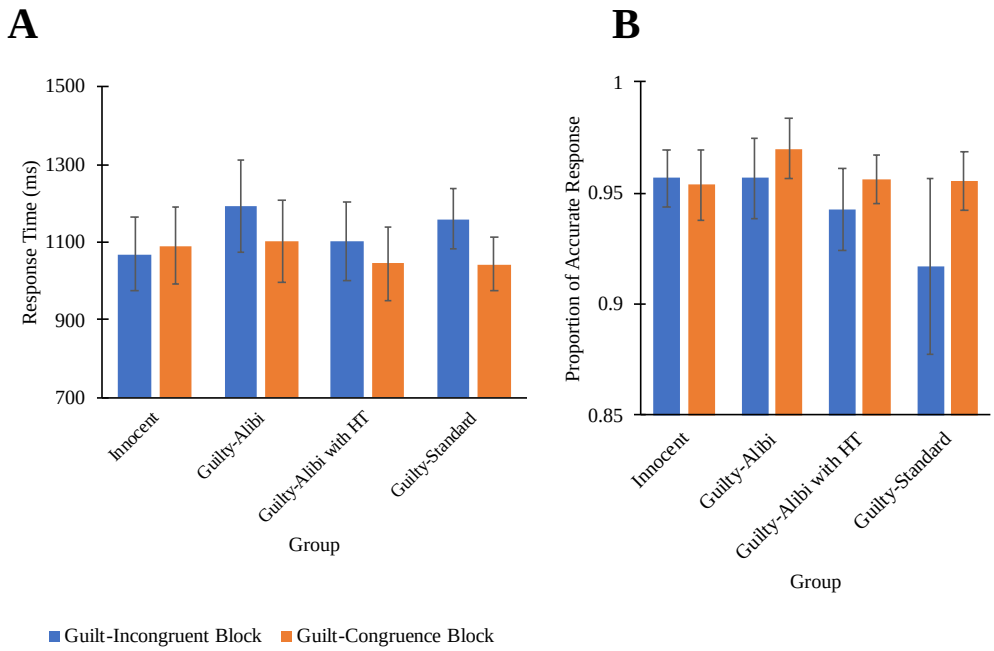

Figure 3.3. Proportion accurate responses and mean response times from guilt-incongruent (True+Exam/False+Ring) and guilt-congruent (True+Ring/False+Exam) blocks of the Mock Crime/Unexperienced event aIAT in Experiment 3. Error bars denote 95% confidence intervals.

A 4 (group: Innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with HT; between groups) x 2 (block: guilt-congruent, guilt-incongruent; within subject) mixed ANOVA were also analysed to examine accuracy (Figure 3.4B). Results suggested that were a significant main effect of block ($F(1, 140) = 7.52$, $p = .007$, $\eta_p^2 = .051$), but no main effect of group ($F(3, 140) = .129$, $p = .099$, $\eta_p^2 = .044$) nor interaction effect ($F(1, 140) = 2.275$, $p = .083$, $\eta_p^2 = 046$). Paired t-tests showed significant accuracy differences between blocks in the Guilty-Standard group ($t(35) = 2.13$, $p = .040$, $d = .446$), but not in Innocent ($t(35) = .415$, $p = .681$, $d = .073$), Guilty-Alibi ($t(35) = 1.66$, $p = .106$, $d = .304$), nor Guilty-Alibi with HT groups ($t(35) = 1.61$, $p = .118$, $d = .283$). Independent t-tests were conducted to investigate accuracy differences between groups in guilt-congruent and guilt-incongruent blocks (Table 3.3). These showed only a trend towards a difference ($p = .054$) in the guilt-incongruent block when comparing Innocent and Guilty-Standard groups, and no other differences between groups in guilt-congruent nor guilt-incongruent blocks. Thus, as in the Mock Crime/Innocent event aIAT version and the previous experiments, the strongest and most consistent effects on RT and accuracy were in the standard guilty group.

Faking analyses

Faking analyses were conducted to also investigate possible unusual response patterns in the mock crime vs. unexperienced event aIAT version. Results showed that there were no differences between Innocent ($M = 1.12$, $SE = .21$) and Guilty-Standard ($M = 1.14$, $SE = 0.19$; $t(70) = 0.475$, $p = .636$, $d = 0.11$), Innocent and Guilty-Alibi ($M = 1.18$, $SE = 0.22$; $t(70) = 1.14$, $p = .259$, $d = 0.27$), or Innocent and Guilty-Alibi with HT groups ($M = 1.13$, $SE = 0.18$; $t(70) = 0.22$, $p = .827$, $d = 0.05$)

in the average faking index. There were also no differences in faking index between Guilty-Standard and Guilty-Alibi ($t(70) = 0.716$, $p = .476$, $d = 0.17$), Guilty-Standard and Guilty-Alibi with HT ($t(70) = 0.279$, $p = .781$, $d = 0.07$), or Guilty-Alibi and Guilty-Alibi with HT groups ($t(70) = 0.992$, $p = .325$, $d = 0.23$).

Using the 1.08 classification cut-off as suggested by Agosta et al. (2011), 56% of Guilty-Standard group and 64% of Guilty-Alibi group were classified as faking and this was not significantly different ($\chi^2(1) = 0.52$, $p = .471$, $\varphi = 0.085$). There was also no difference between Guilty-Standard and Guilty-Alibi with HT (72%; $\chi^2(1) = 2.17$, $p = .141$, $\varphi = 0.173$), and Guilty-Standard and Innocent group (also 56%, so both groups were the same). There was also no significant difference when comparing Innocent to Guilty-Alibi group ($\chi^2(1) = 0.52$, $p = .471$, $\varphi = 0.085$), Innocent to Guilty-Alibi with HT group ($\chi^2(1) = 2.17$, $p = .141$, $\varphi = 0.173$), and Guilty-Alibi and Guilty-Alibi with HT group ($\chi^2(1) = 0.58$, $p = .448$, $\varphi = 0.089$) in faking classification at this threshold.

Threshold independent ROC analysis showed that faking classification was not different from chance when comparing Innocent and Guilty-Alibi group ($AUC = .57$, $SE = .068$, $p = .280$), Innocent and Guilty-Alibi with HT ($AUC = .54$, $SE = .069$, $p = .551$), Innocent and Guilty-Standard ($AUC = .53$, $SE = .069$, $p = .693$) and Guilty-Alibi and Guilty-Alibi with HT ($AUC = .55$, $SE = .07$, $p = .471$). When compared to Guilty-Standard group, the classification of Guilty-Alibi group ($AUC = .56$, $SE = .068$, $p = .375$) and Guilty-Alibi with HT ($AUC = .52$, $SE = .069$, $p = .787$) as fakers was also at chance. Thus, according to the faking index all of the groups showed equal amounts of unusual slowing in double classification blocks in this aIAT version.

Email/Exam aIAT

D-score

Next, I analysed the Email/Exam version of the aIAT which contrasted the innocent/alibi event (involving writing an email) with an event that none of the groups had experience nor knowledge of (exam) in order to assess whether the innocent/alibi event would be detected as true for any of the groups. That is, would learning and rehearsing a false alibi lead that scenario to be detected as true, or would it only be detected as true for the Innocent group who had actually conducted the act? In this test, positive D-scores are indicative of innocence because they suggest that participants associate the email event with the truth, whereas D-scores around zero suggest that participants associate both events equally strongly with the truth (i.e. they associate either both, or neither event with the truth). Because none of the two events is indicative of guilt there is no result that would be diagnostic of guilt in this aIAT version, and no groups were predicted to show negative D-scores. In this test, there was no overall significant difference between the groups in mean D-scores ($F(3, 140) = 1.95$, $p = .124$, $\eta_p^2 = 0.04$; see Figure 3.4), suggesting that this aIAT version did not discriminate between the groups well. The mean D-score of the Guilty-Standard group was not different from zero ($t(35) = 0.11$, $p = .915$, $d = 0.02$) as expected, since this group had no knowledge of either event. In contrast, the Guilty-Alibi ($t(35) = 2.28$, $p = .029$, $d = 0.38$) and Guilty-Alibi with HT groups ($t(35) = 2.23$, $p = .033$, $d = 0.37$) did score significantly above zero, suggesting that the alibi was detected as if true on average in these groups. Surprisingly however, the Innocent group's mean D-score was not significantly above zero ($t(35) = 0.40$, $p = .691$, $d = 0.07$), showing a failure of the test to detect the innocent event even though it was actually true for that group. Comparing differences in mean D-score between

groups using independent t-tests, there were non-significant trends towards more positive D-scores in the two Alibi groups than in the Guilty-Standard group but none of the other differences approached significance (see Table 3.4).

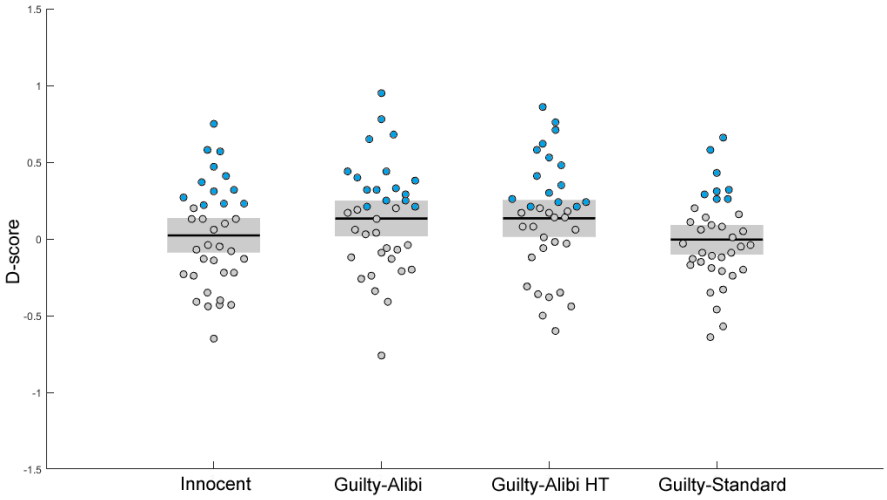

Figure 3.4. D-scores for the four groups on the aIAT contrasting the alibi/innocent act with the unexperienced exam event in Experiment 3. The black lines show the mean score and the grey boxes show the 95% confidence intervals of the mean. D-scores above zero suggest innocence (that the email-related sentences are associated with the truth). D-scores close to zero suggest that the events were equally associated with the truth, but because the test did not include a truly "guilty" event, guilt cannot be classified in this aIAT version. Scores consistent with innocence (>0.2) are marked with blue dots, and inconclusive scores (<0.2) are marked in grey.

Individual classification after excluding participants who scored too close to zero (excluded: Innocent = 11, Guilty-Standard = 20, Guilty-Alibi = 13, and Guilty-Alibi with HT = 13) suggested that only 50% of the Innocent group was correctly identified as associating email-related actions with the truth, whereas 69.9% of the Guilty-Alibi with HT group and 73.9% of the Guilty-Alibi group were identified as associating email-related action more with the truth when compared to exam-related action. Out of the Guilty-Standard group, only 56.3% participants were detected to associate email-related action more with the truth, as would be expected since this group had no knowledge of either of the events. However, there were no significant

differences among the groups in classification rates: (Innocent vs. Guilty-Alibi: $\chi^2(1) = 2.84$, $p = .092$, $\varphi = .246$; Innocent vs. Guilty-Alibi with HT: $\chi^2(1) = 1.87$, $p = .172$, $\varphi = .20$; Guilty-Standard vs. Guilty-Alibi: $\chi^2(1) = 2.24$, $p = .135$, $\varphi = .243$; Guilty-Standard vs. Guilty-Alibi with HT: $\chi^2(1) = 1.325$, $p = .250$, $\varphi = .184$; Guilty-Alibi vs. Guilty-Alibi with HT: $\chi^2(1) = .107$, $p = .743$, $\varphi = .048$).

ROC analyses were further conducted to investigate D-score classification performance independent of a specific cut-off. The results revealed that email/exam d-score classification was not accurate at all. Comparing the Innocent group with the Guilty-Standard group, classification performance was at chance (*AUC* = .52, *SE* = .069, *p* = .787), and it was only slightly better (non-significantly so) when comparing Innocent participants to Guilty-Alibi (*AUC* = .59, *SE* = .067, *p* = .177) and Guilty-Alibi with HT (*AUC* = .59, *SE* = .068, *p* = .169). Thus, D-scores indicated very poor detection of the participants who had actually performed the innocent act, whereas imagining a false alibi seems to have slightly increased detection of this false scenario as true in the two Alibi groups, in that their d-scores were significantly above zero. However, since the groups were not significantly different from each other in d-scores, this slight increase in the Alibi groups is difficult to interpret and not very reliable.

Table 3.4. Independent t-test results comparing group performance on the Email/Exam aIAT version

Variable	D-score			RT						ACC					
				Innocence-congruent			Innocence-incongruent			Innocence-congruent			Innocence-incongruent		
	t	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
Innocent x Guilty-Standard	0.37	0.71	0.09	0.22	0.83	0.05	0.53	0.6	0.13	1.12	0.27	0.27	0.56	0.58	-0.13
Innocent x Guilty-Alibi	1.37	0.18	-0.33	0.04	0.97	0.01	0.8	0.43	-0.19	0.88	0.38	-0.21	0.54	0.59	0.13
Innocent x Guilty-Alibi with HT	1.35	0.18	-0.32	0.77	0.45	-0.18	1.57	0.12	-0.38	0.66	0.52	-0.16	0.98	0.33	-0.24
Guilty-Standard x guilty- alibi	1.82	0.07	0.44	0.16	0.87	0.04	1.39	0.17	0.33	**2.20***	**0.03**	**0.52**	-0.93	0.36	-0.22
Guilty-Standard x Guilty-Alibi with HT	1.8	0.08	0.43	1.09	0.28	0.26	**2.40***	**0.02**	**0.57**	1.65	0.11	0.39	0.46	0.65	0.11
Guilty-Alibi x Guilty-Alibi with HT	0.01	0.99	0	0.79	0.43	0.19	0.64	0.52	0.15	0.03	0.98	-0.01	1.22	0.23	0.29

Note. Significance values below .05 are shown in bold.

Reaction Time and Accuracy

Please refer to Table 3.1 for mean scores and standard deviation of RT and accuracy. For RTs (Figure 3.5), a 4 (group: Innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with HT; between groups) x 2 (block: congruent, incongruent; within subject) mixed ANOVA showed no main effect of block ($F(1, 140) = 2.47$, $p = .118$, $\eta_p^2 = .017$), main effect of group ($F(3, 140) = 1.00$, $p = .394$, $\eta_p^2 = .021$), nor block x group interaction ($F(3, 140) = 1.62$, $p = .188$, $\eta_p^2 = .034$). Follow up paired t-tests comparing the blocks within each groups showed no significant differences in RT between innocence-congruent and innocence-incongruent blocks in Innocent ($t(35) = 0.05$, $p = .960$, $d = 0.01$), Guilty-Standard ($t(35) = 0.64$, $p = .527$, $d = 0.09$), Guilty-Alibi ($t(35) = 1.61$, $p = .116$, $d = 0.19$), or Guilty-Alibi with HT group ($t(35) = 1.83$, $p = .075$, $d = 0.19$). Independent t-tests also showed no differences between groups across either congruent or incongruent blocks, except RT in the innocence-incongruent block, where the Guilty-Alibi with HT group was significantly slower than the Guilty-Standard group (see Table 3.4).

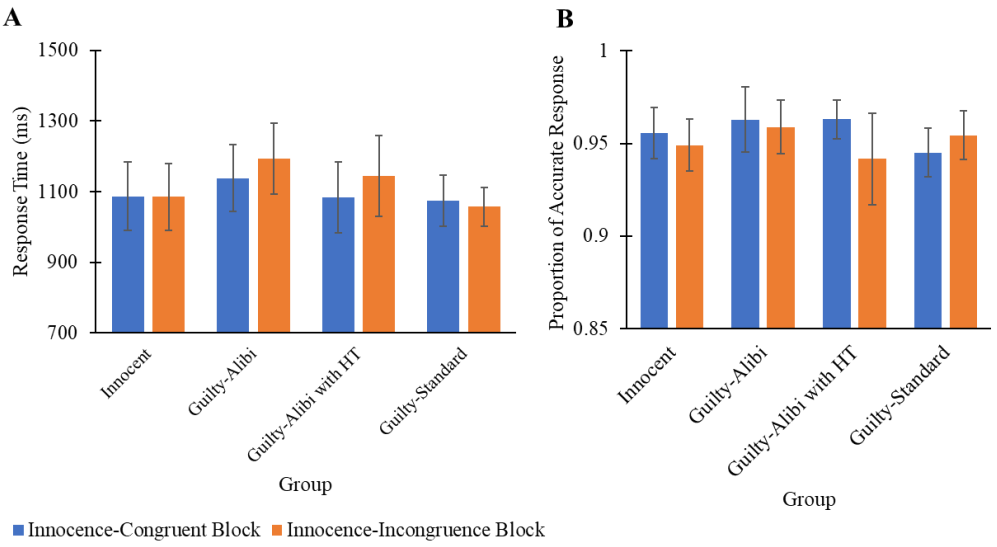

Figure 3.5. Proportion accurate responses and mean response times from innocence-incongruent (True+Exam/False+Email) and innocence-congruent (True+Email/False+Exam) blocks of the Innocent/Unexperienced event aIAT in Experiment 3. Error bars denote 95% confidence intervals.

For accuracy (Figure 3.5), 4 (group: Innocent, Guilty-Standard, Guilty-Alibi, Guilty-Alibi with HT; between groups) x 2 (block: congruent, incongruent; within subject) mixed ANOVA showed no main effect of block ($F(1, 140) = 1.77$, $p = .185$, $\eta_p^2 = .012$), no main effect of group ($F(3, 140) = 0.55$, $p = .647$, $\eta_p^2 = .012$), nor block x group interaction ($F(3, 140) = 2.29$, $p = .081$, $\eta_p^2 = .047$). When comparing each block within the groups, there was no differences in accuracy between blocks in Innocent ($t(35) = 1.00$, $p = .326$, $d = 0.16$), Guilty-Standard ($t(35) = 1.19$, $p = .241$, $d = 0.24$), Guilty-Alibi ($t(35) = 1.61$, $p = .116$, $d = 0.19$), and Guilty-Alibi with HT group ($t(35) = 0.65$, $p = .519$, $d = 0.08$). When comparing the groups within each block, there were no differences between groups in either innocence-congruent or innocence-incongruent blocks, except that the Guilty-Alibi group was more accurate than the Guilty-Standard group in the innocence-congruent block (see Table 3.4). Thus, RT and accuracy differences between blocks and groups were very small and mostly non-significant in the email/exam version of the aIAT, consistent with the main D-score analysis.

Faking analyses

Faking analyses were conducted again to investigate possible unusual response patterns. In the innocent vs. unexperienced event version of the aIAT, there was no difference between Innocent ($M = 1.15$, $SD = .14$) and Guilty-Standard ($M = 1.15$, $SD = 0.16$; $t(70) = .03$, $p = .979$, $d = 0.01$), Innocent and Guilty-Alibi ($M = 1.18$, $SD = 0.17$; $t(70) = 0.66$, $p = .509$, $d = 0.16$), and Innocent and Guilty-Alibi with HT ($M = 1.11$, $SD = 0.19$; $t(70) = 1.11$, $p = .270$, $d = 0.26$) in mean faking score. There were also no differences in mean faking score between Guilty-Standard and Guilty-Alibi ($t(70) = 0.66$, $p = .513$, $d = 0.15$), Guilty-Standard and Guilty-Alibi with

HT ($t(70) = 1.04$, $p = .301$, $d = 0.25$), and Guilty-Alibi and Guilty-Alibi with HT ($t(70) = 1.60$, $p = .114$, $d = 0.38$).

Using the 1.08 cut-off (Agosta et al., 2011), 64% of Innocent group and 67% of Guilty-Standard group were classified as faking and these rates were not significantly different ($\chi^2(1) = 0.06$, $p = .804$, $\varphi = 0.03$), neither were Innocent and Guilty-Alibi groups (83%; $\chi^2(1) = 3.50$, $p = .061$, $\varphi = 0.22$), nor were Innocent and Guilty-Alibi with HT group (64% also). There were also no differences in faking classification between Guilty-Standard and Guilty-Alibi group ($\chi^2(1) = 2.67$, $p = .102$, $\varphi = 0.19$), Guilty-Standard and Guilty-Alibi with HT ($\chi^2(1) = 0.06$, $p = .804$, $\varphi = 0.03$), and Guilty-Alibi and Guilty-Alibi with HT ($\chi^2(1) = 3.50$, $p = .061$, $\varphi = 0.22$).

The ROC analyses showed that faking classification was not different from chance when comparing Innocent and Guilty-Standard group (*AUC* = .53, *SE* = .069, $p = .719$), Innocent and Guilty-Alibi group (*AUC* = .54, *SE* = .069, $p = .547$), and Innocent and Guilty-Alibi with HT (*AUC* = .55, *SE* = .069, $p = .451$). Classification performance was also at chance when comparing Guilty-Standard to Guilty-Alibi (*AUC* = .58, *SE* = .069, $p = .270$), Guilty-Standard to Guilty-Alibi with HT (*AUC* = .54 *SE* = .069, $p = .558$), and Guilty-Alibi to Guilty-Alibi with HT (*AUC* = .59, *SE* = .068, $p = .188$).

Post-Experiment Questionnaire Analysis

Results from the final questionnaire are shown in Table 3.5. The Innocent group rated their memory of the innocent act as less vivid than the three Guilty groups rated their memory for the mock crime act (Innocent vs. Guilty-Standard: $t(70) = 3.46$, $p = .001$, $d = 0.83$; Innocent vs. Guilty-Alibi: $t(70) = 3.39$, $p = .001$, $d = 0.81$; Innocent vs. Guilty-Alibi with HT: $t(70) = 4.45$, $p < .001$, $d = 1.06$) and they

also reported that they remembered fewer details of the act (Innocent vs. Guilty-Standard: ($t(70) = 4.42$, $p < .001$, $d = 1.06$; Innocent vs. Guilty-Alibi: $t(70) = 5.20$, $p < .001$, $d = 1.24$; Innocent vs. Guilty-Alibi with HT: $t(70) = 4.93$, $p < .001$, $d = 1.18$). The Innocent group also reported having been less nervous during the innocent act than the three Guilty groups were when they committed the mock crime (Innocent vs. Guilty-Standard: $t(70) = 2.80$, $p = .007$, $d = 0.67$; Innocent vs. Guilty-Alibi: $t(70) = 2.13$, $p = .037$, $d = 0.51$; Innocent vs. Guilty-Alibi with HT: $t(70) = 3.83$, $p < .001$, $d = 0.92$), and reported thinking about the innocent act less during the aIATs than the three Guilty groups thought about the mock crime during the aIATs (Innocent vs. Guilty-Standard: $t(70) = 3.85$, $p < .001$, $d = 0.92$; Innocent vs. Guilty-Alibi: $t(70) = 2.13$, $p = .037$, $d = 0.51$; Innocent vs. Guilty-Alibi with HT: $t(70) = 3.95$, $p < .001$, $d = 0.94$). There were no significant differences between the three Guilty groups on any of those questions (all ps >0.14).

The Alibi groups and the Innocent group were all more motivated to appear innocent on the aIATs than the Guilty-Standard group (Guilty-Standard vs. Innocent: $t(70) = 2.04$, $p = .045$, $d = 0.49$; Guilty-Standard vs. Guilty-Alibi: $t(70) = 3.09$, $p = .003$, $d = 0.74$; Guilty-Standard vs. Guilty-Alibi with HT: $t(70) = 2.83$, $p = .006$, $d = 0.68$), but did not differ between each other in levels of motivation (all ps >0.39). With regards to the alibi-specific questions, there were no differences between the Albi groups in terms of how much they were thinking of the alibi during the aIATs ($t(70) = 0.75$, $p = .46$, $d = 0.18$), but the Guilty-Alibi with HT group reported being able to imagine the alibi scenario in more details ($t(70) = 2.48$, $p = .016$, $d = 0.59$) and more vividly than the Guilty-Alibi group ($t(70) = 2.36$, $p = .021$, $d = 0.56$). Exploratory correlation analyses were also conducted to investigate whether any of

the self-report measures correlated with performance in the aIAT, but there were no significant correlations.

Table 3.5. Mean and standard deviations of self-reported ratings on the final questionnaire for the four groups. The scale had seven points (0-6), and lower scores always indicate less of the item being measured (e.g. less vividness/nervousness/motivation, etc.) and higher scores always indicate more of the item being measured (e.g. more vividness/nervousness/ motivation, etc.).

	Innocent	Guilty-Alibi	Guilty-Alibi with HT	Guilty-Standard
Remember detail of the act	3.39 (1.25)	4.64 (0.72)	4.64 (0.87)	4.53 (0.91)
Vividness of the act memory	3.50 (1.76)	4.36 (0.83)	4.69 (0.98)	4.44 (1.03)
Nervousness during the act	1.67 (1.29)	2.33 (1.37)	3.05 (1.76)	2.69 (1.79)
Thinking about the act during aIAT	1.58 (1.56)	2.50 (1.68)	3.11 (1.71)	3.08 (1.75)
Motivation to beat the aIAT	3.86 (1.50)	4.14 (1.22)	4.14 (1.50)	3.11 (1.58)
Imagine detail of the alibi	-	3.94 (1.33)	4.57 (0.70)	-
Vividness of the alibi imagination	-	3.89 (1.47)	4.57 (0.88)	-
Thinking about the alibi during aIAT	-	2.83 (1.68)	3.14 (1.78)	-

Note: the "act" refers to the act conducted in the first session (i.e. either mock crime or innocent act, depending on group).

So in sum, the questionnaire data from Experiment 3 suggested that the Innocent group had poorer memory of the innocent act than the Guilty groups' memory of the mock crime, whereas repeated and extended rehearsal of the alibi scenario in the Guilty-Alibi with HT group led to improved ability to imagine the alibi scenario when compared to the Guilty-Alibi group. Furthermore, the Innocent and Alibi groups were more motivated to appear innocent on the aIATs than the Guilty-Standard group.

Discussion

The aim of this study was to further investigate the effect of rehearsing alibi as a countermeasure on the aIAT, which has been suggested as an accurate, inexpensive, and practical way to assess criminal guilt in forensic applications. Previous research suggested that rehearsing an alternative scenario to what actually happened can interfere with and potentially weaken the true memory. In Experiment 3, I investigated whether learning and imagining a false alibi prior to the aIAT would impair the original memory for a mock crime and/or increase the implicit truth value of the alibi itself, and whether these effects would be particularly enhanced when the alibi was repeatedly rehearsed and imagined over an extended time period, in line with the literature on motivated forgetting that has shown gradual weakening of unwanted memories with repetition (see e.g. Anderson & Green, 2001; Anderson & Hanslmayr, 2014). To test these research questions, I included three versions of the aIAT. I contrasted 1) the mock crime vs. the innocent act (also used as a false alibi), 2) the mock crime vs. an unexperienced event, and 3) the innocent act (alibi) vs. an unexperienced event. This design allowed me to assess both the implicit truth of the mock crime (2) and the imagined alibi (3), as well as their relative truth value (1). Expanding on my previous two studies, I also added a time delay of one week between the mock crime/innocent act and aIAT tests and another experimental group, Guilty-Alibi with Home Training (HT), to assess the effect of repeatedly rehearsing an alibi over a longer period of time. The rationale for this new group was that this would be more realistic, as in real life a criminal suspect might come up with an alibi and practice it prior to the investigation, which would usually not occur immediately after the crime.

The results indicated that in the aIAT that tested the relative strength of the mock crime vs. innocent act/alibi, the mock crime was possible to detect after a week delay in Guilty-Standard participants. However, the aIAT could not distinguish which of the two events were true for Innocent participants nor for Guilty-Alibi with HT groups, and in the Guilty-Alibi group that did not receive home training, the test result was more indicative of innocence than guilt. In the aIAT that tested the relative strength of the mock crime vs. an unexperienced event, results suggested that the mock crime was possible to detect in Guilty-Standard and Guilty-Alibi with HT groups, while it was undetectable in Innocent and Guilty-Alibi groups. In the aIAT that tested the relative strength of the innocent/alibi act vs. an unexperienced event, results showed a trend towards detection of the innocent/alibi act as true in both Guilty-Alibi groups, but not in the Guilty-Standard and Innocent groups, however the results were weak in this aIAT version. In fact, in this study I could not detect which event was true for Innocent participants in any version of aIAT. Furthermore, although the test outcome of Guilty-Standard participants was in line with results from Experiments 1 and 2, their results were also somewhat weaker in this study. This weak detection may be related to the one week delay that I introduced between the initial crime/innocent act and the aIAT, compared to the previous two studies in which the test was administered immediately after the mock crime/innocent act, suggesting that time delay may reduce memory detection as found in previous literature (Gronau et al., 2015). Nevertheless overall, the results of this study support my conclusions in the previous chapter that rehearsing a false alibi before an aIAT may distort the test results, but clearly this depends on how the alibi countermeasure is used and also depends on how the aIAT is set up.

The strongest effect of the alibi countermeasure was in the Guilty-Alibi participants who learned and imagined a fabricated alibi one week *after* the mock crime and just prior to the test, without repeated rehearsal. In this group, the results suggested that they associated the imagined false alibi event more with the truth relative to the objectively true mock crime event, and also relative to an unexperienced event. Moreover, the aIAT that contrasted the mock crime with an unexperienced event was not able to distinguish which of the two events was true for these guilty participants, suggesting that the mock crime memories may have been weakened in this group. Thus, the effect of the alibi countermeasure in this group was even stronger than the findings in my previous studies, where the alibi group did not show significant associations between the alibi and truth (Experiment 1) and they also showed evidence of associating the mock crime with truth when contrasted with the unexperienced event (Experiment 2). One possible explanation for this phenomenon could be the enhanced salience of the alibi compared to the mock crime. Mental simulation of the alibi event just before the aIAT may have caused this imagined event to be more salient than the true memory of the mock crime, which may have been weaker in this experiment than in prior studies due to the long time delay between the event and the test. Therefore, rehearsing the false alibi might have strengthened alibi-related thoughts and inhibited the original memory from being triggered. More specifically, when sentences in the aIAT triggered mock crime related memories, the false alibi thoughts may have automatically inhibited responses to mock crime sentences and facilitated responses to false alibi sentences instead. As a consequence, participants result indicated that they associated the alibi more with the truth and the mock crime was not able to be detected.

Surprisingly, a different result pattern was observed in Guilty-Alibi with HT participants. I predicted that extended rehearsing of an imagined alibi for a week prior to the test would be particularly effective at inducing retroactive interference or competitive inhibition of the true memory (Anderson & Hanslmayr, 2014; Gronau et al., 2015). and that this group would therefore be most likely to appear innocent. However, Guilty-Alibi with HT participants who imagined a fabricated alibi for a week before the test were not completely effective at appearing innocent. The aIAT was unable to determine which of the two events were true when comparing the mock crime with the innocent/alibi act, which shows that the false alibi reduced guilt detection, but the alibi was not detected as relatively more truthful than the mock crime (as in the other alibi group without home training). It may be that when Alibi with HT participants were completing this aIAT version, their knowledge about these two events were competing for resources and causing neither block to be truly congruent (i.e. they experienced response conflict in both aIAT blocks), therefore slowing reaction times in both blocks. Therefore, they appeared to associate both events equally to the truth in this aIAT version. Moreover, participants in this Alibi group also seemed to associate the innocent/alibi act more with the truth when it was contrasted to an unexperienced event in the aIAT. This result suggests that imagining and rehearsing an alibi enhanced its implicit truth value, and is in line with previous research. Shidlovski, Schul, and Mayo (2014) found that imagining an event can increases the implicit truth value of the imagined event, even though people acknowledged explicitly that the event was not true. However, in the Alibi with HT group, the original memory of the mock crime was still detected when the aIAT contrasted the mock crime versus an unexperienced event, suggesting that extensive and repeated rehearsal of a false alibi did not impair the original mock crime

memory. Unlike Guilty-Alibi participants, Guilty-Alibi with HT participants learned, rehearsed and imagined a fabricated alibi immediately after the mock crime and also continuously rehearsed the alibi for 7 consecutive days before the test. This repeated alibi rehearsal might have strengthened participants' true memory by reactivating the mock crime details, leading to a facilitation compared to the other alibi group who received no such reminders. Therefore, the mock crime was detected in the Guilty-Alibi with HT group, but not in the Guilty-Alibi group.

Furthermore, previous research had suggested an algorithm that can be used for detecting if the participant is trying to fake their responses (Agosta, Ghirardi, et al., 2011). Consistent with Experiment 2, this faking index was not able to detect the differences between participants who adopted a countermeasure and those who did not in any of the aIAT versions in Experiment 3. Although this faking index could detect unusual response patterns in Experiment 1, the classification rate was not as high as in the literature (only 60%). This result across studies indicates that the faking index might not be very reliable in detecting fakers and should be used with caution.

The aIAT is proposed as an implicit measure of truth, which is claimed to be over 90% accurate in evaluating which of the two contrasting events is true. However, my findings suggest that an imagined event can be detected as a true memory when it is not, and even when participants know it is not true. This indicates that the aIAT may not actually measure implicit associations between the event and the truth, on the contrary it may measure the relative salience of events, such that the detected event is not necessarily the true memory, but it could be any event that participants acknowledged. Moreover, even in the most optimal conditions (my Guilty-Standard group) my detection rates were not nearly as accurate as those found

by Sartori et al., and the one-week delay in this Experiment seemed to lower detection rates compared to in my previous studies. Furthermore, although my standard guilty group was generally detected as guilty, the objectively true event for the Innocent group was not possible to detect. This latter group also rated their memories of the innocent act as less vivid and detailed compared to the guilty group's ratings of their mock crime memories. This finding is interesting as it points towards a role of subjective memory quality in aIAT accuracy – the test may only be able to detect memories that are subjectively detailed and vivid, and any factors that reduce memory quality may also reduce the test's effectiveness. This issue also suggests limitations of laboratory studies that try to investigate memory detection since memories of mock crimes may differ substantially from real criminal memories, for example in their emotional content, and higher emotional arousal is known to enhance the subjective vividness of memories and their durability over time (Kensinger, 2009). Real criminals of course also differ in their motivation to beat the test, which is likely to be relevant. Future research should consider investigating the accuracy of the aIAT when contrasting a real, emotionally arousing autobiographical memory with an alternative scenario to the memory to further investigate the effect of counterfactual thinking in aIAT.

To conclude, the results of Experiment 3 converge with my previous findings to confirm that the aIAT is very vulnerable to countermeasures that involve imagining an alternative scenario before the test, which could be problem in real life applications as criminal suspects may came up with a false alibi and practice this prior to the test. The results also suggest that the aIAT cannot be used as a simple and direct measure of truth, because it seems to measure something rather different. This is a fundamental problem for using the aIAT in real criminal cases – if

researchers do not know what the test is measuring, how can using the test be justified when a false result may have dire real-life consequences? Clearly, practical applications of the aIAT is premature until further research has clarified what the test actually measures, and in what situations it will produce accurate results.

Chapter 4: The effect of imagining a false alibi on the concealed information test

In Experiments 1-3, I investigated the effect of rehearsing a fabricated alibi on the aIAT (Sartori, et al., 2008). I found that the aIAT is very susceptible to a false alibi countermeasure: rehearsing an alibi prior to the test can reduce aIAT effectiveness such that it can fail to detect an objectively true mock crime as true, which seemed to be primarily driven by increased implicit truth value of the alibi itself. Importantly, this result was not due to participants experiencing source confusion about which event was true (Takarangi et al., 2015), because participants knew that the alibi was false (*cf*. Shidlovski et al., 2014). Thus, the aIAT seems unable to distinguish an objectively true memory *that the suspect knows is true* from a counterfactual version of the event *that the subject knows is false*. This could result in a misleading conclusion that a perpetrator has not committed a crime and that the false alibi he/she gives is true, which would be detrimental in real life criminal investigations. The results of these experiments raise serious questions about what exactly the aIAT is measuring. For example, the aIAT could measure the relative salience of two events, regardless of truth value (as discussed in Shidlovski et al., 2014; and similar concerns have been raised regarding the original implicit association test, e.g. Blanton & Jaccard, 2006)**.** Because of these uncertainties regarding what the aIAT is measuring, I decided to use another more direct method for testing the existence of incriminating memories in order to assess whether counterfactual imagination of a false alibi can reduce the accuracy of concealed memory detection tests.

In this chapter I will focus on the main cognition neuroscience technique used in memory detection research, namely the ERP-based concealed information test (CIT).

As discussed in Chapter 1, the CIT was developed to examine physiological responses to concealed information from the crime suspect (Lykken, 1960). It has been used in combination with various autonomic measures, such as electrodermal measures, respiratory measures, and cardiovascular measures, and brain activity measures like electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI). Among these different measures, P300-based Event-Related Potentials (ERPs) seems to outperform other measures in accuracy at discriminating between guilty and innocent suspects, and is a non-invasive technique that is relatively cheap, and easy to implement (Ben-Shakhar, 2012; Meijer et al., 2014). The P300 ERPs component is a brain marker that is elicited when people recognise a rare meaningful stimulus (Farwell & Donchin, 1991; Nasman, Whalen, Cantwell, & Mazzeri, 1987). In the CIT, enlarged P300 signals in response to critical crime-related information ("probes"; information that is known only to the guilty suspect) when compared to control items ("irrelevants") is used as an indication of guilt (Mertens & Allen, 2008; Rosenfeld et al., 2004).

A newly developed P300-based CIT version named the complex trial protocol (CTP) has been claimed to be more resistant to countermeasures and able to detect concealed information at a higher rate than other CIT versions (Rosenfeld et al., 2013, 2008). These studies showed that the CTP was robust against various physical (e.g. participants pressing their thumb on their leg) and mental countermeasures (e.g. participants counting in reverse). However, a recent study showed that the CTP was vulnerable to participants suppressing retrieval during the test, which reduced P300 amplitude for incriminating information and suggesting that incriminating memories can be voluntarily inhibited (Hu et al., 2015). Most research on the CIT has only investigated if it can resist countermeasures that suspects adopt during the test, but as

I argue in this thesis, countermeasures can also be engaged to modify memories in advance of memory detection tests, and this might be a particularly effective strategy for appearing innocent because it does not require engagement of complicated strategies during the test, which might reveal the suspect's guilt if they can be detected (e.g. the faking index used in Experiments 1-3, Agosta et al., 2013). Interestingly, simply generating false information to interfere with the crime-related memory can attenuate CIT detection accuracy in when used with skin conductance and other autonomic measures (Gronau et al., 2015). Since the CIT is a relatively direct test of whether critical information is recognised as meaningful, this finding suggests that the false information may interfere with true memories and somehow weaken them, perhaps by retrieval inhibition (cf. Retrieval-Induced Forgetting or "RIF", Anderson et al., 1994; as discussed previously in this thesis) However, autonomic measures such as skin conductance responses are quite far removed from the brain processes that give rise to recognition, and little is known regarding the effect of rehearsing an alibi on memory-related brain activity, as addressed in my next experiment.

Experiment 4

In the current study, I assessed if repeatedly rehearsing and imagining a false alibi will enable people to "beat" brain-activity based memory detection. Specifically, I investigated the extent to which people can modify crime-related memories through rehearsing an alibi using a P300-based CIT, with the so called "countermeasure-resistant" CTP version (Rosenfeld et al., 2008). I adopted a similar procedure as in previous studies described in this thesis, by first asking participants to commit a mock crime and then asking them to take the memory detection test. However, in this study I only assessed potential differences in memory detection

among three guilty conditions that all conducted a mock crime as the first step. Next, a "Guilty-Immediate" group undertook the CIT straight afterwards within the same session, whereas a "Guilty-Delay" group was dismissed after the mock crime, then returned seven days later for taking the CIT. These two guilty groups were compared against a "Guilty-Alibi with home training (HT) group", who practiced rehearsing an alibi from home once per day for the intervening days, before returning after seven days to take the CIT. The alibi manipulation in this Experiment 4 was thus identical to the "Guilty-Alibi with Home Training" group in Experiment 3, which I found to be less effective than an alibi countermeasure that was applied only once just before the aIAT. Nevertheless, I wanted to assess this version of the alibi manipulation since it is more in line with real life situations, when the suspect is more likely to be interrogated sometime after the actual event than immediately after the event, and may therefore repeatedly practice their false alibi. Since the prior experiments also cast doubts on what exactly the aIAT is measuring, I wanted to assess whether this ecologically valid countermeasure might be effective against an alternative, more direct measure of concealed crime memories. Finally, since repeated rehearsal of false information should theoretically be more likely to inhibit the true memories through a process like RIF (Anderson et al., 1994), I wanted to maximise the likelihood that the alibi manipulation would result in retrieval inhibition. The purpose of including both an immediate and delay guilty group was to assess whether guilt detection in the absence of countermeasures would deteriorate over time, since there was some indication in Experiment 3 that the mock crime and innocent memories may have been weaker after one week's delay (for example, since the single session alibi manipulation was more effective in Experiment 3 than

Experiments 1 and 2, and since the innocent event could not be detected as true in Experiment 3, but it could be detected as true in Experiment 1).

This Experiment 4 thus addressed both the effect of simple passage of time and the effect of a false alibi countermeasure on concealed memory detection with the P300-based CTP. I hypothesised that if the alibi measure was effective at weakening true memories of the crime, participants who were asked to rehearse an alibi with home training should show attenuated P300 differences between crime probes versus irrelevant items, when compared to the other groups. I also hypothesised that the probe-irrelevant difference should be more pronounced in the immediate group compared to the other two groups because time delay between the mock crime and the test may reduce mock crime memory strength and hence CIT sensitivity (Gronau et al., 2015; although see Gamer, Kosiol, & Vossel, 2010; Hu & Rosenfeld, 2012; Lefebvre, Marchand, Smith, & Connolly, 2007; Nahari & Ben-Shakhar, 2011).

Methods

Participants

Seventy-two undergraduate students were recruited through offer of course credits or monetary (20 pounds) compensation in return for their participation. They were randomly assigned to three experimental groups (N = 24 in each group): Guilty-Immediate (21 female and 3 male), Guilty-Delay (20 female and 4 male), or Guilty-Alibi with HT (delay with alibi home training; 19 female and 5 male). The mean age was 19.54 (SD = 1.96, range 18-31) years (Guilty-Immediate M = 19.29, SD = 1.46; Guilty-Delay M = 19.08, SD = 1.59; and Guilty-Alibi with HT M = 20.25, SD = 2.54). The groups were not significantly different in terms of age (F(2, 69) = 2.51, p = .089, η_p^2 = .068) nor in gender (X^2(2) = .60, p = .741, φ = .091). All

participants were right-handed, had English as their first language, have normal or corrected-to-normal vision, were neurologically normal and had no diagnosis of dyslexia. The study was approved by the University of Kent Psychology Ethic committee.

Materials, Design, and Procedure

Mock Crime. Each participant was given brief information about the study and asked to read and sign a consent form. After that, they were instructed to commit a mock crime, which was very similar to previous experiments, except that I varied the object that participants stole in order to be able to counterbalance stimuli assignment to conditions in the CIT. The mock crime involved stealing a box with either a ring, necklace, key, wallet, or phone inside a bag from a kitchen adjacent to staff offices in a university building and return to the laboratory. As in my previous experiments, the object they should steal was not mentioned in the instructions, so they only discovered the identity of the object through enacting the crime. Next, participants who were in the Guilty-Delay group were dismissed and subsequently returned to the laboratory after 7 days to complete the CIT. On the other hand, participants in the Guilty-Alibi with HT group proceeded with the home training procedures (described below) and the Guilty-Immediate group proceeded with the CIT.

Alibi Home-Training. As in my previous experiments, participants in the Guilty-Alibi with HT group were provided with a false alibi scenario. Participants were told that they would later take part in a test that is designed to detect if they are guilty of the crime they committed. However, they should try to appear innocent by adopting the alibi. They were instructed that it would be crucial for them to imagine

the alibi as if it was real and as vividly as possible. The alibi was a short verbal description of a scenario: "You were on your way to find your lecturer. On their door, there was a sheet of paper specifying that you could leave your email address for the lecturer to get back to you. So, you tore off a bit of paper and wrote your email address and left it in the envelope provided and came back here. The envelope has since been destroyed so there is no evidence that your alibi is false" (this was the same alibi scenario as Experiments 1-3). Participants were told to close their eyes and imagine the scenario as vividly as possible for two minutes. Then, they were asked to repeat the alibi verbally and answer a few questions regarding the alibi. However, if they failed to give a correct answer, the alibi and the questions were repeated until they are able to give the correct answer. They then repeatedly practiced the alibi at home for 6 consecutive days by going online to a provided Qualtrics link, where they would read a description of the alibi scenario, then imagine the alibi (during a timed page) and answer a few questions about it (e.g. rate the extent to which they could imagine the alibi). They were instructed to complete the alibi home training every day, otherwise they were not eligible to complete the rest of the study (and compliance was monitored by the experimenter). After receiving instructions, the Guilty-Alibi with HT group were dismissed and returned to the laboratory after 7 days for the CIT.

P300-Based Complex Trial Protocol (CTP) CIT. In the final stage, all participant took an P300-based CTP, which was closely based on prior research (Rosenfeld et al., 2008; 2013). Participant in the Guilty-Immediate group completed the CTP immediately following the mock crime, whilst participants in the other two groups performed the CTP 7 days after the mock crime. The EEG was set up, and

continuous EEG was recorded during the CTP. The CTP involves two stimulus presentations per trial, the first presentation is either a probe or irrelevant item, whereas the second presentation is a target or non-target presentation for a target detection task. During the probe/irrelevant presentation, participants were shown one of six stimuli, which were the words “ring”, “necklace”, “watch”, “key”, “wallet”, or “phone” presented in random order, in white font in the middle of the screen. These words thus consisted of one probe (the object they had stolen, counterbalanced across participants) and five irrelevants (other objects they had not stolen), and participants were required to respond to any stimuli by pressing the ‘X’ button on the keyboard as soon as any word was presented on the screen (Figure 4.1). Thus, participants were not asked to explicitly discriminate between the probe and irrelevants, so that any enlarged P300 for probes would be due to task-unrelated recognition of that probe as meaningful, indicating that participants had incriminating knowledge about the mock crime. The probe and each irrelevant were repeated for 50 times, which made up a total of 300 trials.

Immediately following probe/irrelevant presentations, the target/non-target presentation consisted of strings of numbers (‘111111’ – ‘666666’). Participants were required to press the ‘m’ button on the keyboard when the target (‘111111’) appeared on the screen and press the ‘n’ button when any other string of numbers appeared. Target and non-target strings were randomly presented and repeated for 50 times each, 300 trials in total. There was no systematic relationship between probe/irrelevant and target/non-target presentation orders, and participants were always presented with probe/irrelevant stimuli first, then target/non-target stimuli. Each complex trial began with a fixation cross for 100ms, which was followed by the probe/irrelevant stimulus for 300ms, which was followed by a black screen at

1400-1700ms, then the target/non-target stimulus was presented for 300ms and was followed by another black screen at 2400ms (Figure. 4.1). To maintain attention to probe/irrelevant stimuli, participants were also informed that the experiment would randomly pause every 20~40 trials and require them to report the most recent word they had seen, by typing their response with the keyboard. Participants were instructed to try to avoid excessive blinking, and especially to avoid blinking while the probe/irrelevant words were on the screen, but there was no dedicated "blink pauses". After the CIT, all participants completed a post-questionnaire (see Appendix G-H).

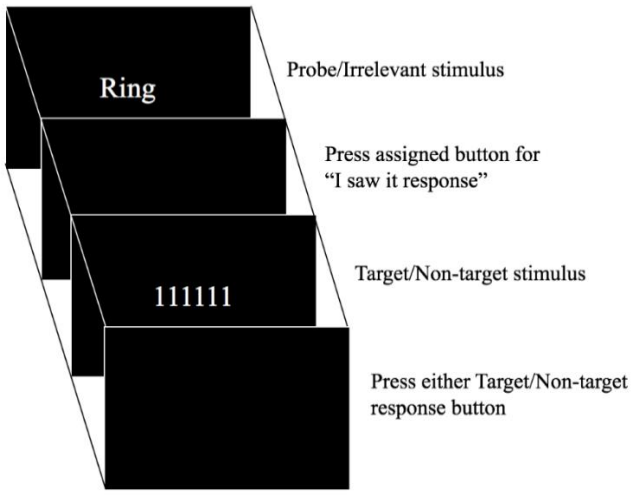

Figure 4.1. Illustration of the Complex Trial Protocol (CTP) procedure.

EEG recording and pre-processing

Continuous EEG was recorded during the CIT phase at 500 Hz with a 0.05 to 70 Hz bandwidth using 64 scalp electrodes placed in an actiCAP, with FCz as a reference electrode (Brain Products GmbH, München, Germany). Channel locations were based on the extended 10-20 system. The electrooculogram (EOG) was recorded vertically from below the left eye (VEOG) and horizontally from outer of the right eye (HEOG). The recorded data were analysed using EEGLAB (UC San

Diego, 2004). The EEG was re-referenced to the mastoids and divided into 300 epochs (-100ms to 1500ms), time-locked to the onset of the probe/irrelevant stimulus presentation. The data were digitally filtered with a 30Hz low-pass and 0.3Hz high-pass. Although a 0.3Hz high-pass can distort the shape of ERPs, it is recommended in the CIT paradigm because it has been shown to increase discrimination between guilty and innocent suspects (Soskins, Rosenfeld & Niendam, 2001). Then, the epochs were concatenated and submitted to Independent Component Analysis (ICA) using runica (infomax). The independent components reflecting eye movements and artefacts were identified by visual inspection of their topography, time-course and spectral profile, following recommendations in the EEGLAB manual. Noise components were subsequently removed from the data, and any trials that still contained artefacts after ICA cleaning were also removed after visual inspection. ERPs were formed separately for Probe items (the crime-relevant word, acting as a reminder of the stolen object) and Irrelevant items (other words used as distractors, that gives a baseline for how large the ERP response is in the absence of crime recognition). The mean trial numbers per group and condition were as follows; Guilty-Immediate: probe (Range = 43-50, *M* = 49.33, *SD* = 1.63), irrelevant (Range = 208-250, *M* = 246.46, *SD* = 8.94; Guilty-Delay: Probe (Range = 45-50, *M* = 48.54, *SD* = 1.59), Irrelevant (Range = 224-250, *M* = 242.21, *SD* = 7.32); Guilty-Alibi with HT: Probe (Range = 41-50, *M* = 48.92, *SD* = 1.98), Irrelevant (Range = 206-250, *M* = 242.42, *SD* = 10.28)).

In an initial targeted analysis, time windows and locations used when analysing ERP effects related to probe recognition were based on the literature (specifically, I followed Hu et al., 2015). In line with previous studies, we examined both the P300 positive peak and the late posterior negativity (LPN) that typically

occurs after the P300 and is typically also enlarged for probes compared to irrelevants. These effects were measured at the Pz site based on their typical scalp distribution and convention in the literature. The amplitude of the P300 (base-to-peak P300 measure, since this amplitude is in relation to the pre-stimulus baseline that was used for baseline correction) was calculated as the mean of the most positive 100ms segment during the 300-800ms post-stimulus time window, identified at an individual level to counteract individual differences in P300 latency. The LPN was calculated as the mean of the most negative 100ms segment following the P300 latency to the end of epoch at 1500ms. Furthermore, the peak-to-peak measure, which is the difference between the P300 and LPN amplitudes, was also analysed since prior literature has found that this is the most effective measure for discriminating guilty from innocent participants (e.g Rosenfeld et al., 2013). For all of these ERP measures, it is expected that guilty suspects should show a larger amplitude effect for probes than irrelevant stimuli (so a positive difference for P300 base-peak measure and the peak-peak measure, and a negative difference for the LPN since this is a negative going ERP component).

Because the targeted analysis described above may miss out on ERP differences that occur in other locations and electrode locations, I also conducted an exploratory "whole head" analysis where I analysed the effect of the group manipulation on mean amplitudes for probe and irrelevant stimuli across seven successive time-windows (0-200ms, 200-400ms, 400-600ms, 600-800ms, 800-1000ms, 1000-1250ms, 1250-1500ms) from a grid of electrodes that were distributed across left and right hemispheres and frontopolar, frontal, central, parietal and occipital sites (FP1, FP2, F3, F4, C3, C4, P3, P4, O1, and O2).

Results

Behaviour results

First, I analysed participants' performance on the probe/irrelevant button pressing task to assess whether there were any behavioural differences between the groups in how they completed this task. For reaction times, a 3 (Group: Guilty-Immediate, Guilty-Delay, vs. Guilty-Alibi with HT; between groups) x 2 (Stimulus type: Probe vs Irrelevant; within participants) mixed ANOVA revealed no main effects of Group ($F(2, 69) = .64$, $p = .531$, $\eta_p^2 = .018$) nor Stimulus type ($F(1, 69) = .177$, $p = .675$, $\eta_p^2 = .003$), nor Group x Stimulus type interaction ($F(2, 69) = 1.34$, $p = .270$, $\eta_p^2 = .037$). For accuracy, a (Group: Guilty-Immediate, Guilty-Delay, vs. Guilty-Alibi with HT; between groups) x 2 (Stimulus type: Probe vs Irrelevant; within participants) mixed ANOVA revealed no main effect of neither Group ($F(2, 69) = 1.66$, $p = .199$, $\eta_p^2 = .046$) nor Stimulus type ($F(1, 69) = .045$, $p = .833$, $\eta_p^2 = .001$), and no Group x Stimulus type interaction ($F(2, 69) = 1.90$, $p = .157$, $\eta_p^2 = .052$). Thus, behavioural performance on the probe/irrelevant button pressing task was not affected by the group manipulation or by the crime-relevance of the word, which may be expected since the task was simply to press a button when *any* word appeared.

ERPs

Grand average ERPs from the mid-parietal site are shown in Figure 4.2.

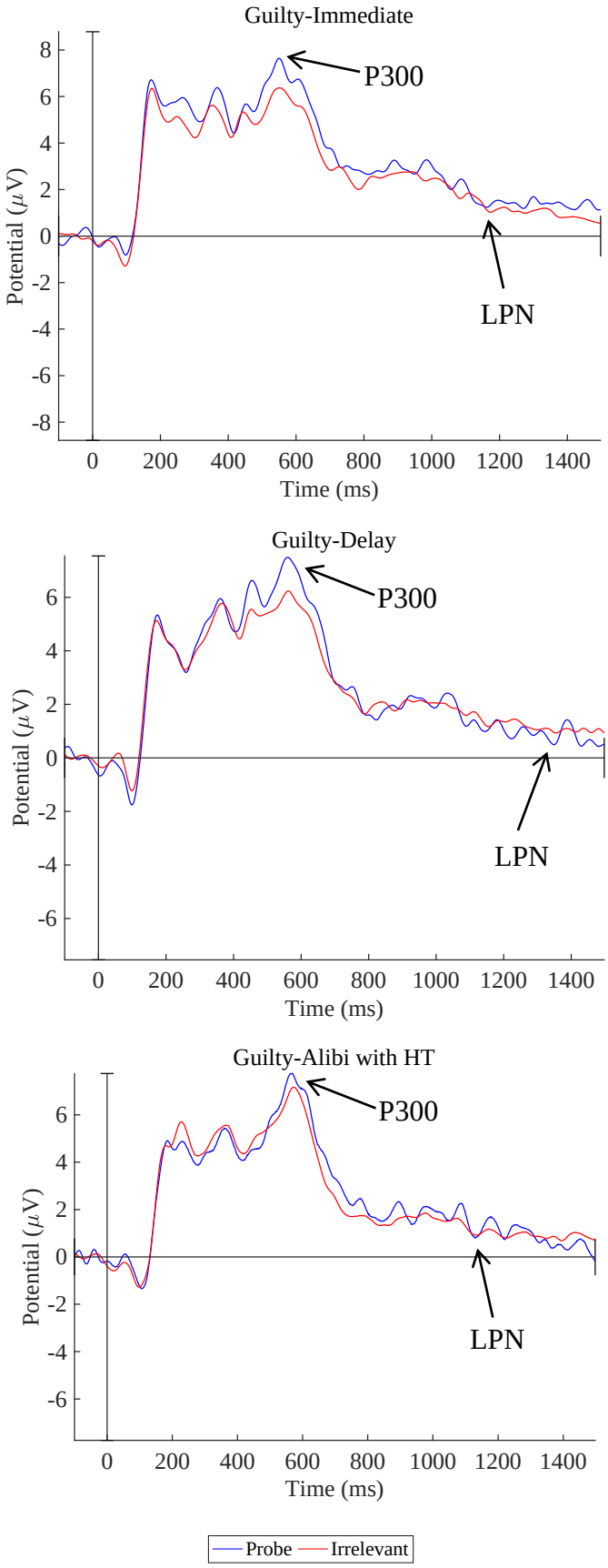

Figure 4.2. Grand-average mid-parietal (Pz site) ERPs for the Probe and Irrelevant conditions in the three groups.

Targeted P300/LPN analyses

For the targeted ERPs analyses, 3 (Group: Guilty-Immediate, Guilty-Delay, vs. Guilty-Alibi with HT; between groups) x 2 (Stimulus type: Probe vs Irrelevant; within participants) mixed ANOVAs were conducted to examine the effects of the Group manipulation on the Probe-Irrelevant difference for the three ERP effects of interest at the mid-parietal site (Pz; see Fig. 4.3). For the base-to-peak P300, there was no significant main effect of Group ($F(2, 69) = 0.313$, $p = .732$, $\eta_p^2 = .009$) but a significant main effect of Stimulus type ($F(1, 69) = 10.43$, $p = .002$, $\eta_p^2 = .131$). There was however no interaction between Group and Stimulus type ($F(2, 69) = 0.812$, $p = .448$, $\eta_p^2 = .023$). Regardless of Group, Probe stimuli ($M = 8.18$, $SD = 3.16$) elicited significantly larger P300 amplitudes when compared to Irrelevant stimuli ($M = 7.48$, $SD = 2.58$), in line with the typical finding when a CIT is performed on a "Guilty" group of participants. Even though this Probe-Irrelevant effect did not interact with Group, I wanted to verify to presence of an effect in each of the groups separately, so paired t-tests were conducted to compare the base-to-peak P300 between Probe and Irrelevant stimuli for each group. Results revealed a significantly larger P300 for Probes ($M = 8.19$, $SD = 3.06$) than Irrelevant stimuli ($M = 7.16$, $SD = 2.45$; $t(23) = 2.96$, $p = .007$, $d = 0.37$) in the Guilty-Immediate group, but not in the Guilty-Delay group ($M = 8.55$, $SD = 3.93$; $M = 7.82$, $SD = 3.16$, respectively; $t(23) = 1.61$, $p = .120$, $d = 0.20$) or Guilty-Alibi with HT group ($M = 7.79$, $SD = 2.38$; $M = 7.44$, $SD = 2.05$, respectively; $t(23) = 1.10$, $p = .282$, $d = 0.16$).

For the LPN, there was again no significant main effect of Group ($F(2, 69) = .33$, $p = .721$, $\eta_p^2 = .009$) but a significant main effect of Stimulus type ($F(1,69) = 9.29$, $p = .003$, $\eta_p^2 = .119$), and no interaction between Group and Stimulus type ($F(2, 69) = .81$, $p = .451$, $\eta_p^2 = .023$). Across groups, Probes ($M = -.52$, $SD = 1.74$) elicited

more negative LPN amplitude when compared to Irrelevant stimuli (*M* = .06, *SD* = 1.33), in line with typical CIT findings. Again, paired t-tests were conducted to compare the LPN between Probe and Irrelevant stimuli separately for each group. There were no differences between Probes (*M* = -.17, *SD* = 1.61) and Irrelevants (*M* = .07, *SD* = 1.32) in the Guilty-Immediate group (*t*(23) = -.96, *p* = .347, *d* = 0.16), and only a non-significant trend towards a difference in the Guilty-Delay group (*M* = -.74, *SD* = 2.12; *M* = .05, *SD* = 1.52, respectively; *t*(23) = 2.06, *p* = .051, *d* = 0.43), and the difference in the Guilty-Alibi with HT group was only marginally significant (*M* = -.63, *SD* = 1.43; *M* = .07, *SD* = 1.19, respectively; *t*(23) 2.09, *p* = .048, *d* = 0.53).

The peak-to-peak measure that consists of the difference between the positive P300 peak and the negative LPN peak showed similar results as the individual peak measures. There was no significant main effect of Group (*F*(2, 69) = .61, p = .549, η_p^2 = .017) but a highly significant main effect of Stimulus type (*F*(1,69) = 29.21, *p* < .001, η_p^2 = .30). However, these two factors did not interact (F(2, 69) = .34, p = .716, η_p^2 = .01). The P300-LPN was significant larger for probe stimuli (*M* = 8.69, *SD* = 3.31) than for irrelevant stimuli (*M* = 7.41, *SD* = 2.33), as would be expected. Follow-up t-tests to assess if this effect was significant within each group separately revealed a significant larger P300-LPN difference for Probes (*M* = 8.36, *SD* = 2.74) than Irrelevants (*M* = 7.09, *SD* = 2.25) in the Guilty-Immediate group (*t*(23) = 3.68, *p* = .001, *d* = 0.51), as well as in the Guilty-Delay (*M* = 9.30, *SD* = 4.15; *M* = 7.77, *SD* = 2.85, respectively; *t*(23) = 2.99, *p* = .007, *d* = 0.43) and Guilty-Alibi with HT groups (*M* = 8.42, *SD* = 2.89; *M* = 7.37, *SD* = 1.83, respectively; *t*(23) = 2.95, *p* = .007, *d* = 0.43). Thus, this initial group level analysis of P300 and LPN effects showed very small differences between the groups, and converged with the

observation in the literature that the P300-LPN peak-to-peak was more sensitive for detecting guilt than the individual peaks alone (e.g. Rosenfeld, et al., 2013).

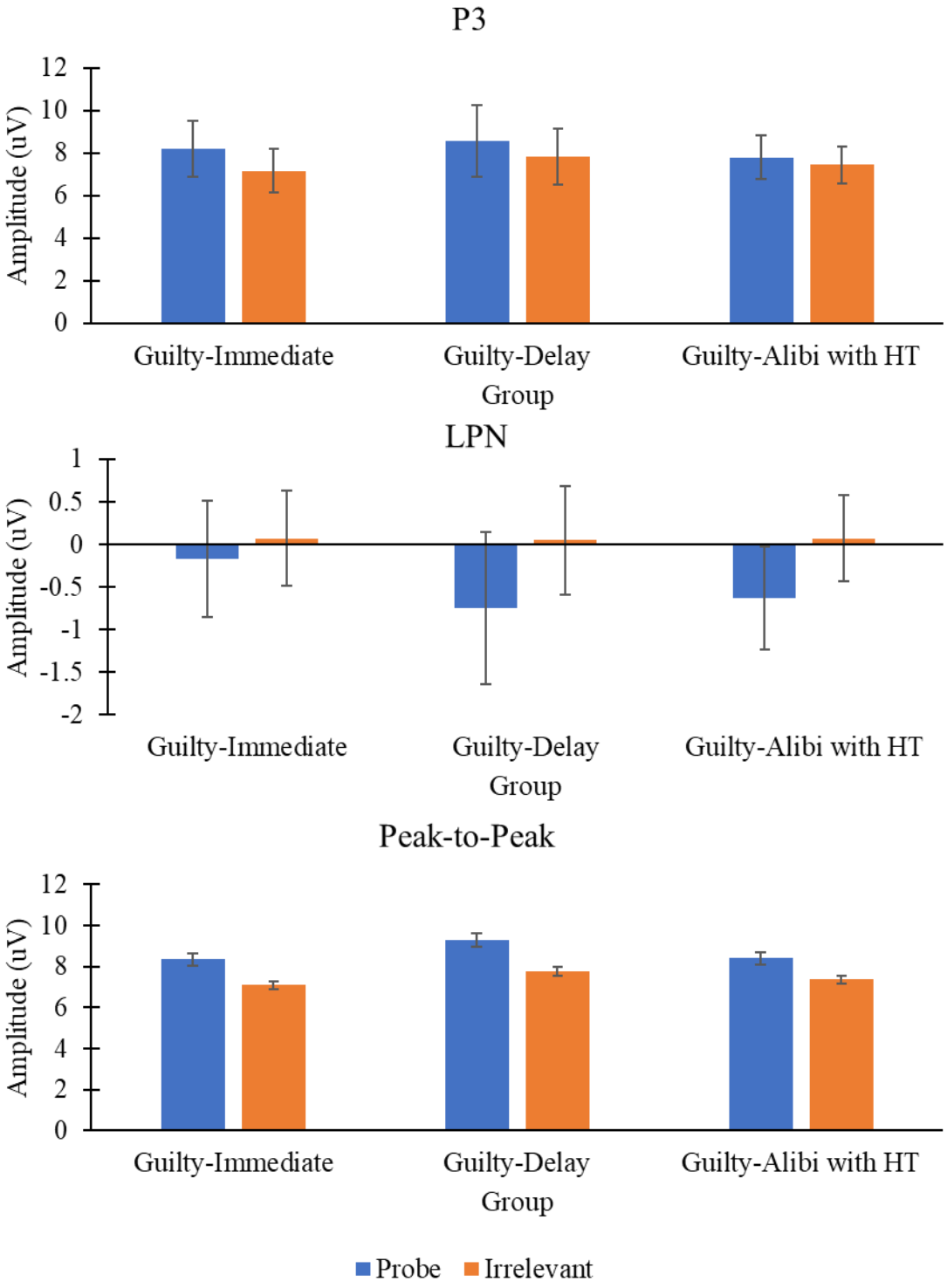

Figure 4.3. Mean amplitudes of the base-to-peak P300, LPN, and peak-to-peak P300-LPN measures for Probe and Irrelevant stimuli in the three groups extracted at the mid parietal site. Error bars represent 95% confidence interval.

Individual guilt diagnoses

Because the CIT is meant to diagnose guilt of individuals, this requires conducting statistical analysis for each individual participant. I used bootstrap analysis to investigate individual guilt classification following the same bootstrap procedure as in Rosenfeld et al. (2008), which is the most commonly used method in the P300 CIT literature. This analysis involves repeated random sampling with replacement from the single EEG trials that are used to construct the ERP for Probe and Irrelevant conditions, and then averaging these together for each stimulus type into bootstrap ERPs. This process was repeated for 1,000 iterations. Then, the three ERP measures of interest (the base-to-peak P300, the LPN, and the peak-to-peak P300-LNP) are extracted for each bootstrap ERP from the mid-parietal site, using the same method as when extracting the actual observed measures (as analysed in the previous section). Next, the proportion of samples for which the Probe bootstrapped ERP measure is larger than the Irrelevant is calculated. If this proportion is higher than 90%, the subject is classified as guilty (see Rosenfeld et al., 2008, for more information).

For the base-to-peak P300, only 25% (18 out of 72) of participants were classified correctly as guilty across all groups (i.e. for those 18 participants, the base-to-peak P300 for probes was more positive than the base-to-peak P300 for irrelevants on at least 90% of bootstrap samples). In the Guilty-Immediate group 8/24 participants were classified as guilty, which was not significantly differ from the Guilty-Delay group (7 participants; $\chi^2(1) = 0.10$, $p = .755$, $\varphi = 0.045$) nor Guilty-Alibi with HT group (3 participants; $\chi^2(1) = 2.95$, $p = .086$, $\varphi = 0.248$). There was also no difference in guilt classification between participants in Guilty-Delay and Guilty-Alibi with HT groups ($\chi^2(1) = 2.02$, $p = .155$, $\varphi = 0.205$). For the LPN

measure, similarly only 26% (19 out of 72) of participants were correctly classified as guilty (as indicated by the LPN for the probes being more negative than for the irrelevants for at least 90% of bootstrap samples) . There were 4 and 7 out of 24 participants correctly classified in Guilty-Immediate and Guilty-Delay group, respectively, which was not statistically different ($\chi^2(1) = 1.06$, $p = .303$, $\varphi = 0.149$). Eight participants in the Guilty-Alibi with HT group were classified as guilty, which was also not significantly different compared to the Guilty-Immediate ($\chi^2(1) = 1.78$, $p = .182$, $\varphi = 0.192$) or Guilty-Delay groups ($\chi^2(1) = .10$, $p = .755$, $\varphi = 0.045$). Guilt detection was somewhat better with the P300-LPN peak-to-peak measure, since 42% (30 out of 72) of participants were correctly classified as guilty across groups (i.e. their P300-LPN difference was larger for probes than irrelevants on at least 90% of bootstrap samples). With this measure, 10 participants in each of the groups were classified as guilty, showing no effect at all of the manipulation on peak-to-peak guilt detection of individuals.

To summarise the targeted ERP analysis: at the group level, we found the strongest Probe-Irrelevant difference using the peak-to-peak P300-LPN measure, and bootstrap analysis confirmed that the peak-to-peak is the most sensitive measure for discriminating guilty from innocent suspects, as previously argued (Rosenfeld, 2013). However, guilt detection in this study was relatively low compared to the literature, and was not affected by the manipulation. One possible reason for the poor detection rate might be that our attention check was not sufficiently effective at making participants attend to the probe/irrelevant stimuli, so I conducted additional correlations between the bootstrap measure of guilt and attention check performance to assess if people who showed smaller evidence of guilt also had poorer performance on the attention check. If the attention check of this study was

ineffective, people who failed the attention check should have a less reliable probe vs. irrelevant difference. However, the results did not support this prediction, as the only significant effect was a *negative* correlation between the number of bootstrap samples for which the P300 was larger for probes than irrelevants (higher numbers indicating better memory detection) and the proportion of correct attention check responses in the Guilty-Immediate group (indicating more attention to probes/irrelevant words, ($r(22) = -.50$, $p = .014$) but not in Guilty-Delay ($r(22) = .23$, $p = .566$) nor Guilty-Alibi with HT ($r(22) = .004$, $p = .987$) groups. Thus within the Guilty-Immediate group only, smaller evidence of guilt was associated with *better* performance on the attention check task, but this pattern did not generalise to the delay groups.

Post-experiment questionnaire results

An additional questionnaire was also used to gain insight into any effects found in the CIT. Participants provided ratings on a scale from 0 to 6 on various questions. Results revealed that there was no difference in the extent to which crime-relevant memories came to mind automatically during the CIT and how nervous participants had felt during the mock crime among the groups ($ps > .11$). However, there was a significant difference in motivation to beat the test. The Guilty-Alibi with HT group ($M = 4.04$, $SD = 1.46$) was more motivated than the Guilty-Delay group ($M = 2.50$, $SD = 1.14$; $t(46) = 4.08$, $p < .001$, $d = 1.20$) and the Guilty-Immediate group ($M = 3.42$, $SD = 1.56$) was also more motivated than the Guilty-Delay group ($t(46) = 2.32$, $p = .025$, $d = .68$).

Additional questions related to the false alibi were submitted to a correlation analysis to investigate factors that might affect CIT detection in the Guilty-Alibi with

HT group. Results showed that self-reported vividness of the imagined alibi was negatively correlated with the amplitude of the observed base-to-peak P300 for the Probe ($r(24) = -.41$, $p = .046$) and the Irrelevant stimuli ($r(24) = -.43$, $p = .035$). The extent to which participants believed that the alibi scenario occurred was negatively correlated with the observed LPN for the Probe ($r(24) = -.48$, $p = .017$) and positively correlated with the observed P300-LPN peak-to-peak for the Probe ($r(24) = .45$, $p = .026$). However, since a large number of correlations were performed and these correlations would not survive correction for multiple comparisons, they were not interpreted any further.

Whole-head ERP analysis

In a final ERP analysis from the CTP phase, I investigated if there were any differences between the groups in terms of their Probe vs. Irrelevant ERP effects across other time-windows and locations that were not considered in the targeted analysis presented above. The mean amplitudes of Probe and Irrelevant ERPs were analysed for seven successive time-windows (0-200ms, 200-400ms, 400-600ms, 600-800ms, 800-1000ms, 1000-1250ms, 1250-1500ms) across 10 electrodes (FP1, FP2, F3, F4, C3, C4, P3, P4, O1, and O2). For each window, an omnibus repeated measure ANOVA was conducted with the factors Group (Guilty-Immediate, Guilty-Delay, Guilty-Alibi with HT) x Stimulus type (Probe vs. Irrelevant) x Hemisphere (Left vs. Right) x Anterior-Posterior region (Frontopolar, Frontal, Central, Parietal, Occipital). Paired t-tests were conducted to follow up significant results that involved interactions with the Stimulus type as a factor (as these are the only results that would be meaningful to interpret). Scalp topographic maps showing differences between Probe and Irrelevant stimulus types are shown in Figure 4.4. As can be seen

in this figure, the probes elicited more positive ERPs than irrelevant in all groups, but this effect seemed to onset earlier and last longer in the Guilty-Immediate group, whereas the Probe-Irrelevant difference seemed to peak in the 400-600ms window in the Guilty-Delay group, and in the 600-800ms window in the Guilty-Alibi with HT group. Grand-average ERPs from the 10 electrodes sites for each group are showed in Figure 4.5 - 4.7.

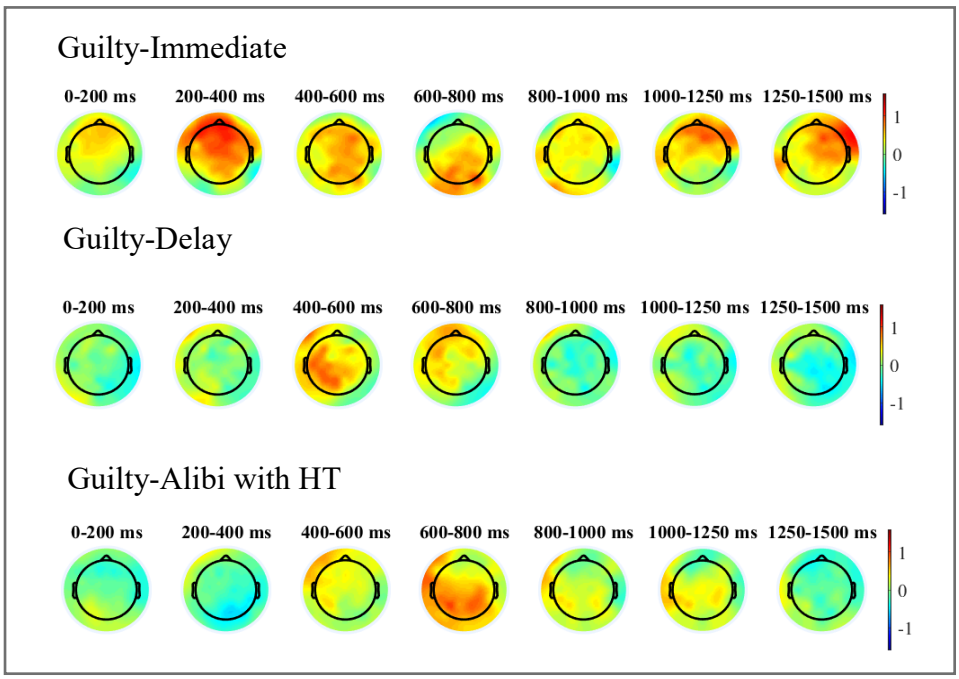

Figure 4.4. Topographic maps showing the scalp distribution of ERP amplitude differences between Probes and Irrelevants for the three groups. The plots were generated by subtracting the mean amplitudes of the Irrelevant condition from the mean amplitudes for the Probes.

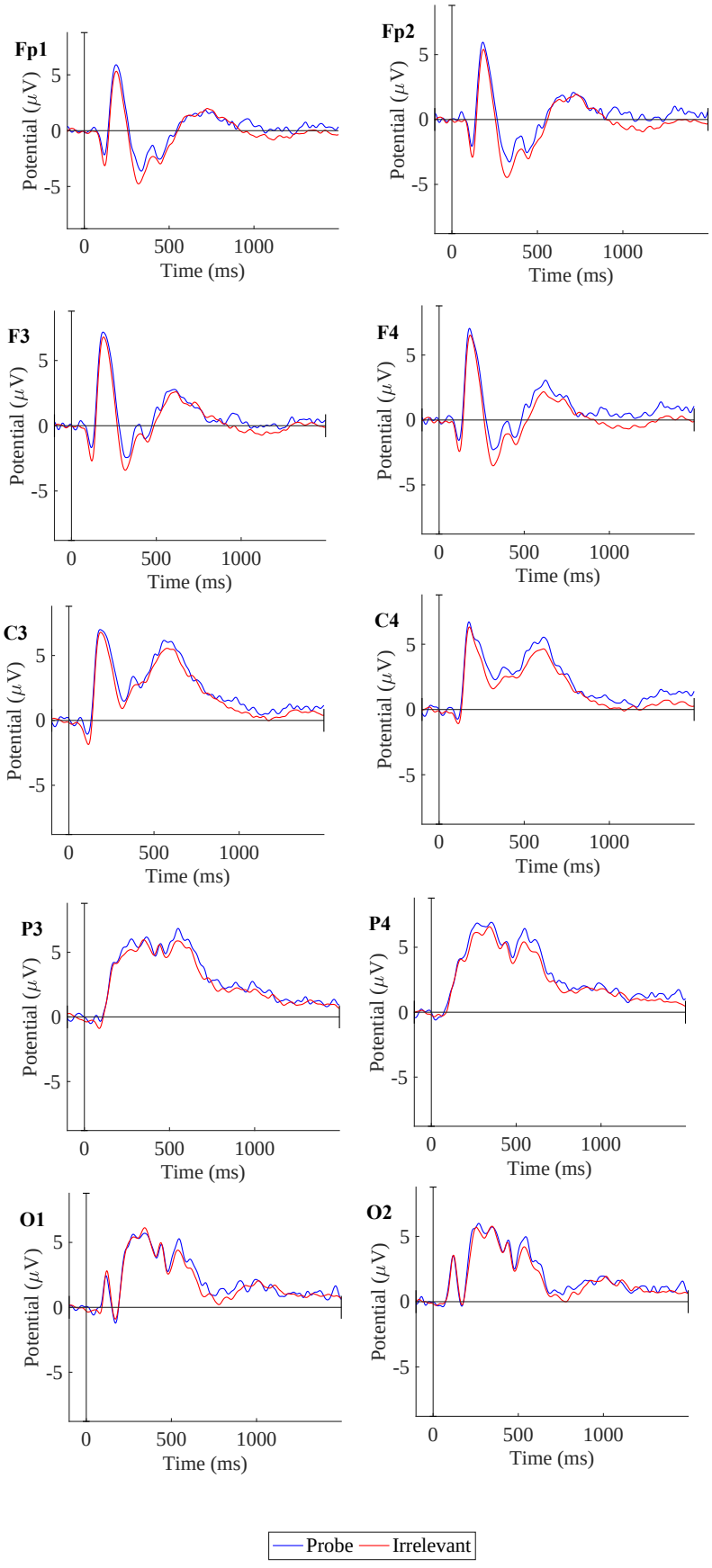

Figure 4.5. Grand-average ERPs from Guilty-Immediate group during the CIT.

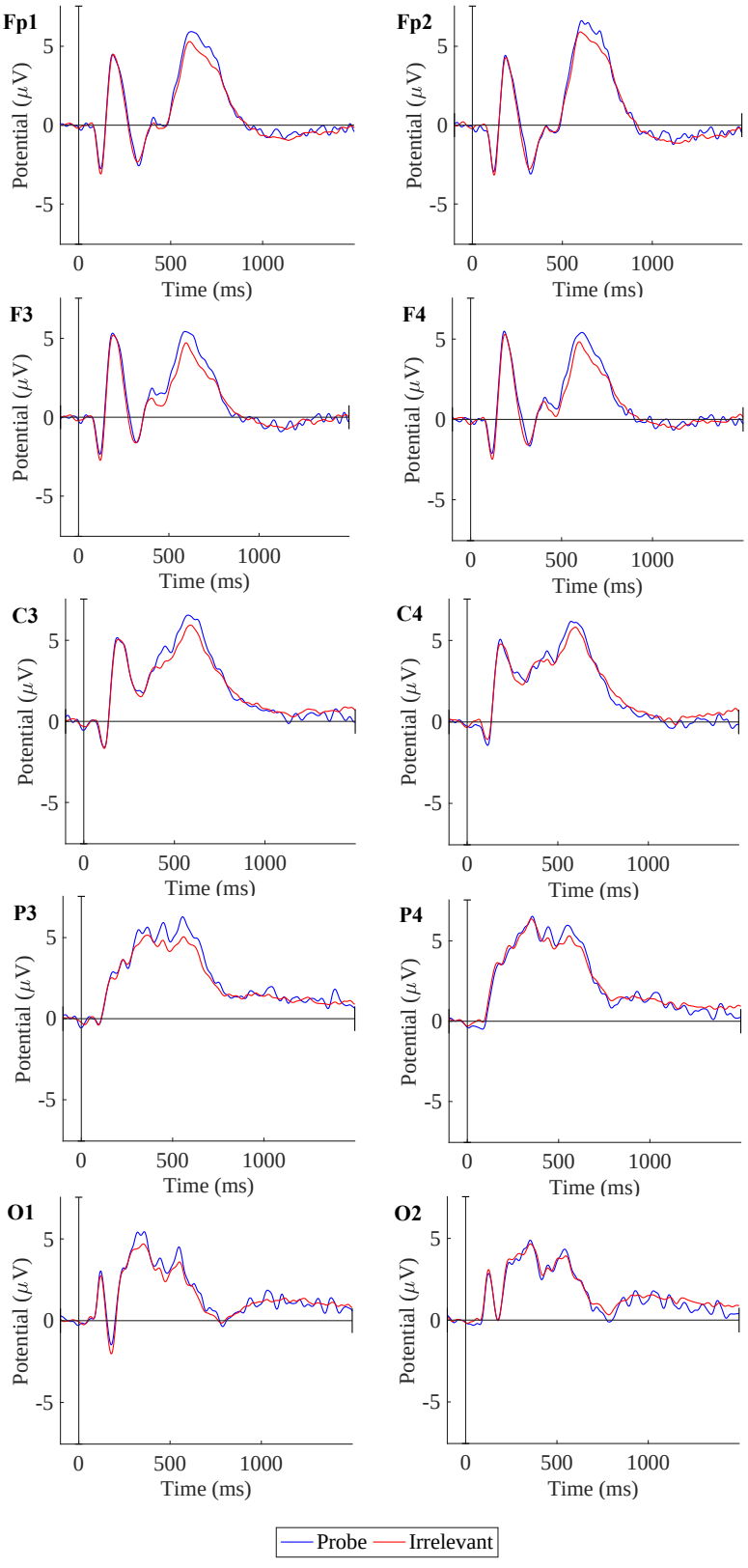

Figure 4.6. Grand-average ERPs from Guilty-Delay group during the CIT.

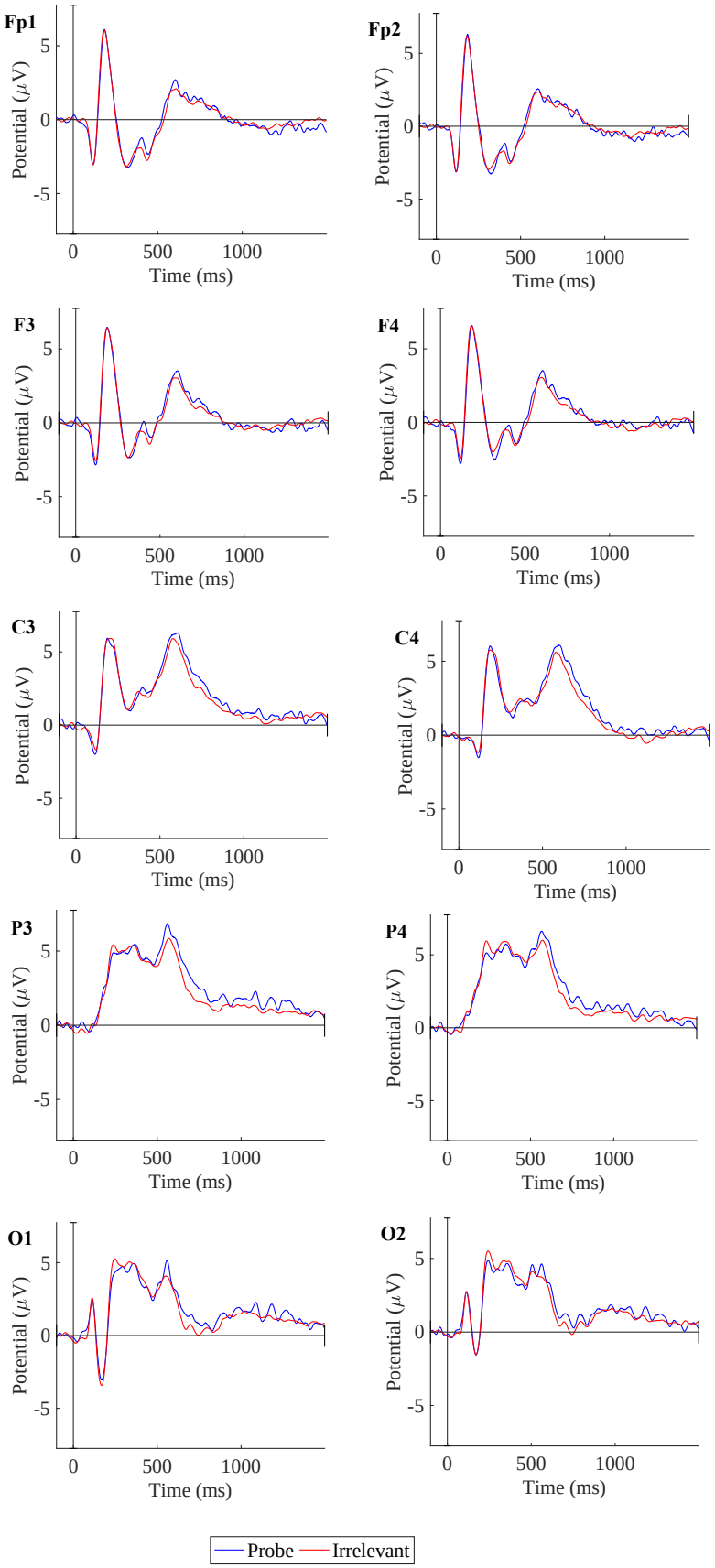

Figure 4.7. Grand-average ERPs from Guilty-Alibi with HT group during the CIT.

The results from the omnibus 5 (Anterior-Posterior: FP, F, C, P, and O electrode regions) x 2 (Hemisphere: Left vs. Right) x 2 (Stimulus Type: Probe vs. Irrelevant) x 3 (Group: Guilty-Immediate, Guilty-Delay, and Guilty-Alibi with HT) mixed ANOVA are shown in Table 4.1. There was a significant three-way Anterior-Posterior x Stimulus type x Group interaction in the 0-200ms time-window. Follow up 2-way Anterior-Posterior x Stimulus type ANOVAs were conducted for each group separately, collapsed across hemisphere. Results indicated that there were no significant main effects of Stimulus type in the Guilty-Delay or Guilty-Alibi with HT groups (Guilty-Delay: $F(1,23) = 0.01$, $p = .926$, $\eta_p^2 = .000$; Guilty-Alibi with HT: $F(1,23) = 0.02$, $p = .901$, $\eta_p^2 = .001$). There was also no significant Stimulus type x Anterior-Posterior interaction in the Guilty-Delay ($F(1.71,39.29) = 0.654$, $p = .502$, $\eta_p^2 = .028$), nor Guilty-Alibi with HT ($F(1.34, 30.81) = 1.44$, $p = .248$, $\eta_p^2 = .059$) groups. However, in the Guilty-Immediate group the main effect of Stimulus type ($F(1,23) = 4.19$, $p = .052$, $\eta_p^2 = .154$) was at trend level, and so was the Anterior Posterior x Stimulus type interaction ($F(2.08, 47.89) = 2.82$, $p = .067$, $\eta_p^2 = .109$).

There was a significant Stimulus type by Group interaction in the 200-400ms time-window. Follow up paired t-tests compared the mean amplitude of Probes vs. Irrelevants separately within each group, collapsed across electrode sites (since there were no interactions with Hemisphere or Anterior-Posterior region). These showed that Probes elicited more positive ERP amplitudes than Irrelevants in the Guilty-Immediate group between 200-400ms ($t(23) = 2.70$, $p = .013$, $d = .43$), whereas there were no Probe vs. Irrelevant ERP differences in this early time window in the Guilty-Delay group ($t(23) = .68$, $p = .502$, $d = .07$) and the Guilty-Alibi with HT group ($t(23) = .91$, $p = .370$, $d = .13$).

In the 400-600ms window, results revealed only a significant main effect of condition for 400-600ms that no longer interacted with the Group factor, caused by a broadly distributed positivity for Probes compared to Irrelevants that was similar across groups (see Figure 4.4). There was a significant three-way Anterior-Posterior x Stimulus type x Group interaction in the 600-800ms time-window. Follow up 2-way ANOVAs were conducted with the factors Anterior-Posterior x Stimulus type, separately for each group and collapsed across the Hemisphere factor. The results indicated that the Probe-Irrelevant main effect was not significant in the Guilty-Immediate and Guilty-Delay groups in this time window (Guilty-Immediate: $F(1,23) = 2.51$, $p = .127$, $\eta_p^2 = .098$; Guilty-Delay: $F(1,23) = 2.24$, $p = .149$, $\eta_p^2 = .089$), and the Probe-Irrelevant difference also did not differ as a function of Anterior-Posterior location in the Guilty-Immediate ($F(1.72, 39.59) = 1.46$, $p = .244$, $\eta_p^2 = .060$) and Guilty-Delay groups ($F(1.62,37.31) = 2.09$, $p = .146$, $\eta_p^2 = .083$). However, there was a significant Stimulus type x Anterior-Posterior interaction in the Guilty-Alibi with HT group ($F(1.89,43.55) = 3.62$, $p = .037$, $\eta_p^2 = .136$) such that the Probes elicited significantly more positive amplitudes than Irrelevants at Frontal ($t(23) = 2.30$, $p = .031$, $d = .20$), Central ($t(23) = 3.73$, $p = .001$, $d = .044$), Parietal ($t(23) = 3.52$, $p = .002$, $d = .64$), and Occipital ($t(23) = 3.07$, $p = .005$, $d = .41$) sites, but not at the Fronto-Polar sites ($t(23) = 0.66$, $p = .514$, $d = .07$) site.

There were no significant effects involving Stimulus type in the 800-1000ms time-window, but there was a significant Anterior-Posterior x Stimulus type x Group interaction in the 1000-1250ms time-window. Follow up test were conducted separately for each group with the Anterior-Posterior x Stimulus type factors, collapsed across Hemisphere. These showed that the Probe-Irrelevant main effect was not significant in any group (Guilty-Immediate: $F(1,23) = 2.13$, $p = .158$, $\eta_p^2 =$

.085; Guilty-Delay: $F(1,23) = 0.01$, $p = .930$, $\eta_p^2 = 0.00$; Guilty-Alibi with HT: $F(1,23) = 1.36$, $p = .256$, $\eta_p^2 = .056$), and the Probe-Irrelevant difference also did not differ as a function of Anterior-Posterior location in any group (Guilty-Immediate: $F(1.58, 36.37) = 2.54$, $p = .104$, $\eta_p^2 = .099$; Guilty-Delay: $F(1.72, 39.57) = 0.96$, $p = .379$, $\eta_p^2 = .040$; Guilty-Alibi with HT: $F(1.35, 31.11) = 1.86$, $p = .181$, $\eta_p^2 = .075$)

In the final 1250-1500ms time-window, there was a significant Hemisphere x Stimulus type x Group interaction in the 1250-1500ms time-window. Follow up test with the Hemisphere x Stimulus type factors within each group (collapsed across Anterior-Posterior location) indicated that there was no main effect of Stimulus Type in either group (Guilty-Immediate: ($F(1,23) = 2.97$, $p = .098$, $\eta_p^2 = .114$), Guilty-Delay: ($F(1,23) = .22$, $p = .642$, $\eta_p^2 = .010$), Guilty-Alibi with HT: ($F(1,23) = .19$, $p = .666$, $\eta_p^2 = .008$). The Probe-Irrelevant difference also did not differ significantly between the left and right hemisphere in the Guilty-Delay or Guilty-Alibi with HT groups (Guilty-Delay: $F(1,23) = 1.72$, $p = .203$, $\eta_p^2 = .070$, Guilty-Alibi with HT: $F(1,23) = .04$, $p = .845$, $\eta_p^2 = .002$), but the interaction was at trend level in the Guilty-Immediate group ($F(1,23) = 3.88$, $p = .061$, $\eta_p^2 = .144$).

So, to summarise the ERP whole-head results, there was some evidence of earlier onset Probe>Irrelevant differences in the Guilty-Immediate group than the other two groups, and evidence that the Probe-Irrelevant P300 difference peaked a bit later in the Guilty-Alibi with HT group compared to the other two groups, since only the Guilty-Alibi with HT group had a significant Probe-Irrelevant effect in the 600-800ms time-window. In the latter part of the epoch, there were tendencies for Probe>Irrelevant differences in the Guilty-Immediate group only, but there were not strictly significant. However, these group differences were subtle and did not

translate into any differences in memory detection with standard measures, as assessed in the previous targeted analysis

Table 4.1. Mixed ANOVA results from the omnibus test during CIT phase

	Time Window													
	0 to 200ms		200 to 400ms		400 to 600ms		600 to 800ms		800 to 1000ms		1000 to 1250ms		1250 to 1500ms	
Within	F	*p*	F	*p*	F	*p*	F	*p*	F	*p*	F	*p*	F	*p*
AP	9.20	<.001	154.81	<.001	55.29	<.001	4.37	0.002	16.18	<.001	50.52	<.001	24.51	<.001
AP * G	1.46	0.17	1.73	0.09	2.24	0.02	1.14	0.34	0.32	0.96	0.21	0.99	0.36	0.94
ST	1.27	0.26	2.35	0.13	**8.44**	**0.005**	**12.15**	**0.001**	1.16	0.29	2.09	0.15	0.35	0.56
ST * G	1.43	0.25	**3.99**	**0.02**	0.27	0.76	0.27	0.76	0.94	0.40	0.81	0.45	1.80	0.17
H	15.03	<.001	0.89	0.35	1.80	0.18	0.50	0.48	2.18	0.14	8.45	0.005	5.21	0.03
H * G	0.26	0.77	0.02	0.98	1.15	0.32	2.45	0.09	2.06	0.14	1.24	0.30	3.57	0.03
AP * ST	0.26	0.90	**2.90**	**0.02**	1.21	0.30	1.22	0.30	0.16	0.96	0.76	0.55	0.34	0.85
AP * ST * G	**2.34**	**0.02**	0.94	0.49	0.32	0.96	**2.76**	**0.01**	0.70	0.69	**2.52**	**0.01**	1.26	0.27
AP * H	10.40	<.001	6.49	<.001	3.25	0.01	3.64	0.01	9.20	<.001	6.84	<.001	4.55	0.001
AP * H * G	0.55	0.82	1.25	0.27	0.32	0.96	0.40	0.92	0.32	0.96	0.43	0.91	0.35	0.94
ST * H	2.28	0.14	0.14	0.71	0.46	0.50	0.54	0.46	0.17	0.68	0.24	0.63	0.27	0.61
ST * H * G	0.44	0.65	1.09	0.34	2.78	0.07	2.41	0.10	0.13	0.88	0.63	0.54	**3.31**	**0.04**
AP * ST * H	1.27	0.28	0.04	1.00	0.54	0.70	1.60	0.17	0.49	0.74	2.24	0.07	1.78	0.13
AP * ST * H * G	0.40	0.92	0.77	0.63	0.22	0.99	0.56	0.81	0.12	1.00	0.18	0.99	0.58	0.79
Between														
G	2.46	0.09	0.32	0.73	1.12	0.33	0.45	0.64	0.42	0.66	0.45	0.64	1.51	0.23

Note. AP = Anterior-Posterior, G = Group, ST = Stimulus Type, H = Hemisphere. Significant results are in bold.

Discussion

This experiment was conducted to investigate the effect of repeatedly rehearsing a false alibi on true memory detection with the ERP-based Concealed Information Test (CIT), with the countermeasure-resistant Complex-Trial Protocol (CTP, Rosenfeld, et al., 2008; 2013). As a previous study had found that rehearsing false information can disrupt CIT memory detection with physiological measures (Gronau, et al., 2015), I predicted that rehearsing a false alibi might also impair ERP-based memory detection. However, contrary to that prior finding and the predictions, the results showed that there were no differences between groups, and guilty suspect who adopted a fake alibi were just as detectable as those who did not rehearse an alibi, regardless of whether the CIT was administered immediately after the mock crime or a week after. Thus, the results suggests that the alibi countermeasure did not reduce ERP markers of concealed information, in contrast to the significant reduction I observed for aIAT memory detection in previous experiments in this thesis.

When compared to the most relevant control group, the Guilty-Delay group (that was matched to the Alibi group in terms of time-delay), the alibi had no effect on CIT sensitivity when using with either of the standard base-peak P300, LPN, or P300-LPN peak-to-peak measures. Although there was a numerical reduction in the Probe-Irrelevant ERP amplitude difference in the Guilty-Alibi with home training group, the difference was not significantly smaller than in the other two groups, and this was also the case when comparing the Guilty-Immediate group with the other two groups. Instead, all three groups seemed to show the standard Probe-Irrelevant difference, however it was rather weak and some of the measures (base-peak P300 and LPN) did

not show significant differences within each group. There was also no significant differences between the groups in behavioural performance on the CIT task.

In addition, the whole head analysis confirmed the classic probe vs. irrelevant effect of the CIT; probes elicited more positive ERPs than irrelevant stimuli across most of the epoch, and the maximal differences was at 200-400ms time-window in the Guilty-Immediate group. However, for the Guilty-Delay group, the difference between Probe and Irrelevant stimuli was found maximal in the right hemisphere in the 400-600ms time-window. For the Guilty-Alibi with HT group, the difference between Probe and Irrelevant stimuli was maximal towards the back of the head around 600-800ms after the stimulus onset. The whole-head analysis thus suggest that there were some differences among the groups in the timing and scalp-distribution of the P300 probe-irrelevant effect, suggesting that the groups processed the probes and irrelevant items differently in some way. The earlier onset Probe-Irrelevant difference and trend towards a more sustained effect in the Guilty-Immediate group suggests that the Probe may have been more salient to participants if they were immediately tested, than when tested after a week's delay. According to previous research, the saliency of stimuli can affect test outcome of the CIT such that more salient stimuli tended to produce a larger Probe effect when compared to stimuli that are less salient (Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003; Gamer & Berti, 2012; Verschuere, Kleinberg, & Theocharidou, 2015). However, although the groups seemed to differ in the onset, duration and scalp distribution of the Probe-Irrelevant effect in my study, there was no indication that the Probe peak amplitude differed across groups. Therefore, there were no differences in detection rates when using the typical method to extract P300 and LPN measures, which

involves using a sliding 100ms window to identify those peaks within a larger window, and is thus relatively resistant to changes in peak timing across participants.

As discussed in previous chapters, I found that the aIAT is extreme susceptible to countermeasures. Rehearsing a false alibi can reduce detection rates with the aIAT regardless of whether suspects have been practicing it just briefly immediately before the test, or repeatedly for a week before the test. The alibi countermeasure also seems to increase the implicit truth value of the alibi scenario and alter the test outcome to be more in line with that of an innocent suspect. Unlike the aIAT, which is a reaction time and accuracy-based measure (Sartori & Agosta, 2008), the P300 CIT is a brain-activity based test of memory that provides a more direct measure of the strength of a true memory, since it directly measures participants' neural recognition response towards probe words, which are central details of the crime. Therefore, participants may be more likely to show an automatic recognition response towards such critical information of the crime (although see Bergström, et al., 2013; Hu et al., 2015), compared to an indirect measure like the aIAT. Thus, the CIT may be a superior method for detecting guilt compared to the aIAT.

In this experiment, although I did not find significant results of the alibi manipulation, the findings converge with the results from the aIAT chapters in showing no reduction in strength of the true memory. Such reductions might have been expected if the alibi rehearsal elicited retrieval-induced forgetting (Anderson et al., 2004) or some kind of memory distortion or retroactive interference that prevented access to the true memory (*cf.* Gronau et al., 2015). Instead, the results in this experiment also suggest that our alibi manipulation primarily involved encoding of the false alibi scenario into a

memory that shares some characteristics with a true memory, consistent with the source monitoring framework that suggests that imagined events can be encoded in similar form to truly perceived events (Lyle & Johnson, 2006; Mitchell & Johnson, 2009).

The results also suggested that the CTP version of the P300 CIT is not only resistant to countermeasures like rehearsing a fake alibi before the test, but is also resistant to time delay. Comparing the Guilty-Immediate group to the Guilty-Delay group, there were no significant differences in group-level ERPs or individual detection rates. This finding is in line with previous research, (Hu et al, 2012) that also found that time delay (1 month) did not have an effect on CTP sensitivity. It also contrast with my findings regarding the aIAT, where time delay did seem to influence aIAT sensitivity such that there was a reduction in detecting true memories after a week's delay (see Chapter 3) compared to when the test was administered immediately after the mock crime (see Chapter 2). Thus, it may be possible to detect guilt with the CTP CIT regardless of when the test is administered, which is likely a necessary condition for real-life applications.

Rosenfeld and colleagues (2013) suggested that the CTP has very high detection accuracy levels at around 90% correct classification of Guilty vs. Innocent suspects. My results from the individual classification analysis were weaker, especially when using solely the base-peak P300 or the LPN, however the detection rate was better when I combined these two effects together into the P300-LPN peak-to-peak measure. These results are consistent with Rosenfeld's (2013) findings that the peak-to-peak measure maximises the chance of accurately detecting guilt.

Nevertheless, my individual detection rates were still quite a lot lower than those described previously in the literature, even when using the optimal measure and standard EEG analysis procedures. The discrepancy may be related to using a slightly different method for encouraging participants to pay attention to the screen. In my study, I paused the experiment at some intervals and asked participants to type out the most recent word that they had seen on the screen as an attention check measure, instead of asking them to respond verbally as introduced by Rosenfeld (2008). However, I found that the proportion of attention check questions that were answered correctly was *negatively* correlated with the reliability of an individual's probe-irrelevant ERP difference as assessed with bootstrap resampling, in Guilty-Immediate group only. Thus, poor attention to the probe/irrelevant words was associated with *increased* accuracy of memory detection in this group. This finding is difficult to interpret, but one speculative explanation could be that the Guilty-Immediate group might have focused more on the attention check task rather than thinking about how the probe word related to the mock crime. That is, focusing on word reading for the attention check task may have disrupted the ability of the probe words to function as retrieval cues to activate the mock crime memory. Thus, perhaps participants did not process the probes as crime-reminders during the experiment, and as a result, our detection rates were weaker than typical findings in the literature. However, there could be plenty of other reasons why the detection rates were lower in this experiment than in the literature, for example the quality of the EEG recording, or EEG pre-processing parameters could also contribute to differences across studies. Nevertheless, despite low detection rates at the individual level, group level statistics did show the typical Probe-Irrelevant effects as typically

found by the developers of the CTP (Rosenfeld et al., 2008; 2013), which have previously only been replicated by one other independent group of authors (Lukács, Weiss, Dalos, Kilencz, Tudja, Csifcsak, 2016).

Another limitation with the current study was that I did not include an Innocent control group, so I was not able to compare discrimination performance for the Guilty groups against an Innocent group, which would have made my results more easily comparable to the literature. However, I was able to test my main hypothesis that rehearsing a false alibi would decrease guilt detection compared to a standard guilty group. Nevertheless, future research on countermeasures should include Innocent participants for a more comprehensive assessment of whether those countermeasures are successful.

To conclude, the key finding in Experiment 4 was that the alibi manipulation did not have an effect on ERP-based memory detection with the CTP version of the Concealed Information Test. Thus, consistent with findings from my first three experiments, there was no evidence that the alibi manipulation had directly impaired participants' true memory of conducting the mock crime, suggesting that neural markers of memory in the Concealed Information Test may be robust against counterfactual imagination effects on memory, at least in the way counterfactual imagination was implemented in my study.

Chapter 5: Cognitive mechanisms underlying counterfactual imagination and its neural correlates

In the previous Experiments, I investigated the effects of counterfactual imagination of a false alibi on detection of a mock crime memory in a forensic setting, using both a behavioural autobiographical implicit association test (aIAT, Experiments 1-3) and an EEG-based concealed information test (CIT, Experiment 4). I found that imagining and rehearsing a false alibi can alter the test outcome of the aIAT memory detection test, although this was not found for the CIT, which was unaffected by the false alibi manipulation. Thus, I suggested that imagining a fake alibi can create a memory for that event that might have similar characteristics as a truthful memory, and as a consequence, to the aIAT was not able to distinguish between the true and imagined memory. However, because those experiments were designed to mimic real life memory detection in a relatively ecologically valid way, they also had several limitations in that they were not optimised to investigate the neurocognitive mechanisms underlying counterfactual imagination effects on memory. For example, the design only permitted one true and one imagined memory per participant, which prevents within-participant comparisons of how different degrees of imagination vividness is related to memory changes. The design also did not enable me to assess brain processes associated with counterfactual imagination, or brain processes that distinguish memories for true events versus those for imagined events that the participant falsely believe are true. In the final experiment in this thesis, I therefore focused on investigating the possible mechanisms underlying memory distortion via imagination, and their neural correlates.

According to the literature, there are several theories that can explain how counterfactual imagination affects memory. Two of the major theories are interference theory and inhibitory-control theory. Interference theory suggests that when multiple memories are associated with a common cue, those memories that are more strongly associated to the cue (for example because they have been repeatedly rehearsed in response to the cue) tend to interfere with and block retrieval of memories that are more weakly associated to the cue (for example because they have not been rehearsed; reviewed in Anderson & Neely, 1996). As a consequence, the stronger memory comes to mind in response to the shared cue and prevents conscious retrieval of the weaker memory. Thus, it could be the case that repeatedly imagining a counterfactual version of an event strengthens that false event memory, which then blocks access to the true memory when tested with a shared cue.

In contrast, inhibitory-control theory argues that repeatedly practiced memories may be strengthen by rehearsal, but that selective rehearsal of those memories can also inhibit other unpractised memories that are associated with the same cue. Evidence for such Retrieval-Induced Forgetting (RIF, reviewed in Levy & Anderson, 2002) comes from the finding that unpractised memories that are associated to the same cues as repeatedly rehearsed memories are more likely to be forgotten even when tested with different, independent cues that are not associated with the strengthened memory. Since strengthened memories are unlikely to be retrieved in response to the independent cues and therefore won't be able to "block" retrieval of the unpractised memories, that is considered strong evidence that the unpractised memories have become inhibited. Likewise, the finding that such unpractised memories are impaired also on recognition

tests where the unpractised memory is cued with a copy of the original stimulus is also considered evidence for inhibition (e.g. Hicks & Starns, 2004), since such recognition cues are very strong reminders of the unpractised memory, and are less likely to elicit retrieval of the strengthened competitor memory and leading to blocking. According to inhibition theory, counterfactual imagination of alternative versions to past events could inhibit the true memory of what really happened, causing the true memory to become forgotten. In my previous experiments I was not able to test these theoretical accounts in an optimal way, but the current experiment was designed to assess evidence for interference versus inhibition of memories as a result of counterfactual imagination, by testing memory with both cues that were shared between the true memory and the counterfactual imagination, and with a recognition test for true memories.

Moreover, in addition to behavioural measures of interference and inhibition, I also recorded ERPs during counterfactual imagination and during subsequent memory testing to assess the neural activity associated with counterfactual imagination effects on memory. Prior research has shown that a posterior ERP positivity that peaked around 600-900ms after participants were cued with a word to imagine an object represented by that word, can predict that participants will later mistake imagined pictures as having been seen (Gonsalves & Paller, 2000). This effect was suggested to be a marker of visual imagination, and it was argued that increased vividness of imagination was the underlying cause of the later false memory because vivid imagination created a false memory with similar perceptual characteristics as a true memory of an object picture. Furthermore, the same study also suggested that it is possible to distinguish true and imagination-induced false memories at subsequent retrieval using ERPs. It was found

that ERP responses were more positive at parietal and occipital sites for true memories than false memories from 900-1200ms after the stimulus onset (Gonsalves & Paller, 2000; see also Gonsalves & Paller, 2002; Gonsalves et al., 2004). However, that line of research did not assess *counterfactual* imagination of alternative versions of past events, as addressed in the current experiment.

As introduced in Chapter 1, other ERP components associated with familiarity, recollection and post-retrieval monitoring are also often observed during retrieval tasks, and I was interested in assessing how counterfactual imagination would affect these ERP effects. The FN400 refers to frontal and central positivities around 300-500ms after stimulus onset, for previously encountered stimuli in recognition tasks. It had been suggested that the FN400 is an index of familiarity, and this effect is often found during successful recognition of stimuli, regardless of whether participants can remember the context where they encountered those stimuli (Rugg & Curran, 2007). The left parietal old/new effect usually correlates with recollection of contextual information associated with a stimulus, and is maximal between 500-800ms after stimulus onset (Duarte, Ranganath, Winward, Hayward, & Knight, 2004). A frontal positivity, often right lateralised, emerging around 500ms after stimulus onset and sustained for up to several seconds, is thought to index the involvement of executive control processes during retrieval (e.g. monitoring and retrieval effort; Leynes, Cairns, & Crawford, 2005; Rosburg, Mecklinger, & Johansson, 2011; see for review Wilding & Ranganath, 2011).

Experiment 5

This study aimed to investigate cognitive and neural mechanisms underlying memory distortion via counterfactual imagination using both behavioural and neural correlates. The study involved three main parts. On the first day, participants were provided with real objects and were asked to perform a simple action involving each object (this task was based on Brandt, Bergstrom, Buda, Henson, & Simons, 2014). Then a week later, participants were presented with photographs of the objects on a computer screen, and were asked to either imagine performing the same action they had done on day 1 for 1/3 of the objects, or imagine performing a new, counterfactual action for another 1/3 of objects. This imagination task was repeated three times per object while continuous EEG was recorded. Participants were asked to imagine performing the action a vividly as they could, and after each trial they were asked to rate the vividness of their imagination. The final 1/3 of objects were not shown, but were included as a behavioural baseline for the subsequent memory tests. Finally, participants took cued recall and recognition test, also with simultaneous EEG recordings. In the cued recall test, they were shown the object photographs for all conditions and asked to recall the action they had performed on day 1 (ignoring what actions they had imagined in the previous task), and in the recognition test they were shown the object photographs together with a description of the action from day 1, or a completely new action they had not previously performed or imagined, and were asked to judge whether the action for each object was the same they had performed on day 1.

I predicted that participants should have better memory for true actions after repeatedly imagining performing the actions that they had previously truly performed on

day 1 (i.e. when they rehearsed a true memory during the imagination task) compared to when they repeatedly imagined performing a new, counterfactual action that conflicted with the true action they had performed on day 1. Importantly, if counterfactual imagination actually impaired memory, participants should be poorer at recalling the true action after counterfactual imagination when compared to the baseline condition that had not been shown in the imagination task and thus assessed simple forgetting over time. Such below-baseline impairments are often found and used to indicate interference- or inhibition-related memory impairments in different paradigms (e.g. AB-AC interference, see Anderson, 2003 or the Think/No-Think paradigm, Anderson & Green, 2001). With regards to contrasting the interference vs. inhibition theories, both theories predict that performance on the cued recall test will be impaired and that participants should sometimes recall a counterfactual imagined action instead of the original action (referred to as an "intrusion error" in the interference literature), because the counterfactual action memory has been strengthened by the imagination manipulation and may be retrieved instead of the true memory in response to the shared cue. Thus, the cued recall test is not able to distinguish whether the impairment is due only to strengthening of the false memory, or also inhibition of the true memory. However in the recognition test, if participants are impaired at recognising the true actions for objects that had previously been associated with counterfactual actions, this would be more consistent with inhibition, since the recognition test provided a strong direct cue for the true action and thus recognition should be less susceptible to interference (Hicks & Starns, 2004). Therefore, the interference account predicts that counterfactual imagination will cause true memory impairments primarily on the cued

recall test and not the recognition test, whereas the inhibition account predicts impairments on both tests.

With regards to the ERP predictions, I analysed ERPs from both the imagination phase and the subsequent cued recall test. For the imagination phase, I assessed ERPs for both rehearsed true actions and counterfactually imagined actions, and divided the latter on the basis of subsequent memory performance. That is, I compared ERPs during counterfactual imagination separately for objects for which the correct true action would later be recalled on the cued recall test, with objects for which the counterfactual action would be incorrectly recalled as true (intrusion errors). If the vividness of counterfactual imagination is a critical factor for producing subsequent misremembering, then ERP effects that predict subsequent intrusion errors might be similar to the posterior ERP positivity described in the previous literature that was linked to vivid imagination and predicted mistaking imagined pictures as perceived (Gonsalves & Paller, 2002). ERPs during subsequent cued recall testing were predicted to differ both on the basis of whether the cue picture had been previously shown, and on the basis of whether the true action had been rehearsed or not. Repeatedly viewing pictures in the imagination phase should elicit recognition when those pictures were shown again in the test. Therefore, I would expect pictures associated with both rehearsal of true actions and counterfactual imagination of false actions to elicit larger ERP correlates of recognition, such as the FN400 and left parietal old/new effects (Rugg & Curran, 2007) compared to the baseline condition for which the pictures had not been repeatedly viewed. More interestingly however, I also wanted to assess how ERP effects associated with cued recall of the associated action were modulated by prior

counterfactual imagination. Based on previous findings that false memories of imagined events elicit lower amplitude parietal ERP positivities (Gonsalves & Paller, 2000), I predicted a reduction in parietal ERP positivity for the counterfactual imagination condition compared to the rehearsal condition. The counterfactual imagination condition might also be more likely to elicit ERP correlates of executive control-related processes, since participants would likely need to monitor their memories closely in this condition, and potentially select between competing true and false memories. Executive control involvement during retrieval is often associated with late right frontal ERP positivities (*cf*. Hanslmayr, Leipold, Pastötter, & Bäuml, 2009; Hayama et al., 2008; Johansson, Aslan, Bauml, Gabel, & Mecklinger, 2007), so I expected to observe such frontal positivities specifically in the counterfactual imagination condition.

Methods

Participants

Thirty undergraduate students were recruited from the School of Psychology's Research Participation Scheme (RPS), and received course credits or monetary compensation (£25) for their participation. However, six participants were excluded from the analyses due to excessive noise in the EEG data. The final sample consisted of 24 participants[1] (18 female and 6 male; age range 18-25; M_{age} = 20.42, SD = 1.93). All

[1] An additional 12 participants completed a behavioural version of the experiment that was identical in design with the only exception that no EEG was recorded. The behavioural results were highly similar across both groups, and combining data from the two groups produced the same results as reported in this chapter, but with more significant effects due to increased power.

participants reported an absence of neurological and psychiatric disorders, and had normal or corrected-to-normal vision. The study was approved by the University of Kent Psychology Ethics committee.

Materials, Design, and Procedures

Stage 1. Learning phase. In the first session, participants learned object-action associations by handling real objects and performing actions with those objects (based closely on Brandt et al., 2014). Stimuli included 120 small every-day objects (or sets of objects), including items such as clothing, kitchen utensils, toys, tools, and stationery and 360 action statements, which involved three potential actions that could naturally be performed with each object. For example, if the objects were a pair of dice, the actions statements could be "roll the dice", "stack the dice" and "place with the 5s facing up". One of the three action statements was presented to be learned with the object, and participants learned each of the 120 object-action associations one-by-one. At the beginning of every trial, the experimenter took out an object from a box and read aloud the action statement. Participants were instructed to perform the action with the object (e.g. "roll the dice"). The order of presenting objects was the same for every participant, but the assignment of objects and action statements to conditions was counterbalanced across participants, thus ensuring no order differences between conditions across the full sample. Participants were told that the experiment was about how people perform and

imagine actions, and there was no mention that their memory for the actions would later be tested. Thus, learning was incidental.

Stage 2. Manipulation phase

Stage 2.1 Practice imagination task (80 trials). In the next step conducted 1 week after the learning phase, the experimenter first set up the EEG recording, and participants were instructed to avoid eye-movements and excessive blinking during all subsequent phases that all included EEG measurements. Then, they proceeded with the practice imagination task (see Figure 5.1A). Participants were first presented with a fixation cross in the centre of the computer screen (duration 0.75s), then they were presented with a picture of the object they had handled in the first session, together with an action statement (duration 6s). For example, the picture presented in the centre of the screen could show a pair of dice, with the action instruction underneath the picture to "roll the dice". For 40 of the object pictures the sentence described the action they had actually conducted in session 1 (the "Rehearsed" condition), but for the other 40 it described one of the alternative actions for that object that they had not previously seen or performed (e.g. "place with the 5s facing up"; the "Imagined" condition). Trials for these two conditions were presented randomly intermixed.

For all objects, participants were instructed to imagine conducting the action listed under the object picture vividly and in as much detail as possible. They were also told that some of the actions might be the same as those they had conducted on day 1 whilst other actions might be different, but that this was not important and they should always just focus on imaging the action that was described under the picture (they were

also not told which actions were same or different) without thinking about the previous session. The purpose of these instructions was to avoid participants intentionally trying to remember the original actions. After the picture disappeared off the screen, participants were shown a rating scale and asked to rate the extent to which they had been able to imagine the action vividly (1-6 scale; 1 = not vivid at all to 6 = extremely vivid). This scale stayed on the screen until participants pressed a button, after which the next trial started. There were two minutes break after every 20 trials. The remaining 40 object-action pairs that had been learned on day 1 were not shown in this phase, so functioned as a "Baseline" condition for the subsequent memory tests.

Stage 2.2 Imagination task with shorten sentences (160 trials). After the initial practice, all participants continued with the same imagination task as described above (Figure 5.1B), but with the only difference that the action statements were shown in an abbreviated form (e.g. "place 5s up"). The purpose of using abbreviations was to reduce the amount of text on the screen to minimise eye movement-related noise in the EEG. Each object-action pair was repeated twice in this phase, in random order.

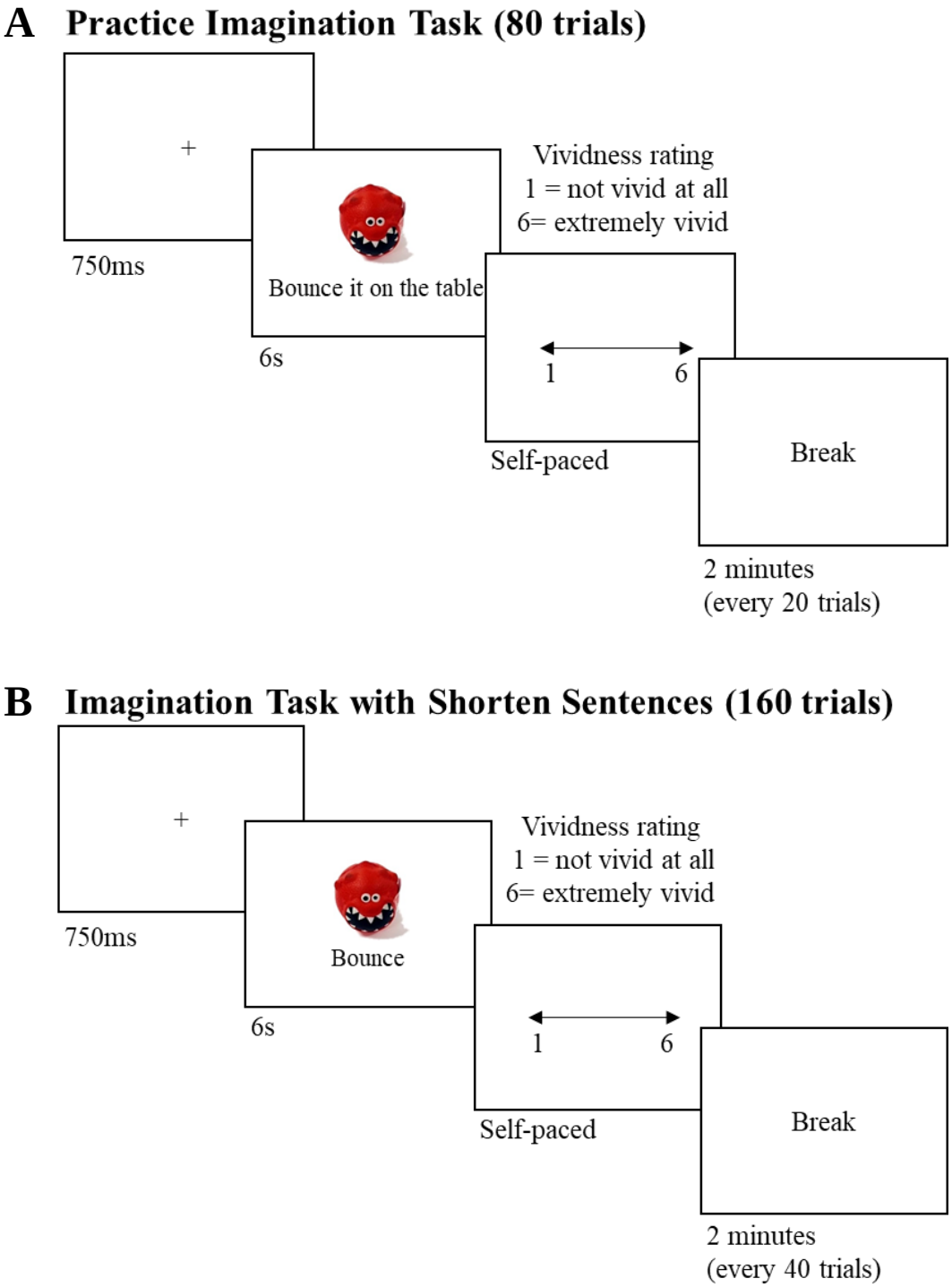

Figure 5.1. Imagination task procedure and example stimuli.

Stage 3. Test phase.

Phase 3.1 Cued recall test (120 trials). After completing the imagination task, all participants were given two surprise memory tests (see Figure 5.2A) to assess whether they could remember the real actions that they had performed on day 1 for all the objects (from Imagined, Rehearsed and Baseline conditions). The first test was a cued recall task, where participants were presented with a picture of an object (6s, preceded by a fixation cross for 0.75s) and asked what action they had really performed with the presented object in the first session. It was made clear that this might have been different from the action they had imagined in the imagination task, and that they should always try to think back to day 1 and report the real action they had conducted. After the object picture went off the screen, participant had 5s to respond verbally and the audio from this phase was recorded and scored off-line. Participants stated what action they had performed and also rated their confidence in the decision by saying out loud a number between 1 to 6, where 6 indicated very high confidence and 1 indicated very low confidence. Object pictures were presented in a random order with two minutes break after every 60 trials.

Phase 3.2. Associative recognition test (120 trials). In a final associative recognition task (Figure 5.2B), participants were presented with a picture and an action statement (60 performed and 60 completely new actions, equally split across the three conditions to create six combinations with 20 trials in each: Rehearsed Old vs. New, Imagined Old vs. New and Baseline Old vs. New), and were asked to decide whether the action shown with the picture was old (original action performed on day 1) or new

(completely new and not shown previously with that object). These pairs were presented for 2.5s and preceded by a fixation cross for 0.75s. After the picture-action pairs went off the screen, participants responded by pressing a button (1-6) in which 1 indicated a very confident "new" response, 2 indicated a moderately confident "new" response, 3 indicated a guess "new" response, 4 indicated a guess "old" response, 5 indicated moderately confident "old" response and 6 indicated a very confident "old" response. Again, participants were given a two minutes break after every 60 trials and stimuli were presented in a random order.

Finally after all the experimental phases, participants also filled in a brief questionnaire to assess their self-reported compliance with instructions (see Appendix I). Participants were asked to rate on a scale from 0-6 the extent to which they were thinking back to the learning phase whilst imagining the actions in the imagination phase (0 = *not at all*; 6 = *always*), how accurate they felt their vividness ratings had been (0 = *never*; 6 = *always*), and whether they had reported their vividness rating inaccurately on purpose (0 = *never*; 6 = *always*).

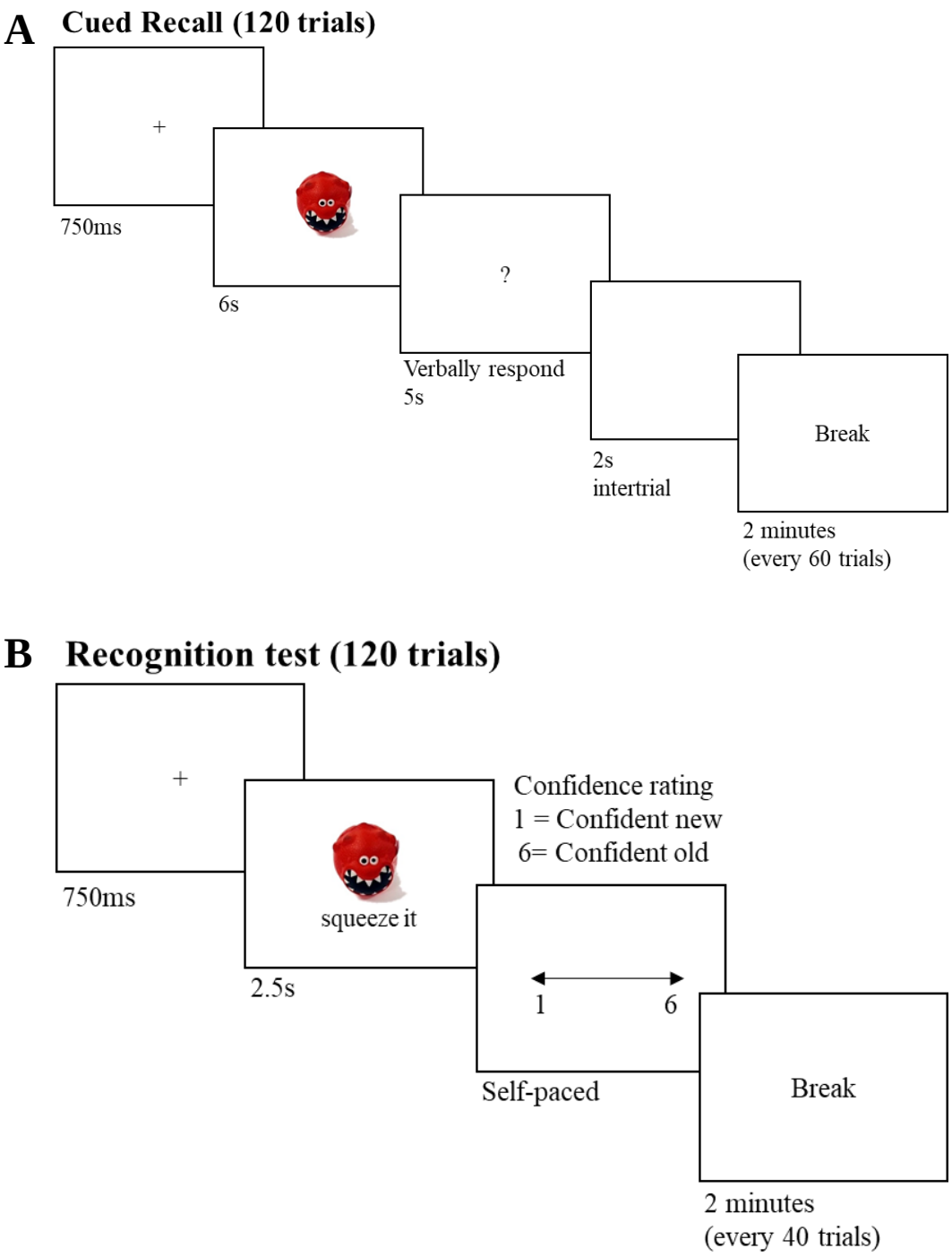

Figure 5.2. Surprise cued recall (A) and associative recognition test (B) procedures.

EEG recording and pre-processing

Continuous EEG were recorded at 500 Hz with a 0.1 to 70 Hz bandwidth, from 32 scalp electrodes (using the extended 10-20 system) placed in an EasyCap, referenced against an average reference. The electrooculogram (EOG) was recorded vertically (VEOG) and horizontally (HEOG) from the left and right eye. The data were analysed using EEGLAB. The EEG was re-referenced to mastoid channels, excluding EOG channels. Then, I divided the continuous EEG into two files, the first containing 360 epochs (-0.75 – 3.5 sec) from the imagination task and cued recall phase, and the second containing 120 epochs (-0.75– 2.5 sec) from the recognition phase, both time-locked to the onset of the object picture + action presentation (or picture only presentation in the cued recall phase, as no action statement was presented there). After that, I merged these two files together and manually scrolled through the file to remove excessive noise segments or channels. Then, the concatenated EEG data were submitted to infomax Independent Component Analysis (ICA) using runica. Independent components reflecting eye movements and other artefacts were identified by visual inspection of their topography, time-course and spectral profile, following recommendations in the EEGLAB manual. Noise components were removed from the data, which subsequently was also low-pass filtered at 40Hz. Finally, any trials that still contained artefacts were removed. However, despite these steps for artefact removal, the EEG data from the recognition test was very noisy due to low trial numbers per conditions coupled with some residual eye-movement noise. Therefore, I only proceeded to statistically analyse and present the EEG data from the imagination task and the cued recall test. The mean trial numbers for ERPs in the imagination phase were as follows: Rehearsed subsequent

correct (Range = 75-119, *M* = 104, *SD* = 13), Imagined subsequent correct (Range = 9-81, *M* = 48, *SD* = 19), Imagined subsequent intrusion (Range = 12-108, *M* = 41, *SD* = 27). The mean trial numbers for the cued recall phase were similar across conditions as they were not conditionalised on responses (Rehearsed (Range = 34-40, *M* = 39, *SD* = 2), Imagined (Range = 33-40, *M* = 39, *SD* = 2), and Baseline (Range = 31-40, *M* = 39, *SD* = 2).

Results

Behavioural Results

Compliance. Self-report questionnaire revealed that most participants complied with the imagination task instructions given by the experimenter. They rarely intentionally thought about the performed action during the imagination task (Range: 0-5; *M* = 2.04, *SD* = 1.65), and were generally accurate when they rated the vividness of their imagination (Range: 2-6; *M* = 4.54, *SD* = 1.10), and only very rarely gave vividness ratings that were inaccurate on purpose (Range: 0-4; *M*= 0.96, *SD* = 1.20). Although a few participants reported that they did not fully comply with the instructions, these were not excluded from the analyses since doing so would lead to an insufficient sample size. In addition, participants often explained in a free text section of the questionnaire that they did comply with the instructions even when their ratings indicated otherwise. For instance, participants explained that they did not think about the learning phase intentionally during the imagination phase (which they were asked not to), but that the action memory from day 1 sometimes automatically came to mind. Therefore, I decided that it would be most appropriate to retain the full sample.

Imagination Phase. An initial analysis compared vividness ratings during the imagination task based on both experimental conditions (Rehearsed vs. Imagined) and subsequent memory on the cued recall task, to investigate whether vividness ratings during the imagination of counterfactual actions would predict that an imagined action would later be misreported as a performed action. To conduct this analysis, Imagined trials were split based on whether the correct action would later be reported on the cued recall task ("Imagined Correct"), or whether participants would instead incorrectly recall the imagined action (henceforth referred to as an "intrusion error", "Imagined Intrusion"), and the mean vividness rating across repetitions was calculated for these categories. I also extracted the mean vividness ratings for rehearsed actions that would go on to be correctly remembered on the cued recall test ("Rehearsed Correct"). However, because cued recall performance was very high in this condition there were insufficient trials for looking at Rehearsed trials associated with later errors.

A one-way repeated measures ANOVA was conducted to investigate the difference in vividness ratings among these conditions (see Figure 5.3). Mauchly's test of sphericity suggested that the assumption of sphericity was met ($\chi^2(2) = 5.55$, $p = .062$). The result revealed that vividness ratings significantly differed among conditions ($F(2,46) = 7.35$, $p = .002$, $\eta_p^2 = .242$). Follow up paired t-tests revealed that vividness ratings for Rehearsed Correct items was significantly higher ($M = 4.82$, $SD = .81$) than for Imagined Correct items ($M = 4.27$, $SD = 1.00$; $t(23) = 3.19$, $p = .004$, $d = .60$). There was also a marginally significant increase in vividness in the Rehearsed Correct compared to Imagined Intrusion trials ($M = 4.50$, $SD = .93$; $t(23) = 2.26$, $p = .034$, $d =$

.37), and a marginally significant increase in vividness for Imagined Intrusion than Imagined Correct trials ($t(23) = 2.08$, $p = .049$, $d = .24$). Thus, this analysis showed that vivid imagination of a counterfactual action predicted later misremembering of that action as performed.

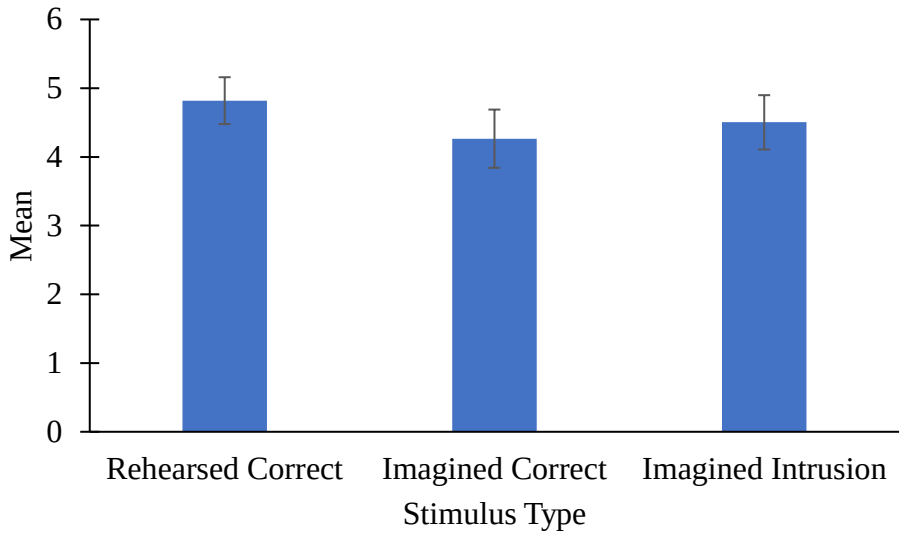

Figure 5.3. Means vividness rating for Rehearsed and Imagined items in the imagination task, as a function of later cued recall accuracy. Error bars denote the 95% confidence interval.

Cued Recall Phase. For the cued recall phase, a one-way repeated measures ANOVA was conducted to examine the differences in proportion of accurate responses for the Rehearsed, Imagined and Baseline conditions, to investigate whether imagining a counterfactual action would reduce memory accuracy (see Figure 5.4). The assumption of sphericity was violated, as assessed by Mauchly's test of sphericity ($\chi^2(2) = 7.40$, $p = .025$). Therefore, a Greenhouse-Geisser correction was applied ($\varepsilon = .78$). There were significant differences in accuracy between conditions ($F(1.56,35.78) = 107.84$, $p < .001$, $\eta_p^2 = .824$). Paired t-tests revealed that participants were more accurate at recalling actions from day 1 for the Rehearsed condition ($M = .86$, $SD = .13$) than both the

Imagined condition (M = .40, SD = .17; $t(23) = 11.31$, $p < .001$, d = 3.04) and the Baseline condition (M = .48, SD = .14; $t(23) = 15.31$, $p < .001$, d = 2.81). Participants were also less accurate at recalling true actions for Imagined than Baseline objects ($t(23) = 2.41$, p = .024, d = .51), thus showing a memory impairment as a result of counterfactual imagination beyond simple forgetting over time.

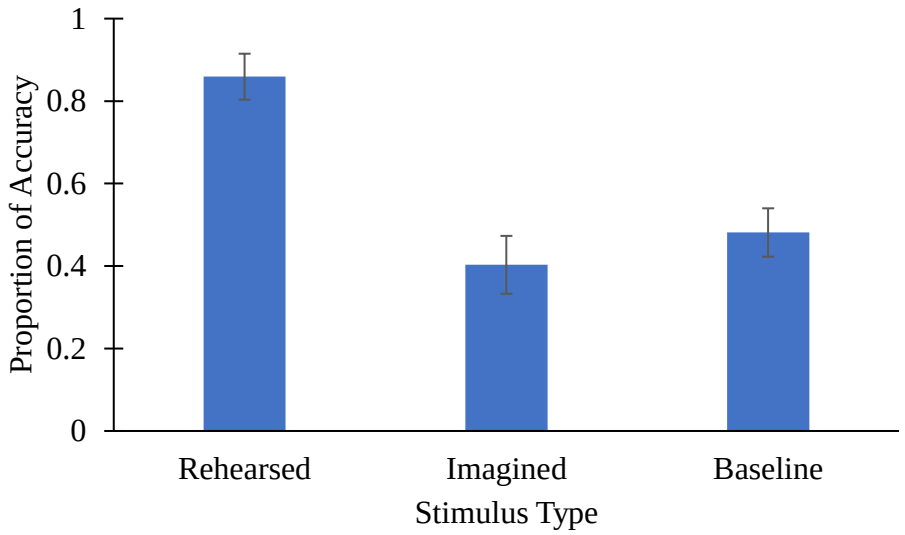

Figure 5.4. Proportion of accurate responses for each condition on the cued recall test. Error bars denote the 95% confidence interval.

Confidence ratings were also analysed to investigate participants' subjective experience of memory for the different conditions in the cued recall phase. Mauchly's test of sphericity suggested that the assumption of sphericity was met ($\chi^2(2) = 1.97$, p = .373). A one-way repeated measures ANOVA showed that confidence ratings were significantly different across conditions ($F(2,46) = 31.09$, $p < .001$, η_p^2 = .58; see Figure 5.5). Follow-up paired t-tests showed that confidence ratings were higher for the Rehearsed condition (M = 5.10, SD = .63) when compared to both the Imagined (M = 4.41, SD = .80; $t(23) = 7.39$, $p < .001$, d = .96), and the Baseline conditions (M = 4.31,

SD = .74; *t*(23) = 6.44, *p* <.001, *d* = 1.15). However, there was no difference in cued recall confidence between Imagined and Baseline conditions (*t*(23) = .83, *p* = .414, *d* = .13). Thus, despite the lower accuracy for Imagined than Baseline items, confidence was similar.

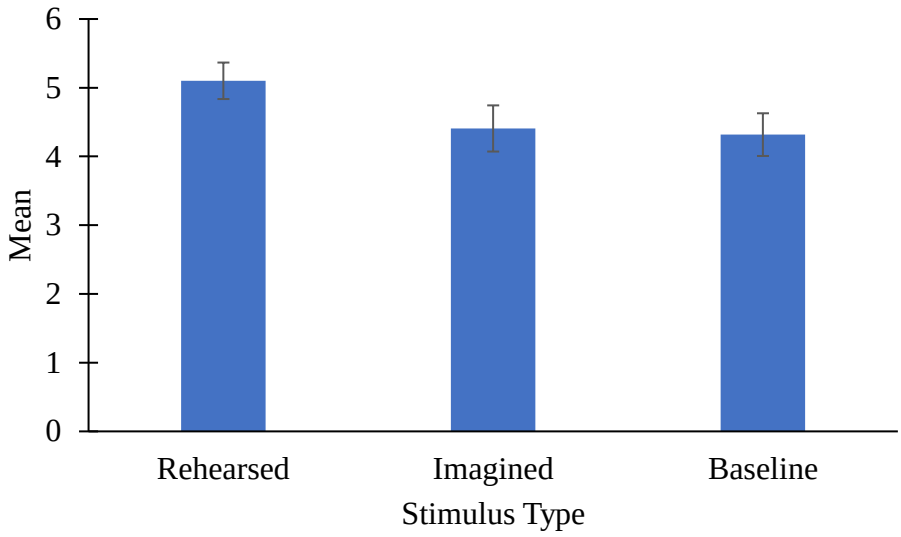

Figure 5.5. Average confidence ratings for each condition in the cued recall phase. Error bars denote the 95% confidence interval.

Associative recognition phase. Proportion accurate responses for each condition on the associative recognition test was first calculated by simply classifying responses as correct or incorrect, regardless of confidence (see Figure. 5.6). A 3 (Manipulation Condition: Rehearsed, Imagined, or Baseline) x 2 (Action Type: Old or New) two-way repeated measures ANOVA was conducted to determine if there were differences in proportion accurate recognition responses among the three conditions, and whether this differed depending on whether the action was Old (conducted on day 1) or New (not seen before in the experiment). Mauchly's test of sphericity revealed that there was sphericity for the interaction term ($\chi^2(2) = 4.43$, $p = .109$). Results revealed a

significant Condition x Action Type interaction ($F(2,46) = 5.17$, $p = .009$, $\eta_p^2 = .183$). One-way repeated measures ANOVAs were conducted to compare Conditions separately for Old and New actions, to follow up the significant interaction. Results showed that there were significant differences between conditions in proportion accurate responses for both old actions ($F(2,46) = 19.00$, $p < .001$, $\eta_p^2 = .452$) and new actions ($F(2,46) = 13.59$, $p < .001$, $\eta_p^2 = .371$). For the Old actions, paired t-tests revealed that accuracy for the Rehearsed condition ($M = .94$, $SD = .08$) was higher than in the Imagined condition ($M = .72$, $SD = .19$; $t(23) = 5.28$, $p < .001$, $d = 1.51$). Recognition accuracy for old actions was also higher in the Rehearsal than Baseline condition ($M = .85$, $SD = .13$; $t(23) = 3.41$, $p = .002$, $d = .83$). Interestingly, participants were also more accurate at recognising Old actions in the Baseline than the Imagined condition ($t(23) = 3.43$, $p = .002$, $d = .80$), despite not having seen those actions since day 1 in either of those two conditions. For new actions, results showed that the accuracy at detecting an action as new was significantly higher in the Rehearsed ($M = .90$, $SD = .08$) than both the Imagined ($M = .80$, $SD = .12$; $t(23) = 4.59$, $p < .001$, $d = .98$), and Baseline conditions ($M = .77$, $SD = .12$; $t(23) = 5.13$, $p < .001$, $d = 1.27$). However, there was no difference in participants ability to detect an action as new between Imagined and Baseline conditions ($t(23) = 1.03$, $p = .312$, $d = .25$). Thus, the associative recognition test showed that the counterfactual imagination manipulation specifically impaired participants' ability to recognise true actions as previously performed for those objects, but did not affect their ability to detect actions as new for those objects, when compared to the Baseline condition.

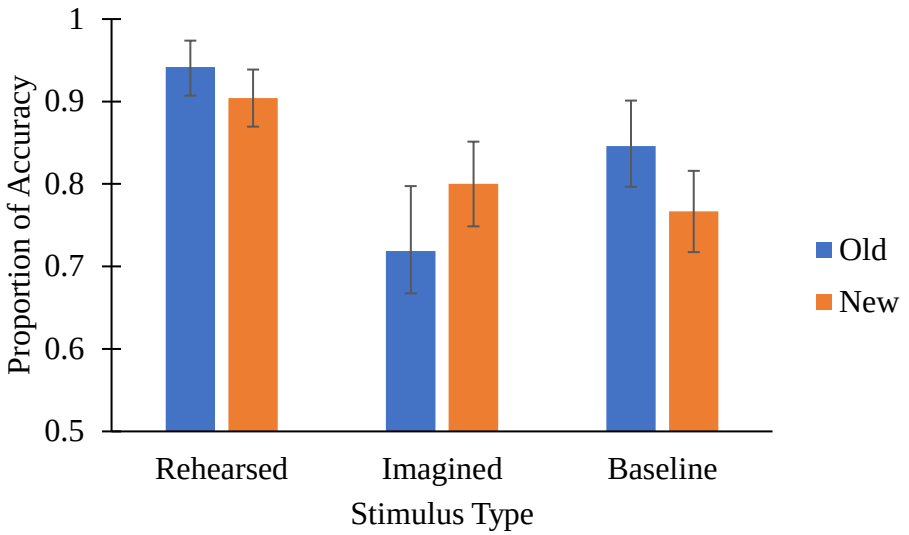

Figure 5.6. Proportion of accurate responses for each condition in the associative recognition phase. Error bars denote the 95% confidence interval.

Confidence ratings were further analysed to gain more insight into participants' subjective experience of recognition by extracting their confidence in decisions for each condition, regardless of accuracy (see Figure 5.7). A 3 (Manipulation Condition: Rehearsed, Imagined, or Baseline) x 2 (Action Type: Old or New) two-way repeated measures ANOVA revealed a significant two-way interaction ($F(2,46) = 18.78$, $p < .001$, $\eta_p^2 = .449$). Two one-way repeated measures ANOVA were conducted separately for Old and New actions to follow up the significant interaction. Results revealed that confidence ratings were significantly different among manipulation conditions for both Old ($F(2,46) = 5.92$, $p = .005$, $\eta_p^2 = .205$) and New actions ($F(2,46) = 15.09$, $p < .001$, $\eta_p^2 = .396$). For the Old actions, paired t-tests showed that confidence ratings were significantly higher for the Rehearsed condition ($M = 2.68$, $SD = .25$) than the Imagined (M = 2.52, $SD = .30$; $t(23) = 2.79$, $p = .010$, $d = .58$) and the Baseline condition ($M = 2.53$, $SD = .30$; $t(23) = 2.97$, $p = .007$, $d = .54$). However, there was no difference between Imagined and Baseline conditions in confidence ratings ($t(23) = 0.31$, $p = .758$,

d = .03) for Old actions. A similar pattern was found for the New actions; decisions in the Rehearsed condition (*M* = 2.63, *SD* = .26) was rated more confident than those in the Imagined (*M* = 2.39, *SD* = .29; *t*(23) = 5.42, *p* < .001, *d* = .87) and the Baseline conditions (*M* = 2.42, *SD* = .21; *t*(23) = 3.97, *p* = .001, *d* = .89). There was no difference in confidence ratings for New actions between Imagined and Baseline conditions (*t*(23) = 0.79, *p* = .440, *d* = .12). Thus, the recognition test showed a similar result as the cued recall test: although participants' true memory accuracy was reduced after counterfactual imagination when compared to a baseline condition measuring simple forgetting over time, their confidence was equivalent in these two conditions.

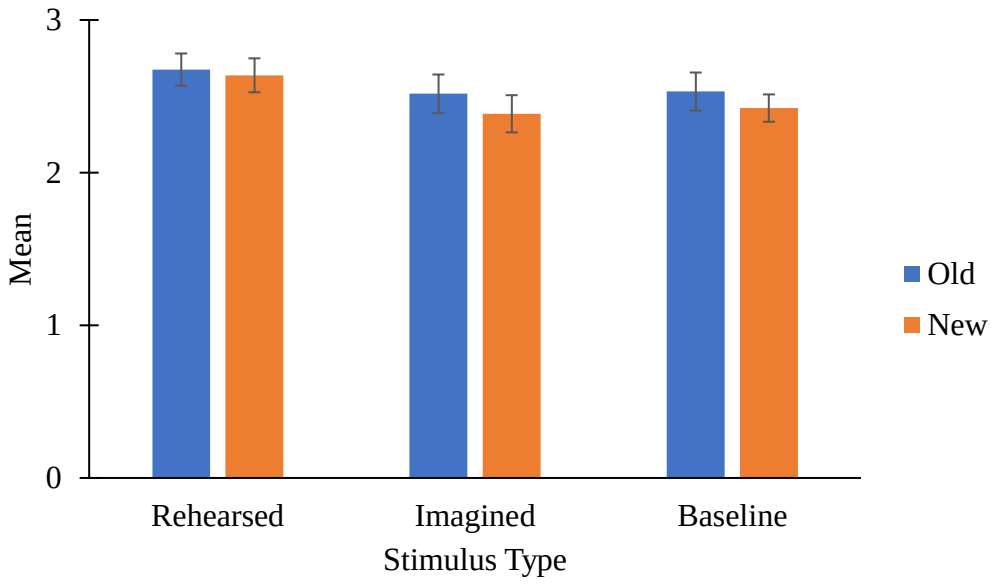

Figure 5.7. Mean confidence ratings for each condition in the associative recognition phase. Error bars denote the 95% confidence interval.

ERP results

ERPs reported here were recorded during the imagination task and the cued recall phase. For both phases, I analysed a grid of electrodes that covered the left and right hemisphere for frontopolar, frontal, central, parietal and occipital electrode sites (FP1, FP2, F3, F4, C3, C4, P3, P4, O1, and O2). The mean amplitudes for each condition at each electrode were calculated for 10 time-windows (0-200ms, 200-400ms, 400-600ms, 600-800ms, 800-1000ms, 1000-1500ms, 1500-2000ms, 2000-2500ms, 2500-3000ms, and 3000-3500ms). Because of the relatively exploratory research questions for ERPs in this novel paradigm, I used this whole-head approach with successive time-windows to assess possible condition differences across locations and time with three-way omnibus repeated measure ANOVAs for each time-window, including the factors Condition (3 levels that differed for the two phases, as explained in the next sections), Anterior-Posterior location (5 levels: Frontopolar, Frontal, Central, Parietal and Occipital) and Hemisphere (2 levels: Left vs. Right). Simple effects analyses with paired t-tests were conducted to follow up significant ANOVA results only if they involved the experimental conditions as a factor (since other effects are not meaningful to interpret).

Imagination phase ERP results. For the imagination phase, I used the same three ERP conditions as in the behavioural analysis of whether vividness ratings predicted subsequent intrusion errors. That is, separate ERPs were created for those Imagined trials where participants subsequently went on to make an intrusion error in the cued recall test (i.e. inaccurately reporting the counterfactual imagined action instead of the action they performed on day 1), versus Imagined trials where participants

subsequently went on to make a correct response in the cued recall test (reporting the action from day 1). ERPs for the Rehearsed condition only included trials where participants later reported the correct action on the cued recall test. Grand-average ERPs from the 10 electrodes sites for each condition are showed in Figure 5.8 for the whole ERP epoch, and in Figure 5.9 for -200 to 800ms. Scalp topographic maps showing the amplitude differences between conditions across 3.5 seconds are shown in Figure 5.10.

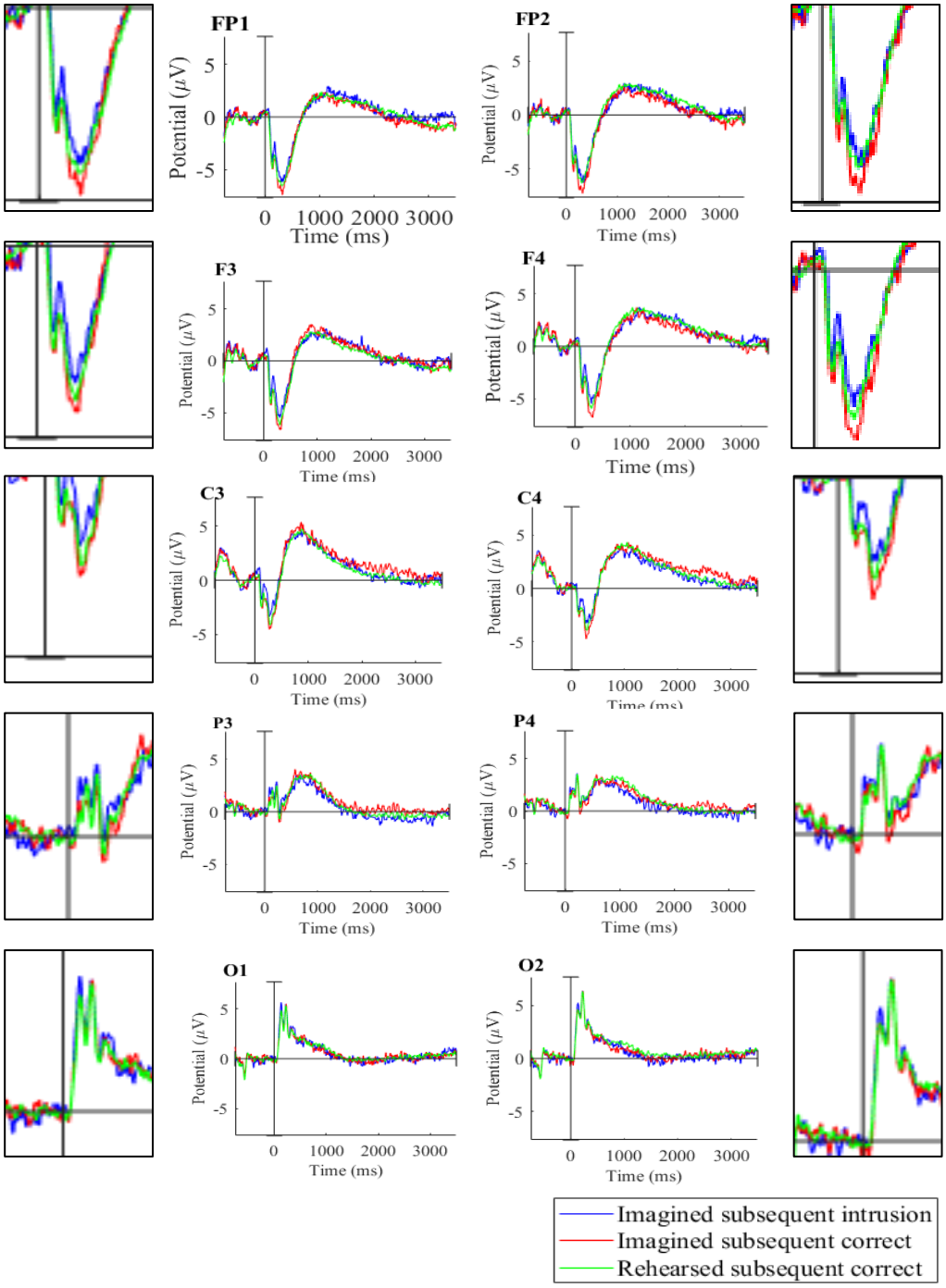

Figure 5.8. Grand-average ERPs from the three conditions in the Imagination task. The boxes on the sides show the first part of the epoch magnified to illustrate the early ERP modulations more clearly.

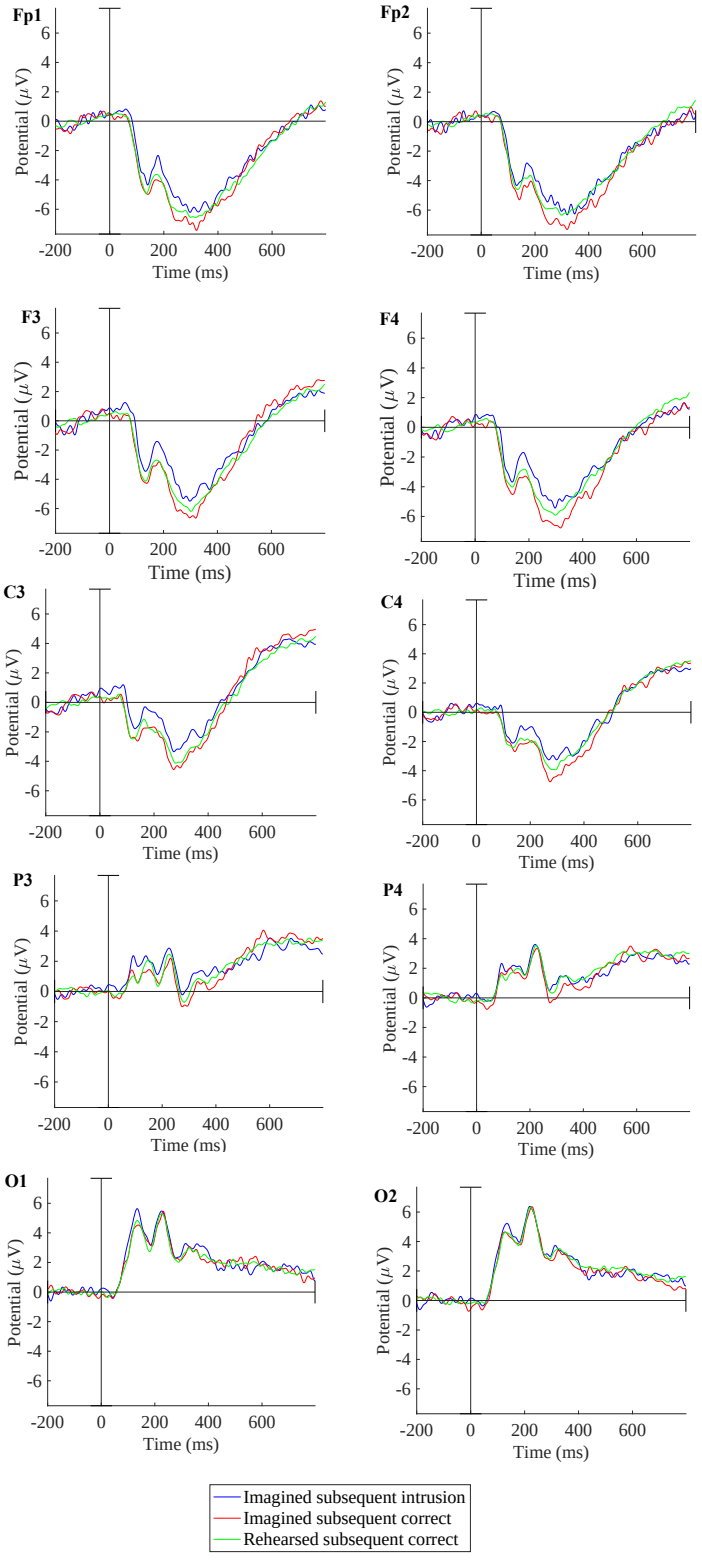

Figure 5.9. Grand-average ERPs from the three conditions in the Imagination task for -200ms to 800ms after stimulus onset.

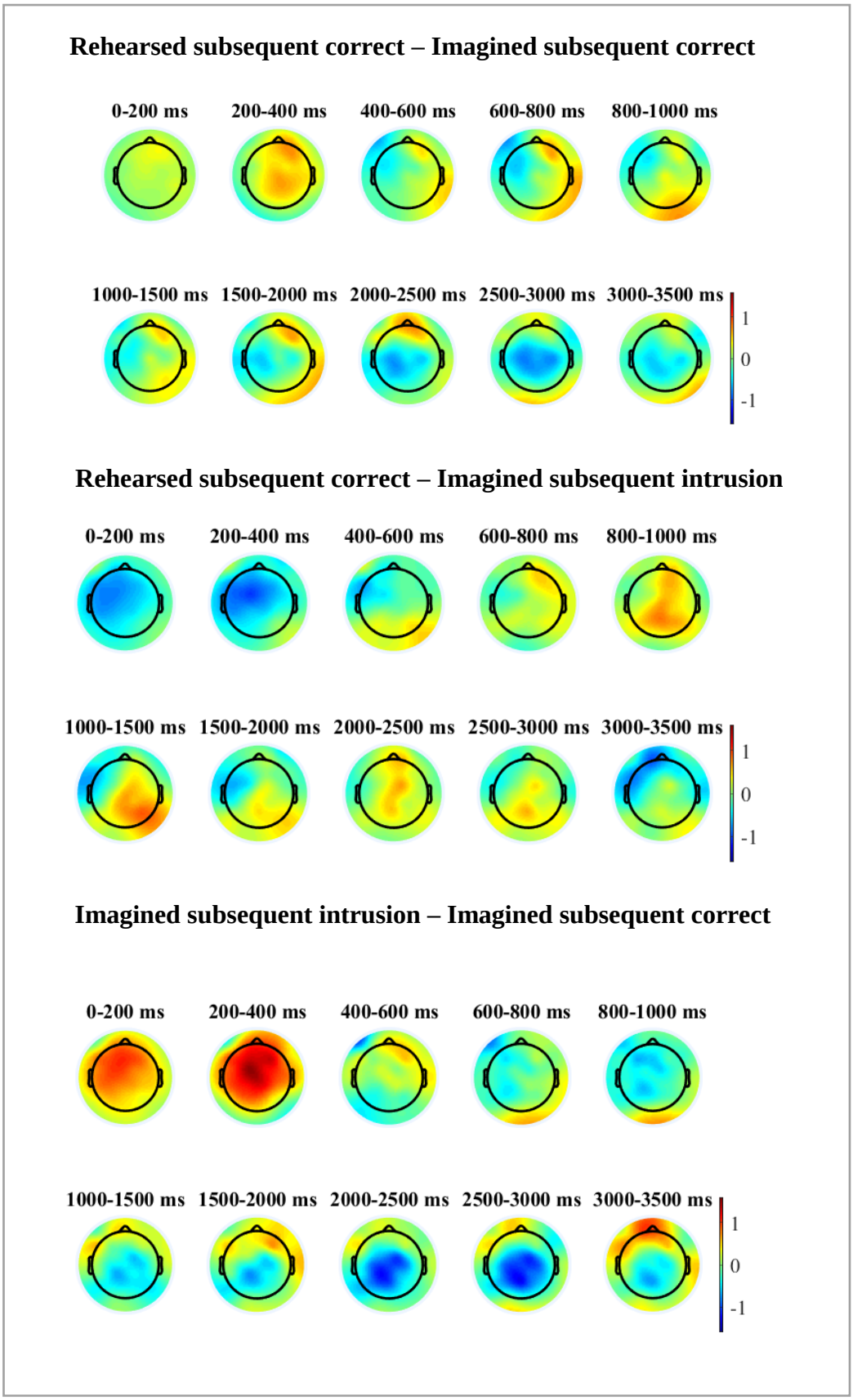

Figure 5.10. Topographic maps showing the scalp distribution of ERP amplitude differences between conditions during the imagination phase.

Results from the omnibus ANOVAs (with factors Anterior-Posterior, Hemisphere and Condition) and follow-up tests are shown in Table 5.1 and Table 5.2 respectively, and effect sizes of significant follow up tests involving condition as a factor are shown in Figure 5.11.

For the imagination phase, ANOVA results revealed significant main effects of condition for the 0-200ms and 200-400ms time-windows (Table 5.1). Follow up paired t-tests compared the conditions collapsed across all electrode sites (since Condition did not interact with Anterior-Posterior or Hemisphere) in these two windows. For the 0-200ms time-window, follow up results indicated that there was no difference between Rehearsed subsequent correct and Imagined subsequent correct conditions. However, Imagined trials where participants subsequently made an intrusion error were associated with a broadly distributed ERP positivity (see Fig. 5.10), that was significantly enhanced compared to both Rehearsed subsequent correct and Imagined subsequent correct conditions (see Table 5.2 and Fig. 5.11). For the 200-400ms time-window, there were also no differences between Rehearsed subsequent correct and Imagined subsequent correct. The ERP positivity for Imagined subsequent intrusions was weaker (see effect sizes in Fig 5.11) but still significantly larger when compared against Imagined subsequent correct, whereas it was only a non-significant trend when compared against Rehearsed subsequent correct (see Table 5.2).

The ANOVA results also showed significant Hemisphere x Condition interactions for the 400-600ms, 600-800ms and 1000-1500ms time-windows (Table 5.1). However, follow-up tests showed no differences between Conditions in either hemisphere (Table 5.2) in either of these time-windows, so these

effects are not interpreted further. There were no significant main effects of Condition or interactions involving the Condition factor in any other time windows.

Table 5.1 .ANOVA results from the omnibus test in imagination phase time-windows.

Omnibus	Time Windows																			
	0 to 200ms		200 to 400ms		400 to 600ms		600 to 800ms		800 to 1000ms		1000 to 1500ms		1500 to 2000ms		2000 to 2500ms		2500 to 3000ms		3000 to 3500ms	
	F	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
AP	47.29	<.001	42.95	<.001	22.15	<.001	11.60	<.001	11.03	<.001	12.83	<.001	10.12	0.00	2.80	0.07	1.12	0.33	2.42	0.10
H	0.13	0.73	0.94	0.34	0.85	0.37	2.66	0.12	0.83	0.37	8.41	0.01	15.56	0.00	11.58	<.001	7.95	0.01	3.22	0.09
C	**6.18**	**<.001**	**4.48**	**0.03**	0.01	0.98	0.14	0.87	0.92	0.40	0.46	0.64	0.02	0.98	0.71	0.50	0.57	0.57	0.59	0.56
AP x H	1.24	0.30	1.05	0.36	2.31	0.10	3.76	0.02	1.56	0.22	0.85	0.44	1.63	0.21	2.31	0.12	1.30	0.28	0.84	0.43
AP x C	1.01	0.43	1.22	0.31	0.76	0.47	0.29	0.73	0.68	0.54	1.38	0.25	1.48	0.22	1.55	0.20	1.69	0.16	1.79	0.15
H x C	2.98	0.07	2.41	0.10	**3.57**	**0.04**	**3.84**	**0.03**	3.06	0.06	**4.82**	**0.01**	1.70	0.19	0.48	0.55	0.31	0.74	0.88	0.42
AP x H x C	0.82	0.52	1.01	0.41	1.23	0.30	1.74	0.14	1.20	0.32	0.86	0.51	1.09	0.37	0.89	0.49	0.73	0.56	0.66	0.62

Note .AP =Anterior-Posterior, H =Hemisphere, C =Condition .Significant results are in bold.

Table 5.2. Results of paired t-tests following up significant omnibus ANOVA effects in the imagination phase time-windows.

	Time Windows																			
	0 to 200ms		200 to 400ms		400 to 600ms		600 to 800ms		800 to 1000ms		1000 to 1500ms		1500 to 2000ms		2000 to 2500ms		2500 to 3000ms		3000 to 3500ms	
H x C	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
left																				
RC vs. IC	-	-	-	-	0.64	0.53	0.85	0.40	-	-	0.50	0.62	-	-	-	-	-	-	-	-
RC vs. II	-	-	-	-	0.48	0.63	0.04	0.97	-	-	0.16	0.88	-	-	-	-	-	-	-	-
IC vs. II	-	-	-	-	0.22	0.83	0.70	0.49	-	-	0.31	0.76	-	-	-	-	-	-	-	-
right																				
RC vs. IC	-	-	-	-	0.93	0.36	1.67	0.11	-	-	1.41	0.17	-	-	-	-	-	-	-	-
RC vs. II	-	-	-	-	0.38	0.71	1.13	0.27	-	-	1.89	0.07	-	-	-	-	-	-	-	-
IC vs. II	-	-	-	-	0.47	0.64	0.26	0.80	-	-	0.65	0.52	-	-	-	-	-	-	-	-
main effects of condition																				
RC vs. IC	0.89	0.38	1.68	0.11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RC vs. II	**2.83**	**0.01**	2.00	0.06	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
IC vs. II	**2.87**	**0.01**	**2.31**	**0.03**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Note. H = Hemisphere, C = Condition, RC = Rehearsed subsequent correct, IC = Imagined subsequent correct, II = Imagined subsequent intrusion. Significant results are shown in bold.

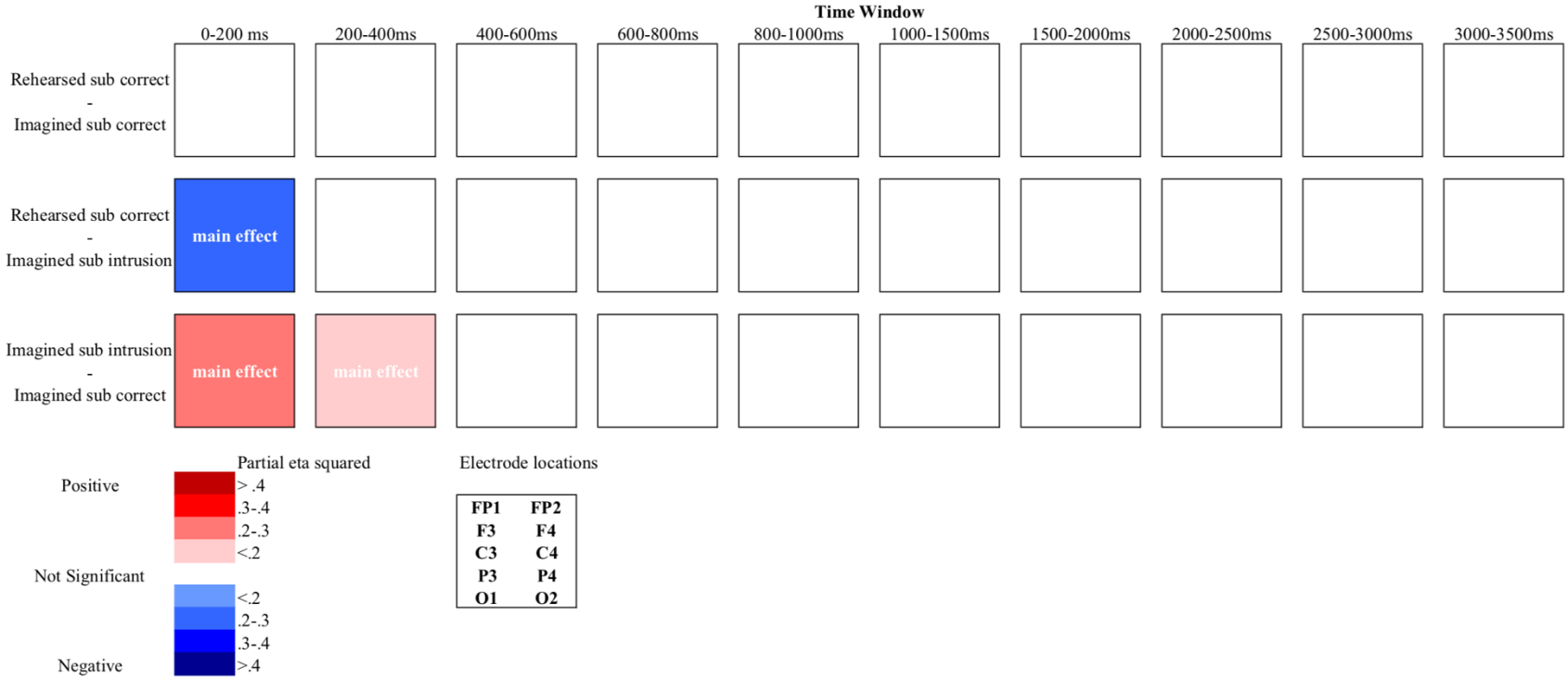

Figure 5.11. Results from the follow-up analyses of significant ANOVA ERP effects in the imagination phase, showing the effect sizes (η_p^2) of pairwise condition differences. Only significant effects involving the factor Condition were followed up. The magnitude and the direction of significant effects across electrodes are illustrated using a colour scale.

Cued recall phase. For the cued recall phase, I used the same three ERP conditions as in the behavioural analysis of cued recall accuracy. That is, separate ERPs were created for object pictures from the Imagined condition (for which participants had previously imagined a counterfactual action instead of the action they performed on day 1), for the Rehearsed condition (for which participants had previously imagined the same action as they performed on day 1) and for the Baseline condition (object pictures that had not been presented before, but depicted objects that participants had performed an action with on day 1), regardless of accuracy. Grand-average ERPs from the 10 electrodes sites for each condition are shown in Figure 5.12 and scalp topographic maps showing the amplitude differences between conditions across 3.5 seconds are shown in Figure 5.13. As can be seen in these figures, in the earlier part of the waveform the Rehearsed and Imagined conditions were associated with increased positive ERPs compared to the Baseline condition, beginning in the 200-400ms time-window, but peaking between around 400-800ms across parietal sites. Later on around 1500-3000ms, the Baseline and Imagined conditions were associated with more positive frontal and central ERPs than the Rehearsal condition.

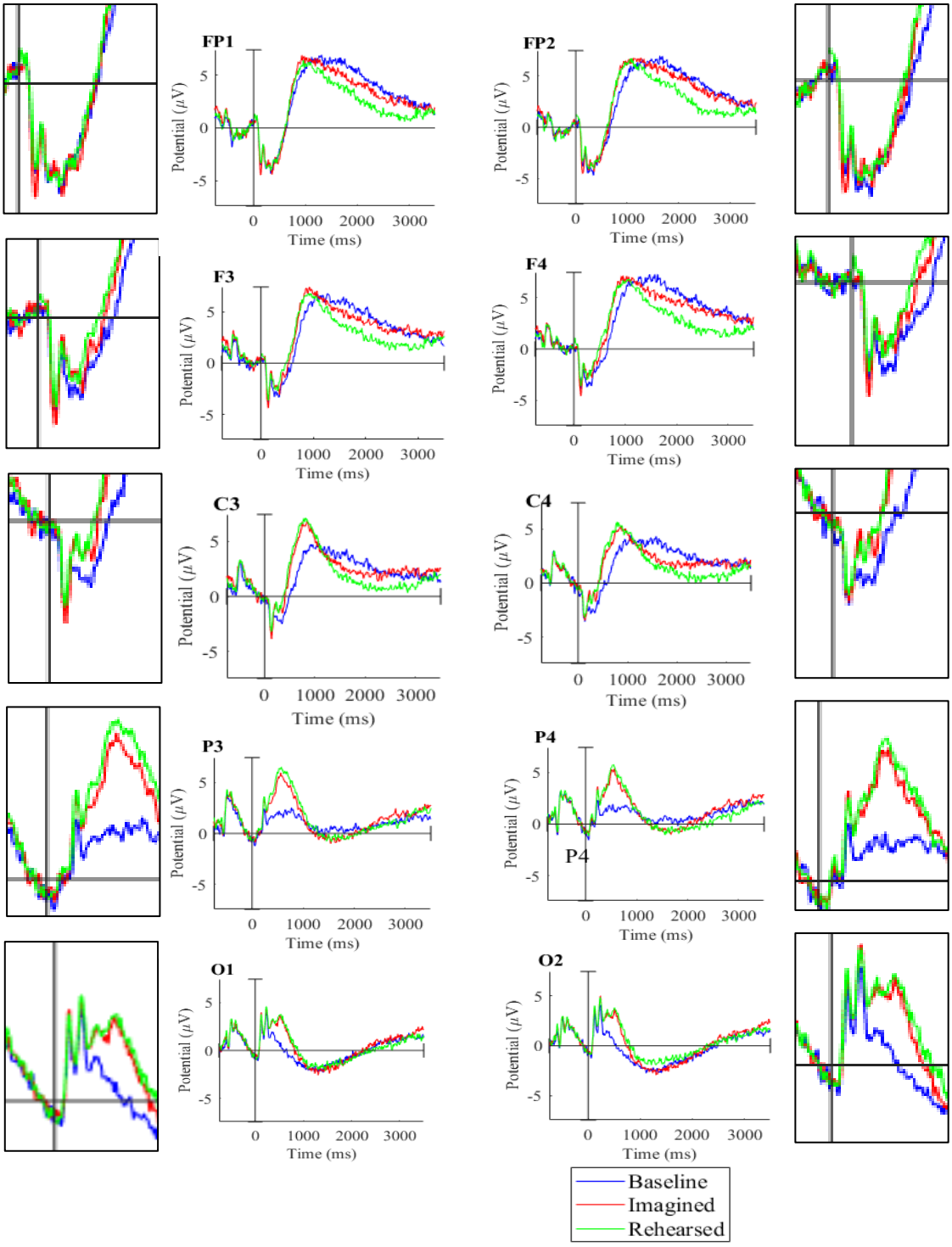

Figure 5.12. Grand-average ERPs from the three conditions in the Cued Recall task. The boxes on the sides show the first part of the epoch magnified to illustrate the early ERP modulations more clearly.

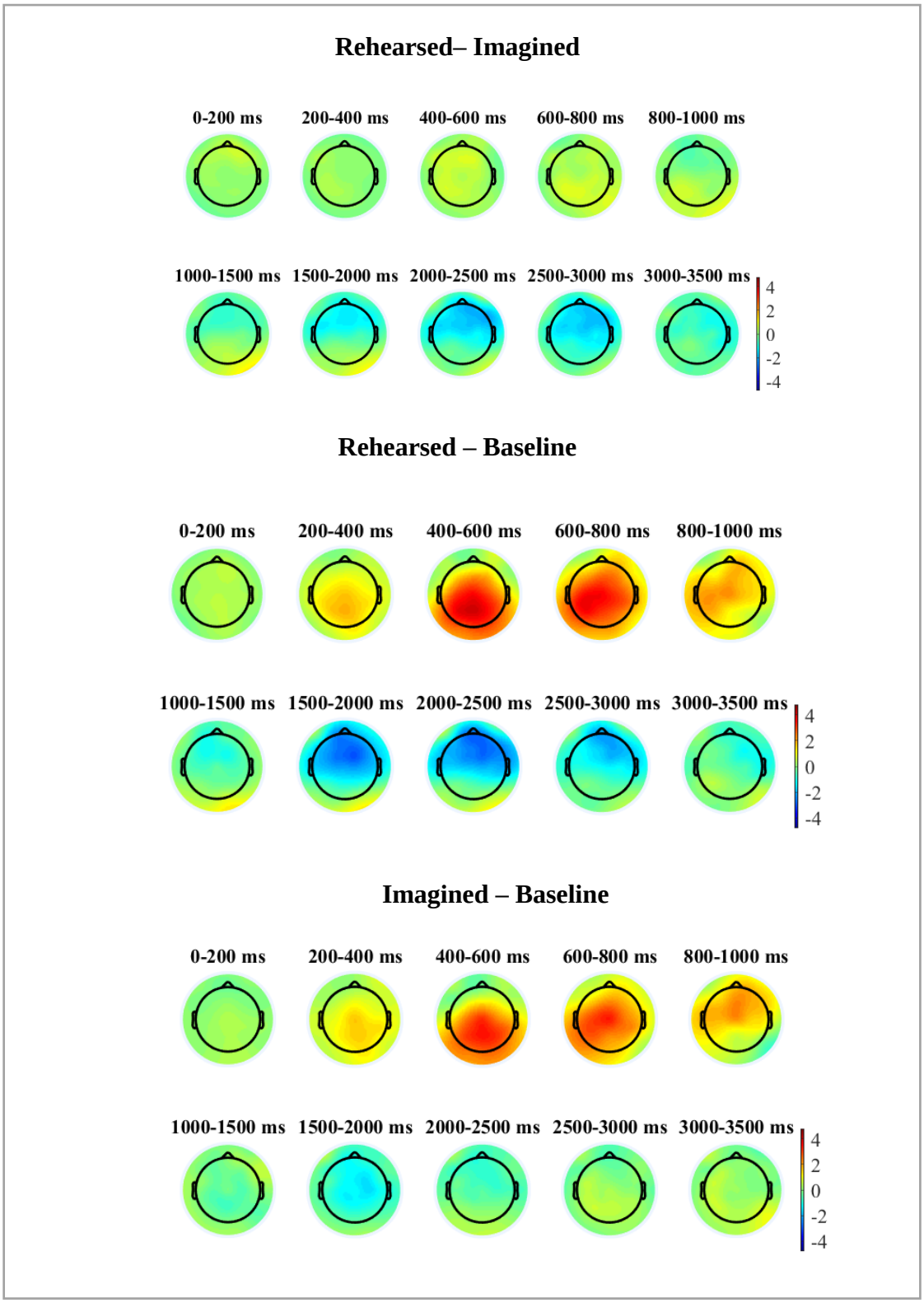

Figure 5.13. Topographic maps showing the scalp distribution of amplitude differences between conditions during the cued recall phase.

Results from the omnibus ANOVAs from the cued recall task (with factors Anterior-Posterior, Hemisphere and Condition) and follow-up tests are shown in Table 5.3 and Table 5.4 respectively, and effect sizes of significant follow up tests involving condition as a factor are shown in Figure 5.14. The earliest significant ANOVA effect was a main effect of Condition in the 200-400ms time-window. Follow up paired t-tests compared the conditions collapsed across all electrode sites (since Condition did not interact with Anterior-Posterior or Hemisphere) showed that the Imagined and Rehearsed conditions both elicited more positive ERPs than the Baseline condition, but there was no significant difference between Rehearsed and Imagined items.

In the 400-600ms time-window, the Condition effect was qualified by an interaction with Anterior-Posterior location. Follow up paired t-tests compared the three conditions separately at each Anterior-Posterior location but collapsed across both hemispheres (since Condition did not interact with Hemisphere). These showed that there were no condition differences at frontopolar sites, but ERPs in the Rehearsed condition were more positive than in the Baseline conditions across all other sites, and ERPs in the Imagined condition were more positive than in the Baseline conditions across central, parietal and occipital sites. This effect peaked at parietal sites (see Figure 5.14 for illustration).

In the 600-800ms and 800-1000ms time-window, there were significant anterior-posterior x hemisphere x condition interactions. Follow up tests comparing conditions at each site separately showed that ERPs in the Rehearsal and Imagination conditions were significantly more positive than ERPs in the Baseline condition across almost all electrodes sites in the 600-800ms, but the differences were strongest at left parietal and left central sites (see Figure 5.14). The Rehearsed

condition was also significantly more positive than the Imagined condition at the P3 site in the 600-800ms time-window. The positive effect for Rehearsed and Imagined items compared to Baseline was a bit weaker but still significant across many sites in the 600-800ms window. In this window, the Rehearsed versus Baseline difference was relatively broad and present across most sites, whereas the positive difference for Imagined versus Baseline was more frontally distributed (see Figure 5.13 and 5.14).

There were no significant effect involving Condition in the 1000-1500ms time-window, but the pattern of ERP effects changed substantially in the later part of the epoch. In the 1500-2000ms, 2000-2500ms and 2500-3000ms time-windows, the ANOVA indicated significant Anterior-Posterior by Condition interactions. Follow up paired t-tests compared the three conditions separately at each Anterior-Posterior location but collapsed across both hemispheres (since Condition did not interact with Hemisphere). These tests showed that in the 1500-2000ms window, the Baseline condition elicited significantly more positive ERPs than Rehearsed and Imagined conditions across frontal central and parietal sites, whereas ERPs in the Imagined condition were also more positive than in the Rehearsed condition across frontal and frontopolar sites (see Figure. 5.14). Across 2000-2500ms and 2500-3000ms windows, both Imagined and Baseline conditions elicited more positive ERPs than the Rehearsed condition across frontopolar, frontal and central sites. There were no significant effects involving Condition in the latest time-window (3000-3500ms).

Thus, the ERPs from the cued recall task showed two different ERP effects. The first effect was an earlier positivity for Rehearsed and Imagined conditions compared to Baseline that peaked across left parietal and central electrodes in the 600-800ms time-window. This effect was similar for Rehearsed and Imagined

conditions, but was slightly larger across the left parietal site between 600-800ms in the Rehearsed condition. The second effect was a late positive sustained effect for Baseline and Imagined conditions compared to the Rehearsed condition that had a frontal maximum and peaked between 1500-3000ms.

Table 5.3. ANOVA results from the omnibus test in cued recall phase time-windows.

	Time Window																			
Omnibus	0 to 200ms		200 to 400ms		400 to 600ms		600 to 800ms		800 to 1000ms		1000 to 1500ms		1500 to 2000ms		2000 to 2500ms		2500 to 3000ms		3000 to 3500ms	
	F	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
AP	22.16	<.001	28.51	<.001	19.76	<.001	7.03	0.01	33.97	<.001	77.07	<.001	66.70	<.001	25.61	<.001	5.32	0.01	1.12	0.33
H	4.15	0.05	0.66	0.42	0.96	0.34	5.10	0.03	4.88	0.04	0.05	0.83	0.18	0.68	0.28	0.60	0.40	0.54	0.16	0.69
C	1.74	0.19	**8.35**	**<.001**	**32.08**	**<.001**	**23.68**	**<.001**	**6.16**	**<.001**	0.56	0.58	**12.06**	**<.001**	**8.71**	**0.001**	**6.19**	**0.004**	1.58	0.22
AP x H	1.50	0.23	0.17	0.89	0.38	0.73	1.05	0.37	0.70	0.51	1.04	0.37	1.39	0.26	1.71	0.15	0.86	0.49	0.50	0.74
AP x C	1.25	0.30	2.43	0.08	**13.30**	**<.001**	**3.83**	**0.01**	1.47	0.23	2.57	0.06	**5.78**	**0.002**	**7.21**	**<.001**	**3.68**	**0.02**	0.97	0.41
H x C	0.68	0.51	1.01	0.37	0.32	0.73	0.52	0.60	0.03	0.97	0.72	0.49	0.11	0.89	0.28	0.73	0.77	0.47	1.56	0.22
AP x H x C	0.56	0.71	0.86	0.49	1.39	0.25	**4.60**	**0.001**	**4.22**	**0.003**	1.31	0.28	1.46	0.21	1.26	0.29	1.22	0.31	1.64	0.17

Note .AP =Anterior-Posterior, H =Hemisphere, C =Condition .Significant results are in bold.

Table 5.4. Results of paired t-tests following up significant omnibus ANOVA effects in the cued recall phase time-windows

	Time Window																			
	0 to 200ms		200 to 400ms		400 to 600ms		600 to 800ms		800 to 1000ms		1000 to 1500ms		1500 to 2000ms		2000 to 2500ms		2500 to 3000ms		3000 to 3500ms	
AP x H x C	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
FP1																				
R vs. I	-	-	-	-	-	-	0.03	0.98	0.76	0.46	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	1.45	0.16	0.94	0.36	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	1.64	0.12	**2.52**	**0.02**	-	-	-	-	-	-	-	-	-	-
FP2																				
R vs. I	-	-	-	-	-	-	0.62	0.54	0.21	0.84	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**3.74**	**0.001**	**2.34**	**0.03**	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**2.76**	**0.01**	**3.31**	**0.003**	-	-	-	-	-	-	-	-	-	-
F3																				
R vs. I	-	-	-	-	-	-	0.65	0.52	0.60	0.56	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**4.16**	**<.001**	1.58	0.13	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**3.79**	**0.001**	**3.15**	**0.004**	-	-	-	-	-	-	-	-	-	-
F4																				
R vs. I	-	-	-	-	-	-	1.50	0.15	0.49	0.63	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**5.17**	**<.001**	**2.61**	**0.02**	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**3.26**	**0.003**	**3.23**	**0.004**	-	-	-	-	-	-	-	-	-	-
C3																				
R vs. I	-	-	-	-	-	-	1.56	0.13	0.54	0.60	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**5.93**	**<.001**	**2.86**	**0.01**	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**4.85**	**<.001**	**3.37**	**0.003**	-	-	-	-	-	-	-	-	-	-
C4																				
R vs. I	-	-	-	-	-	-	1.67	0.11	0.23	0.82	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**5.60**	**<.001**	**2.96**	**0.01**	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**4.18**	**<.001**	**3.13**	**0.01**	-	-	-	-	-	-	-	-	-	-
P3																				
R vs. I	-	-	-	-	-	-	**2.78**	**0.01**	1.68	0.11	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**6.49**	**<.001**	**2.81**	**0.01**	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**4.72**	**<.001**	**2.07**	**0.05**	-	-	-	-	-	-	-	-	-	-
P4																				
R vs. I	-	-	-	-	-	-	1.47	0.16	0.16	0.87	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**5.03**	**<.001**	1.59	0.13	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**3.34**	**0.003**	1.44	0.16	-	-	-	-	-	-	-	-	-	-
O1																				
R vs. I	-	-	-	-	-	-	1.18	0.25	1.17	0.26	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**5.56**	**<.001**	**2.75**	**0.01**	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**4.57**	**<.001**	1.76	0.09	-	-	-	-	-	-	-	-	-	-
O2																				
R vs. I	-	-	-	-	-	-	1.95	0.06	1.70	0.10	-	-	-	-	-	-	-	-	-	-
R vs. B	-	-	-	-	-	-	**4.96**	**<.001**	**2.60**	**0.02**	-	-	-	-	-	-	-	-	-	-
I vs. B	-	-	-	-	-	-	**3.19**	**0.004**	0.97	0.34	-	-	-	-	-	-	-	-	-	-

Table 5.4. Results of paired t-tests following up significant omnibus ANOVA effects in the cued recall phase time-windows (Continued).

	Time Window																			
	0 to 200ms		200 to 400ms		400 to 600ms		600 to 800ms		800 to 1000ms		1000 to 1500ms		1500 to 2000ms		2000 to 2500ms		2500 to 3000ms		3000 to 3500ms	
AP x C	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
FP																				
R vs. I	-	-	-	-	0.72	0.48	-	-	-	-	-	-	**2.71**	**0.01**	**4.00**	**0.001**	**3.38**	**0.003**	-	-
R vs. B	-	-	-	-	1.01	0.32	-	-	-	-	-	-	**3.33**	**0.003**	**3.53**	**0.002**	**3.12**	**0.01**	-	-
I vs. B	-	-	-	-	-0.28	0.78	-	-	-	-	-	-	1.80	0.09	1.20	0.24	0.79	0.44	-	-
F																				
R vs. I	-	-	-	-	1.73	0.10	-	-	-	-	-	-	3.32	0.003	**3.96**	**0.001**	**4.22**	**<.001**	-	-
R vs. B	-	-	-	-	**6.02**	**<.001**	-	-	-	-	-	-	**4.80**	**<.001**	**5.19**	**<.001**	**3.44**	**0.002**	-	-
I vs. B	-	-	-	-	1.89	0.07	-	-	-	-	-	-	**2.66**	**0.01**	1.43	0.17	0.03	0.98	-	-
C																				
R vs. I	-	-	-	-	**1.90**	**0.07**	-	-	-	-	-	-	**2.04**	**0.05**	**3.06**	**0.01**	**3.52**	**0.002**	-	-
R vs. B	-	-	-	-	**7.86**	**<.001**	-	-	-	-	-	-	**4.91**	**<.001**	**4.78**	**<.001**	**2.73**	**0.01**	-	-
I vs. B	-	-	-	-	**5.27**	**<.001**	-	-	-	-	-	-	**3.87**	**0.001**	1.53	0.14	0.53	0.61	-	-
P																				
R vs. I	-	-	-	-	2.04	0.05	-	-	-	-	-	-	0.16	0.88	1.46	0.16	**2.29**	**0.03**	-	-
R vs. B	-	-	-	-	**8.58**	**<.001**	-	-	-	-	-	-	**2.20**	**0.04**	1.51	0.15	0.79	0.44	-	-
I vs. B	-	-	-	-	**7.21**	**<.001**	-	-	-	-	-	-	**2.14**	**0.04**	0.09	0.93	1.33	0.20	-	-
O																				
R vs. I	-	-	-	-	1.45	0.16	-	-	-	-	-	-	1.24	0.23	0.40	0.69	1.16	0.26	-	-
R vs. B	-	-	-	-	**7.36**	**<.001**	-	-	-	-	-	-	0.37	0.72	0.36	0.73	0.55	0.59	-	-
I vs. B	-	-	-	-	**6.25**	**<.001**	-	-	-	-	-	-	0.71	0.49	0.70	0.49	0.62	0.54	-	-
main effect	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
R vs. I	1.11	0.28	0.46	0.65	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R vs. B	1.92	0.07	**3.59**	**0.002**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
I vs. B	0.60	0.55	**3.94**	**0.001**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Note. H = Hemisphere, C = Condition, R = Rehearsed, I = Imagined, B = Baseline. Significant results are in bold.

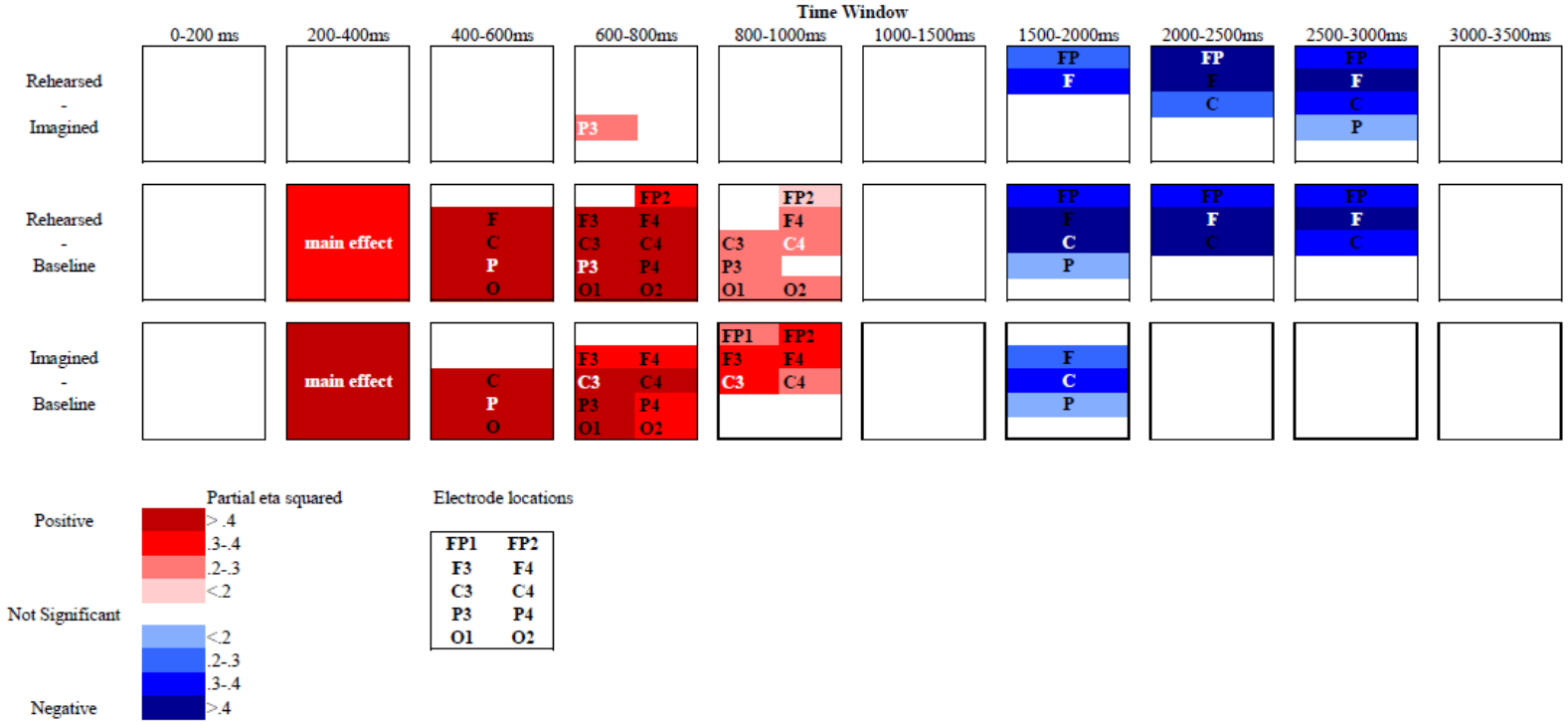

Figure 5.14. Results from the follow-up analyses of significant ANOVA ERP effects in the cued recall phase, showing the effect sizes (η_p^2) of pairwise condition differences. Only significant effects involving the factor Condition were followed up. The magnitude and the direction of significant effects across electrodes are illustrated using a colour scale. The electrode position where the effects had maximal effect size is written in white type font for three-way interactions. Similarly, the region with maximal effect size is written in white type fonts for two-way interactions. Abbreviations: FP1 = frontopolar, F = frontal, C = central, P = parietal, O = occipital.

Discussion

The aim of this study was to investigate the cognitive and neural mechanisms underlying the effect of counterfactual imagination on true memory for an event. As predicted, participants were more accurate at recalling and recognising true actions after they had been rehearsed, but were worse at recalling and recognising true actions after a counterfactual action had been imagined repeatedly. During the imagination phase, there was an early ERP positivity that predicted that participants would subsequently misremember the imagined false action as true. Such subsequent false memories were also predicted by more vivid imagination ratings compared to counterfactually imagined actions that were not subsequently falsely remembered as true. During the subsequent cued recall test, early ERP effects were sensitive to whether the image cue had been previously seen or not, whereas later ERP effects tracked accuracy and confidence at recalling the correct action. Thus, the results provide several lines of evidence on the type of neural and cognitive mechanisms by which counterfactual imagination can cause true memory impairments, and on the consequences of such impairments on subsequent retrieval processes.

The results on the cued recall test suggest that repeatedly imagining a true action associated with an object can strengthen memory for the correct action that participants performed on the first day, while repeatedly imagining performing a counterfactual action with the object causes intrusion errors on the cued recall test due to the false action memory being strengthened by imagination. Consistent with this view, counterfactual imagination was rated significantly more vivid when it was associated with later intrusion errors than when intrusion errors were successfully avoided.

Importantly, counterfactual imagination also produced lower performance on the cued recall test compared to a baseline condition that had not been cued in the imagination phase, confirming a memory impairment beyond simple forgetting over time (Anderson, 2003; Anderson & Green, 2001). In addition, participants rated their confidence on the cued recall task lower for counterfactually imagined and baseline conditions compared to the rehearsed condition. This indicates that participants experienced conscious uncertainty of their answer for the imagined condition, which may have been caused by conflict between the memories for the performed and imagined actions. In contrast, low confidence for the baseline condition may have been produced by a weaker memory of the performed action due to a lack of rehearsal. Interestingly however, confidence did not differ between imagination and baseline conditions, despite less accurate performance in the imagination condition. This latter pattern suggests that at least some of the false memories leading to intrusion errors were associated with a subjective experience of those memories as true.

For the recognition test, recognition of objects together with the original true action was most accurate for the rehearsal condition compared to the imagined and baseline conditions, and recognition of such "old" actions in the baseline condition was also more accurate than in the imagined condition, again showing a below baseline impairment as a result of counterfactual imagination. For objects paired with completely new actions, participants were more accurate in the rehearsal condition than the imagined and baseline conditions, but there was no difference between imagined and baseline conditions. Theoretically, I proposed that repeatedly imagining a new action could reduce participant's ability to remember the original memory via either

interference or inhibition (Anderson, 2003; Anderson & Neely, 1996). The results from the recognition test are potentially consistent with an inhibitory account, because despite providing a very strong and direct cue to the original action (the object with the action sentence from day 1), participants were still impaired at recognising this action as true. If the effects of counterfactual imagination on memory were only due to interference/blocking, I would have expected a below baseline impairment only on the cued recall task and no impairment on the recognition task. However, the results are not conclusive in ruling out interference/blocking as the underlying mechanism, because the recognition test also included the shared object cue (this was necessary for enabling participants to complete the task, as many of the actions would have been too unspecific without an object). Thus, it is possible that the object cue elicited recall of the counterfactually imagined action, and that participants sometimes used this recall to mistakenly reject the true action (a "recall to reject" strategy, Rotello & Heit, 2000; Verde & Rotello, 2004). Thus, the results do not show clear support for either inhibition or interference, but are potentially consistent with both theories.

Considering the ERP results next, I observed a very early positive ERP effect during the imagination phase that predicted later intrusion errors whereby counterfactual actions were mistakenly recalled as true. The same object-actions pairs were also associated with higher subjective vividness ratings during counterfactual imagination than other object-action pairs for which participants did not make later intrusion errors, but rather managed to correctly recall the original true actions. This effects thus has some similarities to the findings by Gonsalves and Paller (2000) who described an ERP positivity related to visual imagination that predicted later false memory that an

imagined object had been seen as a picture, when in fact it had only been cued by an object word. However, Gonsalves and Paller's (2000) ERP effect occurred 600-900ms after the object word was presented, and my effect was significant in the 0-200 and 200-400ms time-windows. This early positivity was not manifest as a modulation of a specific ERP peak (e.g. P2 or N2) but rather as a general positive shift of the waveform in the earliest part of the post-stimulus epoch. Thus, the ERP effect in my data seems too early to be related to imagination of the action, since it was present in the time-windows when participants would likely have still been processing the cue (as visually recognising the object and reading the sentence takes some time).

Instead, the early ERP positivity could be related to some process that enhanced encoding of the counterfactual action. For example, if participants' attention to stimuli varied across trials (as might be expected in a long and repetitive task, where participants may also experience mind wandering/fatigue), enhanced attention to some stimuli could result in stronger encoding of those actions, and increasing the likelihood that those actions would be recalled later and "intrude" on the cued recall test. ERP correlates of processes that enhance encoding can be manifest very early after stimulus onset (reviewed in Paller & Wagner, 2002), or even before a stimulus is presented (Otten et al., 2006). However, this explanation is highly tentative, and it is difficult to conclude with certainty what processes might be associated with this ERP effect. Future research should attempt to replicate this effect during counterfactual imagination, and test whether it is sensitive to attentional manipulations vs. the vividness of imagination, for example.

ERPs from the cued recall phase showed what appeared to be two different effects. In the first part of the epoch, ERPs for the rehearsed and imagined conditions were associated with enhanced positive ERPs compared to the baseline condition, starting in the 200-400ms window where the effect was broadly distributed, and peaking across left central and parietal sites between around 400-800ms after stimulus onset. This effect thus highly resembles typical old>new ERP effects that are thought to index familiarity and recollection during successful recognition (reviewed in Rugg & Curran, 2007), indicating that participants were recognizing the pictures of object that they had seen before, whereas they had not seen the baseline pictures before in the same session (and had only seen the actual objects on day 1). Although this positivity was highly similar for rehearsed and imagined conditions, interestingly it was slightly larger in the rehearsed condition at the left parietal site between 600-800ms, which is where and when recollection-related ERP activity typically peaks (Rugg & Curran, 2007). Thus, consistent with Gonsalves and Paller's (2000) findings, there was some evidence for less successful retrieval in the imagined condition than the rehearsal condition.

After the initial old>new ERP positivities, I also observed a late frontal positivity that showed a different pattern between conditions, since it was more positive in the imagined and baseline conditions than the rehearsed condition. This frontal positivity could reflect executive control-related activity during retrieval, such as retrieval monitoring or retrieval effort, which is required in situations where retrieval is not immediately successful or participants have to evaluate retrieved information in relation to the task goals (e.g. Hayama et al., 2008; reviewed in Voss & Paller, 2017). In my study, for the imagined condition, participants would have likely needed to evaluate

their retrieved memories to assess if they were of the original action that they performed on the first day or of the new action that they have only imagined during the imagination phase. For the baseline condition, participants may have needed to put more effort in to remember what action they performed on the first day, because they did not have an opportunity to rehearse any action during the imagination phase. In both these situations, it was difficult for participants to decide on the correct answer, as shown in participants' behavioural performance which was lower in these two conditions than in the rehearsal condition. Furthermore, the confidence ratings results for imagined and baseline conditions indicate that participants were aware of their low accuracy, since they rated their confidence much lower than in the rehearsed condition, where there was also no right frontal ERP positivity.

The results of this study though show some interesting behavioural and neural effects that begin to explain how counterfactual imagination affects true memories. However, these findings should be interpreted with caution until they have been replicated, and possible limitations of the study should also be considered together with future directions. First, it should be noted that in this study I did not record EEG during the learning phase. Therefore, I cannot draw a conclusion on what is happening in the brain during initial encoding, which might be relevant in terms of predicting which true actions will later be subject to false memories of the counterfactual actions (for example, is it only poorly encoded actions from day 1 that are subject to being supplanted with the counterfactual actions?). Second, because the cued recall phase was always conducted before the recognition test, cued recall responses might have biased recognition test results. To measure participants' recognition independently of cued recall, future

research should vary the type of test between participants rather than using both tests for the same participants. Finally, in this study, I did not have enough trials to create separate ERPs for high/low confidence trials or high/low vividness trials to be able to compare which aspects of the ERP activity was related to confident vs. less confident retrieval or vivid vs. less vivid imagination, which would have helped interpret the effects. It has been suggested that ERP researchers should aim for at least 30 trials per condition in order to compare ERPs between experimental conditions (Wilding & Ranganath, 2011). However, in this study it was not possible to achieve that many trials, as the study would be too long and participants would experience fatigue. In addition, future studies should measure or manipulate participant's attention during the imagination phase, since the amount of attention participants paid to the stimulus may affect encoding of the counterfactual imagination and subsequent retrieval.

In conclusion, the current study demonstrates that imagination of counterfactual actions can impair participants' ability to remember what truly happened, even for sensorimotor rich true memories involving interacting with real objects. During counterfactual imagination, the ERPs showed evidence of an early brain process that predicted subsequent misremembering of the imagined false action instead of the true action, and during subsequent retrieval, ERP markers of executive control involvement were elicited when retrieval was less accurate and less confident, indicating that prior counterfactual imagination affected retrieval processing by making it more effortful and/or requiring additional monitoring. Thus, the current is starting to shed light on the underlying cognitive and brain mechanisms by which counterfactual imagination affects true memory.

Chapter 6: General Discussion

As argued throughout this thesis, our memory is vulnerable to distortion. Engaging in simple cognitive processes like imagining an alternative version of a past event can distort the original memory of the event (Mitchell & Johnson, 2009). Although memory modification may be beneficial in certain circumstances (e.g. to those who suffer from childhood trauma or those who had negative experiences in the past), it could be a serious drawback in forensic settings. Real criminals may imagine a false alibi as a countermeasure strategy to help them evade and hide their guilt. In the past, "lie detection" protocols were used with polygraphs as a technique to detect whether or not the suspects were telling the truth by measuring their physiological reactions when supposedly lying, however researchers have raised serious concerns regarding the validity of trying to detect lies (e.g. Ben-Shakhar, 1991, 2002; Iacono & Lykken, 2002). Recently, an alternative indirect method used for detecting guilt, concealed memory detection, was introduced and showed promising results, for example with the aIAT (Sartori et al., 2008) and the CIT (Rosenfeld et al., 2008). However, these methods require further research, for example to evaluate whether they are susceptible to countermeasures and also whether they should be assumed to work in real life, and not just the laboratory. To my knowledge, there are only a few studies (e.g. Gronau et al., 2015; Shidlovski et al., 2014) that have investigated the impact of rehearsing alternative versions of past events on either aIAT and CIT memory detection, as addressed in this thesis. Here, I present evidence from both behavioural and ERP measures on the issue of how true memories change as a consequence of imagining a counterfactual version of a

past event, using ecologically valid methods that are relevant to real-life, such as forensic settings.

6.1. Summary of Empirical Findings

Experiment 1 and 2 investigated the consequences of rehearsing and imagining a false alibi after committing a mock crime, and whether this can be used as a countermeasure strategy to help participants evade subsequent memory detection using aIAT. Experiment 1 suggested that aIAT memory detection was substantially impaired when contrasting the mock crime to the alibi event directly, as a result of imagining a false alibi. The aIAT was not able to distinguish whether the mock crime or a false alibi was true. However, it was not possible in this design to determine *why* the mock crime and false alibi were indistinguishable, that is, whether this occurred because the detection of truth of the mock crime *decreased*, or the detection of truth of the false alibi *increased* as a result of the manipulation. Experiment 2 therefore contrasted the mock crime with an unexperienced event that was independent from the false alibi, to examine if the same pattern would be found regardless of which event was contrasted with the mock crime in the aIAT. Results from Experiment 2 revealed that the mock crime was detectable on average, even though participants had rehearsed an alibi before the test. This evidence therefore suggested that imagining a false alibi did not impaired the original memory of the mock crime. However, there was a subtle reduction of truth detection of the mock crime memory. Thus, these experiments indicated that the low discrimination between the mock crime and imagined alibi in Experiment 1 was primarily due to the alibi manipulation increasing the detected truth value of the alibi

scenario, with only a subtle non-significant reduction in detection of the mock crime as true.

Extending from Experiment 1 and 2, Experiment 3 adopted the same procedure with an additional 1-week delay in between the lab act (mock crime or innocent act) and the aIAT test. This experiment aimed to replicate and extend on findings from the previous experiments. It focused on strengthening the alibi manipulation to be more realistic, since in real life suspect may prepare and imagine a false alibi repeatedly over a period of time prior to the investigation. In this study, both aIAT versions from the previous two experiments were used, together with an additional aIAT version where the alibi was contrasted with an unexperienced event. The results showed that the alibi manipulation was most effective at making guilty participants appear innocent when participants rehearsed the false alibi just once prior to the test, not repeatedly over a week. Furthermore, the results also suggested that aIAT versions that contrast the true event with a novel, unexperienced event that suspects have no prior knowledge about is the strongest and most optimal aIAT method for detecting guilt. To summarise, converging evidence from Experiments 1-3 showed that the aIAT is very vulnerable to countermeasures involving imagining an alternative scenario of the event prior to the test and this could be a serious problem for real life applications.

Since the previous experiments in this thesis indicated that the aIAT is very susceptible to a false alibi countermeasure, this raised concerns regarding what exactly the aIAT is measuring if it cannot distinguish an objectively true event from a false event that the suspect knows is false. Therefore, a more direct method for detecting existing memories was used in the next study. Experiment 4 used a P300-based CIT to

investigate the extent to which rehearsing a false alibi can modify crime-related memories as assessed with more direct, neural markers of memory. This experiment showed that the alibi manipulation was not effective at modulating memory detection with the P300 CIT, although the detection results I obtained in the standard guilty groups were relatively poor compared to the literature, suggesting that my implementation of the CIT may have been subject to some weaknesses. My results did however converge with findings in the literature, in that the peak-to-peak P300-LPN ERPs measure seems to be a better measure compared to P300 or LPN in isolation for detecting guilt, as suggested by Rosenfeld and colleagues (e.g. 2008; 2013). Furthermore, it was also found that time-delay between the mock crime and the test did not have an effect on accuracy at detecting guilt with the CIT, as previously (Hu et al., 2015; Hu & Rosenfeld, 2012; Rosenfeld et al., 2013). Experiment 4 thus suggested that P300-based CIT may be robust against the false alibi countermeasure, contrary to the aIAT. Across Experiment 1-4, consistent findings were obtained in that there was only little evidence that the alibi manipulation directly impaired the mock crime memory.

Experiment 5 was then conducted to investigate possible cognitive and neural mechanisms underlying memory distortion via counterfactual imagination. Participants were asked to either repeatedly imagine an action performed on the first day or imagine performing another, counterfactual action. Results showed that repeatedly imagining a performed action helped participants later recall that true action better than when they repeatedly imagined another action, since the latter condition produced intrusion errors where the imagined action was mistakenly reported as true. In addition, counterfactually imagined actions that would later be falsely remembered as true were associated with

higher vividness of imagination. Furthermore, cued recall and recognition of performed actions was lower after counterfactual imagination than in a baseline condition that measured simple forgetting over time, thus confirming a memory *impairment* as a result of counterfactual imagination. ERP results revealed two key findings. First, there was an early ERP positivity during counterfactual imagination that predicted intrusion errors that may index encoding-related processes. Second, ERPs recorded during the final recall test showed that the two conditions that were associated with lowest recall performance (baseline and counterfactual imagination conditions) elicited similar late frontal ERP positivities that I hypothesized may reflect the recruitment of executive control processes when recall is particularly difficult. Thus, the final experiment revealed some novel evidence regarding the neurocognitive mechanisms that produce memory distortions after counterfactual imagination, and the processes that are involved during subsequent retrieval when people need to overcome counterfactual imagination effects on memory to recall the true version of what actually happened.

6.2. Theoretical Implications

Overall, the key findings reported in this thesis suggested that imagining counterfactual alternatives of past events has a strong effect on memory, such that it can lead to memory distortions and errors, at least in some circumstances. However, it is less clear on the basis of this new evidence whether the effects of counterfactual imagination on memory are driven by interference or inhibition (see Anderson et al., 1994; Camp et al., 2007; Hellerstedt & Johansson, 2013), or potentially both mechanisms. Results from my Experiments 1-4 indicated that the alibi manipulation was primarily creating

memories for information in the alibi event that had some shared characteristics with true memories, such that repeatedly rehearsing an imagined alibi strengthened the detection of the alibi scenario as true rather than decreased detection of the mock crime. In Experiment 1-3, although the mock crime was detected for those who did not adopt a countermeasure (in line with Sartori et al, 2008; Agosta & Sartori, 2013), results showed that the false alibi countermeasure significantly reduced memory detection. In addition, the results also indicated that a false alibi can be detected as a true memory when in fact it is not. In line with the literature, imagining a false alibi may have created memories that had implicit associations with the truth, despite the fact that participants knew explicitly that the alibi was false (Shidlovski et al., 2014; Takarangi et al., 2015, 2013). This could have emerged because vivid imagination can lead to encoding of various perceptual features in addition to semantic information, which can cause imagined memories to become similar to memories of perceived events in terms of their features, as shown in the reality monitoring literature (Mitchell & Johnson, 2009). Perhaps this similarity to a true memory was sufficient for the imagined alibi to be detected as true by the aIAT, despite participants themselves being able to distinguish it from a true memory.

Results from Experiment 4, however, provided a novel finding that a false alibi did not impair retrieval of the original memory when cued more directly (i.e. the cue being an object word for a stolen object) and retrieval was measured in terms of EEG brain activity associated with recognition in a P300-based CIT. This result thus suggested that there was no reduction in strength of the true mock crime memory, which was expected if rehearsing a false alibi had led to inhibition of the true memory in line

with typical findings in the Retrieval-Induced Forgetting literature (Anderson et al., 1994 for review see Anderson, 2003). According to this literature and a prior study that investigated CIT memory detection with polygraph measures (Gronau et al., 2015), selective rehearsal of an alibi could have inhibited the true memory if it competed during retrieval practice of the alibi, which should have reduced the P300 response for crime reminders (see also Bergstrom et al., 2013; Hu et al., 2015, for related evidence that P300 responses can be suppressed). However, according to my findings, the mock crime memory was not impaired, suggesting that inhibition mechanisms are not involved during alibi rehearsal, at least in the way the false alibi manipulation was implemented here. However of course, just because I did not find evidence of inhibition in my studies does not mean inhibition is not recruited during counterfactual imagination in real life, where suspects may rehearse false alibis much more extensively and with more motivation to succeed than in a staged lab study.

Nevertheless, results from Experiment 5 did produce some evidence that was consistent with the inhibition account, but the evidence was also consistent with an interference/blocking account. On the cued recall test, participants were highly likely to recall the counterfactually imagined actions as true. Thus on this test, memories for imagined actions may have come to mind more easily than the memories for true actions from day 1, which could have blocked access to the true memories without those memories being inhibited (Camp et al., 2007). So, the memory impairment on the cued recall test could be explained by a combination of interference and reality monitoring errors, since in order to falsely believe the imagined actions were truly performed, participants would have had fail to detect their source as being imagination (Mitchell &

Johnson, 2009). In this experiment, it was however also found that participants not only failed to recall performed actions, but they were also impaired at recognizing the true action as a consequence of counterfactual imagination. This pattern is potentially more consistent with inhibition of true memories since the recognition test did provide a very strong and direct cue to the true memory, and this test should therefore be less susceptible to interference/blocking than the cued recall test (e.g. Hicks & Starns, 2004).

However, it is difficult to conclude with certainty that the recognition impairment was due to inhibition, because I was not able to design the recognition test without including the shared object cue that was also associated with the counterfactually imagined action. That is, many of the action sentences would have been ambiguous without an object and there were many actions that were quite similar across the different objects, so I needed to present participants with the object-action pairing from day 1 in order for them to make an associative recognition judgement. However, previous research has shown that participants can sometimes solve associative recognition tasks with a "recall to reject" strategy (Rotello & Heit, 2000; Verde & Rotello, 2004). In my task, participants may have used the object cue to recall the action they had just produced on the cued recall task, which may have actually brought to mind the counterfactual action instead of the true action. As a result, participants may have mistakenly rejected the true action as "new". Taking into account these alternative suggestions, it is inconclusive whether counterfactual imagination effects on true memories was driven by interference or inhibition, or both. Nevertheless, despite the theoretical uncertainty, I did find in this experiment that counterfactual imagination can produce strong distorting effects on memories of interacting with real objects,

suggesting that counterfactual imagination is an important source of memory distortion in our everyday life.

6.3. Practical Implications

A key practical implication of Experiments 1-3 in this thesis is that the aIAT is not a valid method for detecting the objective truthfulness of autobiographical events, contrary to what has been claimed (Agosta et al., 2013; Marini et al., 2012; Sartori et al., 2008). My results thus converge with a growing body of evidence showing that aIAT results cannot be trusted (e.g. Hu et al., 2015; Hu & Rosenfeld, 2012; Shidlovski et al., 2014; Takarangi et al., 2015, 2013; Verschuere et al., 2009). As discussed earlier, it is possible that the aIAT detection failed because counterfactual imagination created a memory for the imagined event that had some implicit association with the truth. Alternatively however, the aIAT may not even measure the association between the event and the truth, but could be sensitive to more trivial factors such as the relative salience of the two contrasted events, so that the detected strength of the association between the truth and an autobiographic event depends upon which of the two events is more obvious or accessible to the suspect (see Shidlovski et al., 2014). In my studies, the alibi and mock crime might have been equally highly accessible to the participants, and therefore, the aIAT was not able to determine which of the event was true. This uncertainly regarding what the aIAT is measuring means that the application of this method in real life should be done with great caution, and is especially concerning since there are already reports that the aIAT has been used in real court cases in Italy (Sirgiovanni et al., 2016). On the other hand, P300-based CIT seems to be a better

measure to use to detect guilty memories, as it was found to be more resistant to the countermeasure in my Experiment 4. It appears that the neural markers of memory as measured with P300 responses are robust against the counterfactual imagination countermeasure and time-delay. Nevertheless, since other research has found that the P300 CIT is vulnerable to countermeasures such as retrieval suppression (Bergstrom et al., 2013; Hu et al., 2015), it is still doubtful whether this test can produce sufficiently valid results for real life applications.

In addition to the above issues with countermeasures, there are several other methodological limitations that should be taken into account when evaluating the validity of forensic memory detection. Laboratory studies have not fully examined how emotional factors, such as anxiety and stress, or motivation to conceal the truth, might affect the aIAT or CIT test outcome. The actual crime may elicit stronger emotional arousal than mock crimes in laboratory settings, and emotional arousal is known to enhance the subjective vividness of memories and their durability over time (Kensinger, 2009). Thus, true memories of crimes may be qualitative different from the types of memories that are usually studied in the lab, which means they may or may not be as susceptible to distortion as memories of mock crimes.

The results of my Experiment 5 also has some practical implications in terms of aiding our understanding of how memory distortions may be manifest in real life. The majority of research on interference and inhibition mechanisms in memory has used rather artificial stimuli such as word pairs or semantic categories and exemplars (e.g. Anderson, 2003; Anderson et al., 1994; Anderson & Green, 2001; Levy & Anderson, 2008; Storm & Levy, 2012), I investigated whether ecologically valid memories of

interacting with real objects would also be susceptible to memory distortions, and found strong effects of counterfactual imagination on these action memories. These findings thus suggest that counterfactual imagination could be a source of memory distortions also for real life memories of potentially important actions such as whether one has remembered to switch off a stove, or taking ones medication. My findings thus add to the literature that even sensorimotor rich, ecologically valid memories can become distorted (see Hu et al., 2015), despite these types of memories often being stronger and better remembered than more arbitrary stimuli such as word pairs (Engelkamp & Zimmer, 1989). This finding emphasizes the importance of this research, in that phenomena that generalise to real life situations are very important for psychologists to try to understand. It will be practically important to understand whether similar distortions are observed as a result of counterfactual imagination when that is done in other situations, such as by eyewitnesses of a crime, or during therapy, for example.

6.4. Limitations and future directions

There were a number of limitations with the current research that should be addressed in future studies. For example, although my Experiments 1-4 did not find evidence that the false alibi rehearsal inhibited the true memory, it is possible that my alibi manipulation was rather weak and that adapting the alibi manipulation to make it more potent could result in inhibition. Specifically, future research should consider explicitly training participants to suppress thoughts of the mock crime while completing the alibi imagination task, which might be an effective strategy for reducing mock crime memory strength whilst simultaneously strengthening memory for the alibi (cf.

Anderson & Green, 2001; Bergström et al., 2013; Hu et al., 2015). It would also be interesting to investigate the effects of rehearsing a false alibi on memories in different populations, to assess if different individuals are more or less able to appear innocent. Since my research was based on university students, it should be confirmed that the results generalize to other populations of young adults who do not attend university and may differ in terms of cognitive abilities, motivations, etc. Furthermore, future research should investigate the effects of counterfactual imagination on crime memories in younger and older populations. Previous research suggested that it is more difficult for older adults to suppress unwanted true memories (Anderson, Reinholz, Kuhl, & Mayr, 2011), but also that older adults are more susceptible to false memories than young adults (Devitt & Schacter, 2016), including false memories arising from counterfactual imagination (Gerlach et al., 2014). Thus, older adults may be less likely to be able to conceal their true crime memories, but may also be less able to discriminate between true and imagined memories. Furthermore, this research can also be extended to investigate memory distortions in clinical populations that may have altered memory processing (e.g. depression, anxiety, and ADHD).

Extending on the previous point, future research should also consider investigating the effect of counterfactual imagination in the third-person perspective. Although research has suggested that imagining an event in first-person perspective is more vivid than imagining it in third-person perspective, it was also found that retrieving memory of the past from the perspective of one's own eyes is different from retrieving it as an observer, and this can effect subsequent memory (Jacques et al., 2017). It is possible that a criminal suspect might try to blame an innocent person for

their crime and that person might be falsely accused. In this scenario, the suspect might rehearse and imagine the other innocent person committing the crime from a third-person perspective, which might affect their memory of committing the act themselves. It would be interesting to know whether forensic memory detection is affected differently by the suspect imagining counterfactual events in first versus third person perspective, and whether counterfactually imagined events can be detected as true memories if they have been imagined from a third-person perspective. Also, it will be important to know whether imagining oneself or another person in an event is more effective for beating the memory detection test. If repeatedly imagining someone else committing a crime can be detected as a true memory in eyewitness testimony, then this casts serious doubts on whether memory detection can be used in forensic settings.

Related to the above point, future research should also investigate memory distortions and memory detection in witnesses of a crime. It is important to make sure that the eyewitness is telling the truth and remember the event accurately. However, it is known that eyewitness testimony is fallible and prone to errors (Helm, Ceci, & Burd, 2016; Loftus, 2003; Shaw & Porter, 2015). It might be possible that if the eyewitness is given a cue that remind them of the event, they might imagine various false details and those details might become confused with true details from the event, which could impair forensic memory detection. It would also be interesting to test whether forensic memory detection tests such as the aIAT and the CIT can detect memories accurately in eyewitnesses after they have been repeatedly exposed to false information of the crime via media (i.e. news).

Moreover, future research should consider using aIAT and the CIT to detect autobiographical memories from participant's real life. Autobiographical memories might be different from a new memory created in the lab because autobiographical memories are usually associated with rich specific knowledge of the event and emotion (Ben-Shakhar & Elaad, 2003; Meijer et al., 2014). Therefore, such memories might be more resistant to distortions from counterfactual imagination.

Future research could also consider building on the new paradigm I developed in Experiment 5, as there are many interesting novel strands of research one could investigate with this paradigm. For example, similar to the above suggestions, it would be interesting to use this paradigm in different populations to examine whether the same cognitive and neural mechanisms underlie counterfactual imagination effects on memory across different groups, such as different ages or different clinical conditions. Furthermore, this paradigm is not limited to the use of counterfactual imagination. It can be adapted to use memory suppression as a manipulation, by asking participants to suppress the associated action memories when presented with object cues. This could enable researchers to examine whether the effect of counterfactual imagination is similar to the effect of memory suppression (Bergstrom et al., 2013; Hu et al., 2015). According to the memory suppression literature, suppressing retrieval can also distort the original memory of an event. However, it is unknown whether counterfactual imagination and memory suppression might share similar neural and cognitive mechanisms. It would thus be interesting to compare the effects of counterfactual imagination and memory suppression directly in this paradigm.

6.5. Conclusions

The research presented in this thesis examined the effect of counterfactual imagination on memories of the past, both as a countermeasure against memory detection, but also in terms of the underlying neurocognitive mechanisms. In line with previous research, my findings show that memories are vulnerable and easy to distort, even when those memories relate to real life actions. Simply rehearsing and imagining a counterfactual version of a past event can distort how we remember that event and can make us appear as if we did not experience it. These findings thus illustrate the malleable nature of episodic memories, consistent with suggestions that memories are not objective records of the past but are subject to constant updating and distortions. These distortions may appear to be flaws, but are thought to arise as consequences of our effective and flexible memory systems (Schacter et al., 2011) that help us function and adapt to our changing environments.

References

Addis, D. R. (2018). Are episodic memories special? On the sameness of remembered and imagined event simulation. *Journal of the Royal Society of New Zealand*, *48*(2–3), 64–88. https://doi.org/10.1080/03036758.2018.1439071

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia*, *45*(7), 1363–1377. https://doi.org/10.1016/j.neuropsychologia.2006.10.016

Agosta, S., Castiello, U., Rigoni, D., Lionetti, S., & Sartori, G. (2011). The detection and the neural correlates of behavioral (prior) intentions. *Journal of Cognitive Neuroscience*, *23*(12), 3888–3902. https://doi.org/10.1162/jocn_a_00039

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., & Sartori, G. (2011). Detecting fakers of the autobiographical IAT. *Applied Cognitive Psychology*, *25*(2), 299–306. https://doi.org/10.1002/acp.1691

Agosta, S., Pezzoli, P., & Sartori, G. (2013). How to detect deception in everyday life and the reasons underlying It. *Applied Cognitive Psychology*, *27*(2), 256–262. https://doi.org/10.1002/acp.2902

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology*, *4*(AUG), 1–12. https://doi.org/10.3389/fpsyg.2013.00519

Allan, K., & Rugg, M. D. (1997). An event-related potential study of explicit memory on tests of cued recall and recognition. *Neuropsychologia*, *35*(4), 387–397. https://doi.org/10.1016/S0028-3932(96)00094-2

Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The Identification of Concealed

Memories Using the Event-Related Potential and Implicit Behavioral Measures: A Methodology for Prediction in the Face of Individual Differences. *Psychophysiology*, *29*(5), 504–522. https://doi.org/10.1111/j.1469-8986.1992.tb02024.x

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*(4), 415–445. https://doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin and Review*, *7*(3), 522–530. https://doi.org/10.3758/BF03214366

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering Can Cause Forgetting: Retrieval Dynamics in Long-Term Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, *410*(6826), 366–369. https://doi.org/10.1038/35066572

Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences*, *18*(6), 279–282. https://doi.org/10.1016/j.tics.2014.03.002

Anderson, M. C., & Levy, B. J. (2007). Theoretical issues in inhibition: Insights from research on human memory. In *Inhibition in cognition.* (pp. 81–102). Washington: American Psychological Association. https://doi.org/10.1037/11587-005

Anderson, M. C., & Neely, J. H. (1996). Interference and Inhibition in Memory

Retrieval. *Memory*, 237–313. https://doi.org/10.1016/B978-012102570-0/50010-0

Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., … Gabrieli, J. D. E. (2004). Neural Systems Underlying the Suppression of UnwantedMemories. *Science*, *303*(5655), 232–235. https://doi.org/10.1126/science.1089504

Anderson, M. C., Reinholz, J., Kuhl, B., & Mayr, U. (2011). Intentional Suppression of Unwanted Memories Grows More Difficult as We Age. *Psychology and Aging*, *26*(2), 397–405. https://doi.org/10.1037/a0022505

Ben-Shakhar, G. (1991). Clinical judgment and decision-making in CQT-polygraphy. *Integrative Physiological and Behavioral Science*, *26*(3), 232–240. https://doi.org/10.1007/BF02912515

Ben-Shakhar, G. (2002). A critical review of the Control Questions Test (CQT). In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 103–126). London: Academic Press.

Ben-Shakhar, G. (2012). Current research and potential applications of the concealed information test: An overview. *Frontiers in Psychology*, *3*(SEP), 1–11. https://doi.org/10.3389/fpsyg.2012.00342

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology*, *88*(1), 131–151. https://doi.org/10.1037/0021-9010.88.1.131

Benoit, R. G. G., & Anderson, M. C. (2012). Opposing Mechanisms Support the Voluntary Forgetting of Unwanted Memories. *Neuron*, *76*(2), 450–460. https://doi.org/10.1016/j.neuron.2012.07.025

Bergstrom, Z. M., Anderson, M. C., Buda, M., Simons, J. S., & Richardson-Klavehn, A. (2013). Intentional retrieval suppression can conceal guilty knowledge in ERP memory detection tests. *Biological Psychology*, *94*(1), 1–11. https://doi.org/10.1016/j.biopsycho.2013.04.012

Bergstrom, Z. M., de Fockert, J. W., & Richardson-Klavehn, A. (2009). ERP and behavioural evidence for direct suppression of unwanted memories. *NeuroImage*, *48*(4), 726–737. https://doi.org/10.1016/j.neuroimage.2009.06.051

Bergstrom, Z. M., Velmans, M., de Fockert, J. W., & Richardson-Klavehn, A. (2007). ERP evidence for successful voluntary avoidance of conscious recollection. *Brain Research*, *1151*(1), 119–133. https://doi.org/10.1016/j.brainres.2007.03.014

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

Brandt, V. C., Bergstrom, Z. M., Buda, M., Henson, R. N. A., & Simons, J. S. (2014). Did i turn off the gas? Reality monitoring of everyday actions. *Cognitive, Affective and Behavioral Neuroscience*, *14*(1), 209–219. https://doi.org/10.3758/s13415-013-0189-z

Bridger, E. K., Bader, R., Kriukova, O., Unger, K., & Mecklinger, A. (2012). The FN400 is functionally distinct from the N400. *Neuroimage*, *63*(3), 1334-1342. https://doi.org/10.1016/j.neuroimage.2012.07.047

Camp, G., Pecher, D., & Schmidt, H. G. (2007). No Retrieval-Induced Forgetting Using Item-Specific Independent Cues: Evidence Against a General Inhibitory Account. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(5), 950–958. https://doi.org/10.1037/0278-7393.33.5.950

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the Validity of the Guilty Knowledge Test From Simulated Experiments: The External Validity of Mock Crime Studies. *Journal of Experimental Psychology: Applied*, *9*(4), 261–269. https://doi.org/10.1037/1076-898X.9.4.261

Clancy, S. A., Schacter, D. L., McNally, R. J., & Pitman, R. K. (2000). False recognition in women reporting recovered memories of sexual abuse. *Psychological Science*, *11*(1), 26–31. https://doi.org/10.1111/1467-9280.00210

Curran, T. (2004). Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia*, *42*(8), 1088–1106. https://doi.org/10.1016/j.neuropsychologia.2003.12.011

Curran, T., L. Tepe, K., & Piatt, C. (2006). *Event-related potential explorations of dual processes in recognition memory*. (H. D. Zimmer, A. Mecklinger, & U. Lindenberger, Eds.), *Handbook of Binding and Memory: Perspectives from Cognitive Neuroscience*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198529675.003.0018

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test Is Statistically Detectable and Partly Correctable. *Basic and Applied Social Psychology*, *32*(4), 302–314. https://doi.org/10.1080/01973533.2010.519236

de Brigard, F. (2017). *Memory and imagination*. (S. Bernecker & K. Michaelian, Eds.), *The Routledge Handbook of Philosophy of Memory*. London: Routledge. https://doi.org/10.4324/9781315687315

de Brigard, F., Szpunar, K. K., & Schacter, D. L. (2013). Coming to Grips With the

Past: Effect of Repeated Simulation on the Perceived Plausibility of Episodic Counterfactual Thoughts. *Psychological Science*, *24*(7), 1329–1334. https://doi.org/10.1177/0956797612468163

Devitt, A. L., & Schacter, D. L. (2016). False memories with age: Neural and cognitive underpinnings. *Neuropsychologia*, *91*, 346–359. https://doi.org/10.1016/j.neuropsychologia.2016.08.030

Duarte, A., Ranganath, C., Winward, L., Hayward, D., & Knight, R. T. (2004). Dissociable neural correlates for familiarity and recollection during the encoding and retrieval of pictures. *Cognitive Brain Research*, *18*(3), 255–272. https://doi.org/10.1016/j.cogbrainres.2003.10.010

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Dzulkifli, M. A., Sharpe, H. L., & Wilding, E. L. (2004). Separating item-related electrophysiological indices of retrieval effort and retrieval orientation. *Brain and Cognition*, *55*(3), 433–443. https://doi.org/10.1016/j.bandc.2004.03.004

Engelkamp, J., & Zimmer, H. D. (1989). Memory for action events: A new field of research. *Psychological Research*, *51*(4), 153–157. https://doi.org/10.1007/BF00309142

Fabiani, M., Karis, D., & Donchin, E. (1986). P300 and Recall in an Incidental Memory Paradigm. *Psychophysiology*, *23*(3), 298–308. https://doi.org/10.1111/j.1469-8986.1986.tb00636.x

Farwell, L. A., & Donchin, E. (1991). The Truth Will Out: Interrogative Polygraphy

("Lie Detection") With Event-Related Brain Potentials. *Psychophysiology*, *28*(5), 531–547. https://doi.org/10.1111/j.1469-8986.1991.tb01990.x

Foerster, A., Wirth, R., Herbort, O., Kunde, W., & Pfister, R. (2017). Lying upside-down: Alibis reverse cognitive burdens of dishonesty. *Journal of Experimental Psychology: Applied*, *23*(3), 301–319. https://doi.org/10.1037/xap0000129

Gamer, M. (2011). Detecting of deception and concealed information using neuroimaging techniques. In *Memory detection: Theory and application of the Concealed Information Test.* (pp. 90–113). New York, NY, US: Cambridge University Press. https://doi.org/10.1017/CBO9780511975196.006

Gamer, M., & Berti, S. (2012). P300 amplitudes in the concealed information test are less affected by depth of processing than electrodermal responses. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00308

Gamer, M., Klimecki, O., Bauermann, T., Stoeter, P., & Vossel, G. (2012). fMRI-activation patterns in the detection of concealed information rely on memory-related effects. *Social Cognitive and Affective Neuroscience*, *7*(5), 506–515. https://doi.org/10.1093/scan/nsp005

Gamer, M., Kosiol, D., & Vossel, G. (2010). Strength of memory encoding affects physiological responses in the Guilty Actions Test. *Biological Psychology*, *83*(2), 101–107. https://doi.org/10.1016/j.biopsycho.2009.11.005

Gerlach, K. D., Dornblaser, D. W., & Schacter, D. L. (2014). Adaptive constructive processes and memory accuracy: Consequences of counterfactual simulations in young and older adults. *Memory*, *22*(1), 145–162. https://doi.org/10.1080/09658211.2013.779381

Goff, L. M., & Roediger, H. L. (1998). Imagination inflation for action events: Repeated imaginings lead to illusory recollections. *Memory and Cognition, 26*(1), 20–33. https://doi.org/10.3758/BF03211367

Gonsalves, B., & Paller, K. A. (2000). Neural events that underlie remembering something that never happened. *Nature Neuroscience, 3*(12), 1316–1321. https://doi.org/10.1038/81851

Gonsalves, B., & Paller, K. A. (2002). Mistaken memories: Remembering events that never happened. *Neuroscientist*, *8*(5), 391–395. https://doi.org/10.1177/107385802236964

Gonsalves, B., Reber, P. J., Gitelman, D. R., Parrish, T. B., Mesulam, M. M., & Paller, K. A. (2004). Neural evidence that vivid imagining can lead to false remembering. *Psychological Science, 15*(10), 655–660. https://doi.org/10.1111/j.0956-7976.2004.00736.x

Granhag, P. A., Vrij, A., & Verschuere, B. (2015). *Detecting deception: Current challenges and cognitive approaches*. John Wiley & Sons.

Gray, N. S., MacCulloch, M. J., Brown, A. S., Smith, J., & Snowden, R. J. (2005). An implicit test of the associations between children and sex in pedophiles. *Journal of Abnormal Psychology, 114*(2), 304–308. https://doi.org/10.1037/0021-843X.114.2.304

Gray, N. S., MacCulloch, M. J., Smith, J., Morris, M., & Snowden, R. J. (2003). Forensic psychology: Violence viewed by psychopathic murderers. *Nature, 423*(6939), 497–498. https://doi.org/10.1038/423497a

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual

differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Gregg, A. P. (2007). When vying reveals lying: the timed antagonistic response alethiometer. *Applied Cognitive Psychology*, *21*(5), 621–647. https://doi.org/10.1002/acp.1298

Gronau, N., Elber, L., Satran, S., Breska, A., & Ben-Shakhar, G. (2015). Retroactive memory interference: A potential countermeasure technique against psychophysiological knowledge detection methods. *Biological Psychology*, *106*, 68–78. https://doi.org/10.1016/j.biopsycho.2015.02.002

Hanslmayr, S., Leipold, P., Pastötter, B., & Bäuml, K.-H. (2009). Anticipatory Signatures of Voluntary Memory Suppression. *The Journal of Neuroscience*, *29*(9), 2742 LP-2747. https://doi.org/10.1523/JNEUROSCI.4703-08.2009

Hayama, H. R., Johnson, J. D., & Rugg, M. D. (2008). The relationship between the right frontal old/new ERP effect and post-retrieval monitoring: Specific or non-specific? *Neuropsychologia*, *46*(5), 1211–1223. https://doi.org/10.1016/j.neuropsychologia.2007.11.021

Hellerstedt, R., & Johansson, M. (2013). Electrophysiological correlates of competitor activation predict retrieval-induced forgetting. *Cerebral Cortex*, *24*(6), 1619–1629.

https://doi.org/10.1093/cercor/bht019

Helm, R. K., Ceci, S. J., & Burd, K. A. (2016). Can Implicit Associations Distinguish True and False Eyewitness Memory? Development and Preliminary Testing of the IATe. *Behavioral Sciences and the Law*, *34*(6), 803–819. https://doi.org/10.1002/bsl.2272

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin & Review*, *11*(1), 125–130. https://doi.org/10.3758/BF03206471

Hu, X., Bergstrom, Z. M., Bodenhausen, G. V., & Rosenfeld, J. P. (2015). Suppressing Unwanted Autobiographical Memories Reduces Their Automatic Influences: Evidence From Electrophysiology and an Implicit Autobiographical Memory Test. *Psychological Science*, *26*(7), 1098–1106. https://doi.org/10.1177/0956797615575734

Hu, X., & Rosenfeld, J. P. (2012). Combining the P300-complex trial-based Concealed Information Test and the reaction time-based autobiographical Implicit Association Test in concealed memory detection. *Psychophysiology*, *49*(8), 1090–1100. https://doi.org/10.1111/j.1469-8986.2012.01389.x

Hu, X., Rosenfeld, J. P., & Bodenhausen, G. V. (2012). Combating Automatic Autobiographical Associations. *Psychological Science*, *23*(10), 1079–1085. https://doi.org/10.1177/0956797612443834

Iacono, W. G., & Lykken, D. T. (2002). The scientific status of research on polygraph techniques: The case against polygraph tests. *Modern Scientific Evidence: The Law and Science of Expert Testimony*, *2*, 483–538.

Jacques, P. L. S., Carpenter, A. C., Szpunar, K. K., & Schacter, D. L. (2018). Remembering and imagining alternative versions of the personal past. *Neuropsychologia*, *110*, 170-179. https://doi.org/10.1016/j.neuropsychologia.2017.06.015

Jacques, P. L. S., Szpunar, K. K., & Schacter, D. L. (2017). Shifting Visual Perspective During Retrieval Shapes Autobiographical Memories. *NeuroImage*, *148*, 103–114. https://doi.org/10.1016/j.neuroimage.2016.12.028

Johansson, M., Aslan, A., Bauml, K.-H., Gabel, A., & Mecklinger, A. (2007). When Remembering Causes Forgetting: Electrophysiological Correlates of Retrieval-Induced Forgetting. *Cerebral Cortex*, *17*(6), 1335–1341. https://doi.org/10.1093/cercor/bhl044

Johansson, M., & Mecklinger, A. (2003). The late posterior negativity in ERP studies of episodic memory: action monitoring and retrieval of attribute conjunctions. *Biological psychology*, *64*(1-2), 91-117. https://doi.org/10.1016/S0301-0511(03)00104-2

Johnson, M. K. (1997). Source monitoring and memory distortion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1362), 1733–1745. https://doi.org/10.1098/rstb.1997.0156

Kensinger, E. A. (2009). Remembering the details: Effects of emotion. *Emotion Review*, *1*(2), 99–113. https://doi.org/10.1177/1754073908100432

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in cognitive sciences*, *4*(12), 463-470. https://doi.org/10.1016/S1364-6613(00)01560-6

Lanciano, T., Curci, A., Mastandrea, S., & Sartori, G. (2013). Do automatic mental associations detect a flashbulb memory? *Memory, 21*(4), 482–493. https://doi.org/10.1080/09658211.2012.740050

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs). *Psychophysiology*, *44*(6), 894–904. https://doi.org/10.1111/j.1469-8986.2007.00566.x

Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences*, *6*(7), 299–305. https://doi.org/10.1016/S1364-6613(02)01923-X

Levy, B. J., & Anderson, M. C. (2008). Individual differences in the suppression of unwanted memories: The executive deficit hypothesis. *Acta Psychologica*, *127*(3), 623–635. https://doi.org/10.1016/j.actpsy.2007.12.004

Leynes, P. A., Cairns, A., & Crawford, J. T. (2005). Event-Related Potentials Indicate That Reality Monitoring Differs from External Source Monitoring. *The American Journal of Psychology*, *118*(4), 497–524.

Loftus, E. F. (2003). Make-Believe Memories. *American Psychologist*, *58*(11), 867–873.

Loftus, E. F., & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals*. https://doi.org/10.1016/S0193-953X(05)70059-9

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique*. *The MIT Press* (2nd ed.). Cambridge, Massachusetts. https://doi.org/10.1118/1.4736938

Lui, M., & Rosenfeld, J. P. (2008). Detection of deception about multiple, concealed,

mock crime items, based on a spatial-temporal analysis of ERP amplitude and scalp distribution. *Psychophysiology*, *45*(5), 721–730. https://doi.org/10.1111/j.1469-8986.2008.00683.x

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385–388. https://doi.org/10.1037/h0046060

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology*, *44*(4), 258–262. https://doi.org/10.1037/h0044413

Lykken, D. T. (1988). Detection of Guilty Knowledge: A Comment on Forman and McCauley. *Journal of Applied Psychology*, *73*(2), 303–304. https://doi.org/10.1037/0021-9010.73.2.303

Lyle, K. B., & Johnson, M. K. (2006). Importing perceived features into false memories. *Memory*, *14*(2), 197–213. https://doi.org/10.1080/09658210544000060

Mecklinger, A., Rosburg, T., & Johansson, M. (2016). Reconstructing the past: The late posterior negativity (LPN) in episodic memory studies. *Neuroscience & Biobehavioral Reviews*, *68*, 621-638. https://doi.org/10.1016/j.neubiorev.2016.06.024

Marini, M., Agosta, S., Mazzoni, G., Barba, G. D., & Sartori, G. (2012). True and false DRM memories: Differences detected with an implicit task. *Frontiers in Psychology*, *3*(AUG), 1–7. https://doi.org/10.3389/fpsyg.2012.00310

Marini, M., Agosta, S., & Sartori, G. (2016). Electrophysiological Correlates of the Autobiographical Implicit Association Test (aIAT): Response Conflict and Conflict Resolution. *Frontiers in Human Neuroscience*, *10*(August), 1–9.

https://doi.org/10.3389/fnhum.2016.00391

Marsh, B. U., Pezdek, K., & Lam, S. T. (2014). Imagination perspective affects ratings of the likelihood of occurrence of autobiographical memories. *Acta Psychologica*, *150*(C), 114–119. https://doi.org/10.1016/j.actpsy.2014.05.006

Meijer, E., Selle, N. K., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*, *51*(9), 879–904. https://doi.org/10.1111/psyp.12239

Mertens, R., & Allen, J. J. B. (2008). The role of psychophysiology in forensic assessments : Deception detection , ERPs , and virtual reality mock crime scenarios, *45*, 286–298. https://doi.org/10.1111/j.1469-8986.2007.00615.x

Mitchell, K. J., & Johnson, M. K. (2009). Source Monitoring 15 Years Later: What Have We Learned From fMRI About the Neural Mechanisms of Source Memory? *Psychological Bulletin*, *135*(4), 638–677. https://doi.org/10.1037/a0015849

Nahari, G., & Ben-Shakhar, G. (2011). Psychophysiological and behavioral measures for detecting concealed information: The role of memory for crime details. *Psychophysiology*, *48*(6), 733–744. https://doi.org/10.1111/j.1469-8986.2010.01148.x

Nasman, V. T., Whalen, R., Cantwell, B., & Mazzeri, L. (1987). Late Vertex Positivity in Event-Related Potentials as a Guilty Knowledge Indicator: A New Method of Lie Detection AU - Rosenfeld, J. Peter. *International Journal of Neuroscience*, *34*(1–2), 125–129. https://doi.org/10.3109/00207458708985947

Osugi, A. (2011). Daily application of the Concealed Information Test: Japan. In

Memory detection: Theory and application of the Concealed Information Test (p. 253). Cambridge University Press.

Otgaar, H., & Baker, A. (2018). When lying changes memory for the truth. *Memory, 26*(1), 2–14. https://doi.org/10.1080/09658211.2017.1340286

Otten, L. J., Quayle, A. H., Akram, S., Ditewig, T. A., & Rugg, M. D. (2006). Brain activity before an event predicts later recollection. *Nature Neuroscience*, *9*(4), 489–491. https://doi.org/10.1038/nn1663

Paller, K. A. (1990). Recall and stem-completion priming have different electrophysiological correlates and are modified differentially by directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1021–1032. https://doi.org/10.1037/0278-7393.16.6.1021

Paller, K. A., Kutas, M., & Mayes, A. R. (1987). Neural correlates of encoding in an incidental learning paradigm. *Electroencephalography and Clinical Neurophysiology*, *67*(4), 360–371. https://doi.org/10.1016/0013-4694(87)90124-6

Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, *6*(2), 93–102. https://doi.org/https://doi.org/10.1016/S1364-6613(00)01845-3

Petrocelli, J. V., & Crysel, L. C. (2009). Counterfactual thinking and confidence in blackjack: A test of the counterfactual inflation hypothesis. *Journal of Experimental Social Psychology*, *45*(6), 1312–1315. https://doi.org/10.1016/j.jesp.2009.08.004

Petrocelli, J. V., Rubin, A. L., & Stevens, R. L. (2016). The Sin of Prediction: When Mentally Simulated Alternatives Compete With Reality. *Personality and Social*

Psychology Bulletin, *42*(12), 1635–1652. https://doi.org/10.1177/0146167216669122

Petrocelli, J. V, & Harris, A. K. (2011). Learning Inhibition in the Monty Hall Problem: The Role of Dysfunctional Counterfactual Prescriptions. *Personality and Social Psychology Bulletin*, *37*(10), 1297–1311. https://doi.org/10.1177/0146167211410245

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, *118*(10), 2128-2148. https://doi.org/10.1016/j.clinph.2007.04.019

Ranganath, C. (2004). The 3-D prefrontal cortex: Hemispheric asymmetries in prefrontal activity and their relation to memory retrieval processes. *Journal of Cognitive Neuroscience*, *16*(6), 903–907. https://doi.org/10.1162/0898929041502625

Rosburg, T., Mecklinger, A., & Johansson, M. (2011). Electrophysiological correlates of retrieval orientation in reality monitoring. *NeuroImage*, *54*(4), 3076–3084. https://doi.org/10.1016/j.neuroimage.2010.10.068

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J.-H. (1991). An ERP-Based, Control-Question Lie Detector Analog: Algorithms for Discriminating Effects Within Individuals' Average Waveforms. *Psychophysiology*, *28*(3), 319–335. https://doi.org/10.1111/j.1469-8986.1991.tb02202.x

Rosenfeld, J. P., Cantwell, B., Nasman, V. T., Wojdac, V., Ivanov, S., & Mazzeri, L. (1988). A Modified, Event-Related Potential-Based Guilty Knowledge Test. *International Journal of Neuroscience*, *42*(1–2), 157–161. https://doi.org/10.3109/00207458808985770

Rosenfeld, J. P., Hu, X., Labkovsky, E., Meixner, J., & Winograd, M. R. (2013). Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information. *International Journal of Psychophysiology*, *90*(2), 118–134. https://doi.org/10.1016/j.ijpsycho.2013.08.012

Rosenfeld, J. P., Labkovsky, E., Winograd, M. R., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology*, *45*(6), 906–919. https://doi.org/10.1111/j.1469-8986.2008.00708.x

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, *41*(2), 205–219. https://doi.org/10.1111/j.1469-8986.2004.00158.x

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, *28*(6), 907–922. https://doi.org/10.3758/BF03209339

Rugg, M. D. (1995). Event-related potential studies of human memory. In *The cognitive neurosciences.* (pp. 789–801). Cambridge, MA, US: The MIT Press.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, *11*(6), 251–257. https://doi.org/10.1016/j.tics.2007.04.004

Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., & Lindsley, D. B. (1980). Electrocortical Signs of Levels of Processing: Perceptual Analysis and Recognition

Memory. *Psychophysiology*, *17*(6), 568–576. https://doi.org/10.1111/j.1469-8986.1980.tb02299.x

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, D., Castiello, U., & Castielo, C. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*(8), 772–780. https://doi.org/10.1111/j.1467-9280.2008.02156.x

Schacter, D. L. (1999). The Seven Sins of Memory: Insights From Psychology and Cognitive Neuroscience. *American Psychologist*, *54*(3), 182–203. https://doi.org/10.1037//0003-066X.54.3.182

Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *The American Psychologist*, *67*(8), 603–613. https://doi.org/10.1037/a0029869

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, *15*(10), 467–474. https://doi.org/10.1016/j.tics.2011.08.004

Schacter, D. L., & Slotnick, S. D. (2004). The cognitive neuroscience of memory distortion. *Neuron*, *44*(1), 149–160. https://doi.org/10.1016/j.neuron.2004.08.017

Shaw, J., & Porter, S. (2015). Constructing Rich False Memories of Committing Crime. *Psychological Science*, *26*(3), 291–301. https://doi.org/10.1177/0956797614562862

Shidlovski, D., Schul, Y., & Mayo, R. (2014). If I imagine it, then it happened: The Implicit Truth Value of imaginary representations. *Chemical Geology*, *387*(1), 517–529. https://doi.org/10.1016/j.cognition.2014.08.005

Sirgiovanni, E., Corbellini, G., & Caporale, C. (2016). A recap on Italian neurolaw: epistemological and ethical issues. *Mind and Society*, *4*(519), 1–19.

https://doi.org/10.1007/s11299-016-0188-1

Soskins, M., Rosenfeld, J.P., & Niendam, T. (2001). The case for peak to-peak measurement of P300 recorded at .3 hz high pass filter settings in detection of deception. *Int. J. Psychophysiology, 40,* 173180. https://doi.org/10.1016/S0167-8760(00)00154-9

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory and Cognition*, *40*(6), 827–843. https://doi.org/10.3758/s13421-012-0211-7

Suengas, A. G., & Johnson, M. K. (1988). Qualitative Effects of Rehearsal on Memories for Perceived and Imagined Complex Events. *Journal of Experimental Psychology: General*, *117*(4), 377–389. https://doi.org/10.1037/0096-3445.117.4.377

Takarangi, M., Strange, D., & Houghton, E. (2015). Event familiarity influences memory detection using the aIAT. *Memory*, *23*(3), 453–461. https://doi.org/10.1080/09658211.2014.902467

Takarangi, M., Strange, D., Shortland, A., & James, H. (2013). Source confusion influences the effectiveness of the autobiographical IAT. *Psychonomic Bulletin and Review*, *20*(6), 1232–1238. https://doi.org/10.3758/s13423-013-0430-3

Teachman, B. A., Gregg, A. P., & Woody, S. R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology*, *110*(2), 226–235. https://doi.org/10.1037/0021-843X.110.2.226

Van Hooff, J. C., Brunia, C. H. M., & Allen, J. J. B. (1996). Event-related potentials as indirect measures of recognition memory. *International Journal of*

Psychophysiology, *21*(1), 15–31. https://doi.org/10.1016/0167-8760(95)00043-7

Vargo, E. J., & Petróczi, A. (2013). Detecting cocaine use? The autobiographical implicit association test (aIAT) produces false positives in a real-world setting. *Substance Abuse: Treatment, Prevention, and Policy*, *8*(1), 1–13. https://doi.org/10.1186/1747-597X-8-22

Verde, M. F., & Rotello, C. M. (2004). Strong memories obscure weak memories in associative recognition. *Psychonomic Bulletin & Review*, *11*(6), 1062–1066. https://doi.org/10.3758/BF03196737

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (2011). *Memory detection: Theory and application of the Concealed Information Test*. Cambridge University Press.

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition*, *4*(1), 59–65. https://doi.org/https://doi.org/10.1016/j.jarmac.2015.01.001

Verschuere, B., Prati, V., & Houwer, J. De. (2009). Cheating the lie detector: Faking in the autobiographical implicit association test: Research Report. *Psychological Science*, *20*(4), 410–413. https://doi.org/10.1111/j.1467-9280.2009.02308.x

Voss, J. L., & Paller, K. A. (2017). Neural Substrates of Remembering: Event-Related Potential Studies. In *Learning and Memory: A Comprehensive Reference* (Third Edit, pp. 81–98). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21070-5

Wilding, E. L., & Ranganath, C. (2011). Electrophysiological Correlates of Episodic Memory Processes. In S. J. Luck & E. Kappenman (Eds.), *The oxford handbook of ERP components* (pp. 373–396). Oxford: Oxford University Press.

Winograd, M. R., & Rosenfeld, J. P. (2011). Mock crime application of the Complex Trial Protocol (CTP) P300-based concealed information test. *Psychophysiology*, *48*(2), 155–161. https://doi.org/10.1111/j.1469-8986.2010.01054.x

Appendix

Appendix A: Post-Questionnaire for Innocent Group (Experiment 2)

Lab Crime task

1. How nervous were you during the lab act (i.e. while writing email on a piece of paper)?

Not nervous at all Somewhat nervous Extremely nervous

0 1 2 3 4 5 6

Sentence classification task

2. Please rate how often you were thinking about the lab crime that you conducted (i.e. stealing the ring) whilst taking part in the sentence classification task?

Not at all Occasionally All the time

0 1 2 3 4 5 6

3. The purpose of the sentence classification task was to detect if you were guilty of stealing the ring. Please rate your motivation to beat the test and appear innocent.

Not motivated at all Somewhat motivated Extremely motivated

0 1 2 3 4 5 6

In the sentence classification test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Appendix B: Post-Questionnaire for Guilty-Standard Group (Experiment 2)

Lab Crime task

1. How nervous were you during the lab crime (i.e. while stealing the ring)?

Not nervous at all			Somewhat nervous			Extremely nervous
0	1	2	3	4	5	6

Sentence classification task

2. Please rate how often you were thinking about the lab crime that you conducted (i.e. stealing the ring) whilst taking part in the sentence classification task?

Not at all			Occasionally			All the time
0	1	2	3	4	5	

3. The purpose of the sentence classification task was to detect if you were guilty of stealing the ring. Please rate your motivation to beat the test and appear innocent.

Not motivated at all			Somewhat motivated			Extremely motivated
0	1	2	3	4	5	6

In the sentence classification test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Appendix C: Post-Questionnaire for Guilty-Alibi Group (Experiment 2)

Lab Crime task

1. How nervous were you during the lab crime (i.e. while stealing the ring)?

Not nervous at all		Somewhat nervous			Extremely nervous	
0	1	2	3	4	5	6

Alibi imagination task

2. Please rate how vividly you imagined the alibi event.

Not vivid at all		Somewhat vivid			Extremely vivid	
0	1	2	3	4	5	6

Sentence classification task

3. Please rate how often you were thinking about the lab crime that you conducted (i.e. stealing the ring) whilst taking part in the sentence classification task?

Not at all			Occasionally			All the time
0	1	2	3	4	5	6

4. Please rate how often you were thinking about the alibi scenario that you imagined (i.e. writing your email) whilst taking part in the sentence classification task?

Not at all			Occasionally			All the time
0	1	2	3	4	5	6

5. The purpose of the sentence classification task was to detect if you were guilty of stealing the ring, but we asked you to try to appear innocent. Please rate your motivation to beat the test and appear innocent.

Not motivated at all		Somewhat motivated			Extremely motivated	
0	1	2	3	4	5	6

In the sentence classification test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Appendix D: Post-Questionnaire for Innocent Group (Experiment 3)

Post-Questionnaire

Questions about the simulated real-life act

In the first part of this study, we asked you to conduct a simulated real-life act. Please answer the following questions about this event.

1. What part of the building did you go to during the act? _____

2. Where did you find the notice to write your email address? _____

3. What colour was the envelope? _____

4. What colour was the paper? _____

5. In how much detail can you remember conducting the act?

Few details						Many details
0	1	2	3	4	5	6

6. How vivid is your memory for the act you conducted?

Not vivid at all			Somewhat vivid			Extremely vivid
0	1	2	3	4	5	6

7. How nervous you were during the act?

Not nervous at all			Somewhat nervous			Extremely nervous
0	1	2	3	4	5	6

Questions about the sentence classification task

In the second part of this study, we asked you to classify sentences on a computer by pressing different buttons. Please answer the following questions about this task.

8. The purpose of the sentence classification task was to detect if you were guilty of a (mock) crime. Please give a rating on your motivation to beat the test (i.e. prove that you are innocent).

Not motivated at all			Somewhat motivated			Extremely motivated
0	1	2	3	4	5	6

9. Please rate how often you were thinking about the act that you conducted (i.e. writing your email address) whilst taking part in the sentence classification task

Not at all			Occasionally			All the time
0	1	2	3	4	5	6

In the sentence classification test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Appendix E: Post-Questionnaire for Guilty-Standard Group (Experiment 3)

Questions about the mock crime

In the first part of this study, we asked you to conduct a mock crime (theft). Please answer the following questions about this event.

1. What room did you go to during the mock crime? _____

2. In the room, where did you find the bag? _____

3. What colour and type of bag was it? _____

4. What object was inside it? _____

5. In how much detail can you remember conducting the mock crime?

Few details						Many details
0	1	2	3	4	5	6

6. How vivid is your memory for the mock crime?

Not vivid at all			Somewhat vivid			Extremely vivid
0	1	2	3	4	5	6

7. How nervous you were during the mock crime?

Not nervous at all			Somewhat nervous			Extremely nervous
0	1	2	3	4	5	6

Questions about the sentence classification task

In the second part of this study, we asked you to classify sentences on a computer by pressing different buttons. Please answer the following questions about this task.

8. The purpose of the sentence classification task was to detect if you were guilty of the mock crime. Please rate your motivation to beat the test and appear innocent.

Not motivated at all			Somewhat motivated			Extremely motivated
0	1	2	3	4	5	6

9. Please rate how often you were thinking about the mock crime that you conducted whilst taking part in the sentence classification task

Not at all			Occasionally			All the time
0	1	2	3	4	5	6

In the sentence classification test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Appendix F: Post-Questionnaire for Guilty-Alibi Group (Experiment 3)

Questions about the mock crime

In the first part of this study, we asked you to conduct a mock crime (theft). Please answer the following questions about this event.

1. What room did you go to during the mock crime? _____

2. In the room, where did you find the bag? _____

3. What colour and type of bag was it? _____

4. What object was inside it? _____

5. In how much detail can you remember conducting the mock crime?

Few details						Many details
0	1	2	3	4	5	6

6. How vivid is your memory for the mock crime?

Not vivid at all			Somewhat vivid			Extremely vivid
0	1	2	3	4	5	6

7. How nervous you were during the mock crime?

Not nervous at all			Somewhat nervous			Extremely nervous
0	1	2	3	4	5	6

Questions about the alibi

After the mock crime, we asked you to rehearse and imagine a false alibi scenario to appear innocent. Please answer the following questions about this alibi.

8. What part of the building did you go to, according to the alibi? _____

9. Where did you find the notice to write your email address, according to the alibi? _____

10. What colour was the envelope in your imagined alibi? _____

11. What colour was the paper in your imagined alibi? _____

12. In how much detail can you imagine the alibi scenario?

Few details						Many details
0	1	2	3	4	5	6

13. How vividly can you imagine the alibi scenario?

Not vivid at all			Somewhat vivid			Extremely vivid
0	1	2	3	4	5	6

Questions about the sentence classification task

In the last part of this study, we asked you to classify sentences on a computer by pressing different buttons. Please answer the following questions about this task.

14. The purpose of the sentence classification task was to detect if you were guilty of the mock crime. Please rate your motivation to beat the test and appear innocent.

Not motivated at all			Somewhat motivated			Extremely motivated
0	1	2	3	4	5	6

15. Please rate how often you were thinking about the mock crime that you conducted whilst taking part in the sentence classification task

Not at all			Occasionally			All the time
0	1	2	3	4	5	6

16. Please rate how often you were thinking about the alibi scenario whilst taking part in the sentence classification task

Not at all			Occasionally			All the time
0	1	2	3	4	5	6

In the sentence classification test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Appendix G: Post-Questionnaire for Guilty-Immediate and Guilty-Delay with HT Group (Experiment 4)

Thank you for participating. The next questions are very important for us to analyse the experiment correctly. Please answer these questions as honest as possible, and please note that your responses are completely anonymous and will not affect your credits.

1. Please rate the extent to which the crime-relevant memories came to mind automatically (i.e. easily, without effort) upon seeing the stolen item?

Not automatic at all						Extremely motivated
			Not sure			
0	1	2	3	4	5	6

2. Please give a rating on your motivation to beat the test (i.e. prove that you are innocent).

Not motivated at all						Extremely motivated
			Somewhat motivated			
0	1	2	3	4	5	6

3. How nervous you were during the lab crime?

Not nervous at all						Extremely nervous
			Somewhat nervous			
0	1	2	3	4	5	6

4. Did you try any strategy during the brainwave test aiming to beat the test?

Yes No

If yes, please list the strategies you use during the brainwave test. If you used more than one strategies, please list all of them.

5. If you used any strategies during the brainwave test aiming to beat it, please rate your confidence level in beating the test:

Not confident at all						Extremely confident
			Not sure			
0	1	2	3	4	5	6

6. In the behavioural test that follows the brainwave test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Appendix H: Post-Questionnaire for Guilty-Alibi with HT Group (Experiment 4)

Thank you for participating. The next questions are very important for us to analyse the experiment correctly. Please answer these questions as honest as possible, and please note that your responses are completely anonymous and will not affect your credits.

Questions about the brainwave test, mock crime, and behavioural test

1. Please rate the extent to which the crime-relevant memories came to mind automatically (i.e. easily, without effort) upon seeing the stolen item?

Not automatic at all			Not sure			Extremely motivated
0	1	2	3	4	5	6

2. Please give a rating on your motivation to beat the test (i.e. prove that you are innocent).

Not motivated at all			Somewhat motivated			Extremely motivated
0	1	2	3	4	5	6

3. How nervous you were during the lab crime?

Not nervous at all			Somewhat nervous			Extremely nervous
0	1	2	3	4	5	6

4. Did you try any strategy during the brainwave test aiming to beat the test?

Yes No

If yes, please list the strategies you use during the brainwave test. If you used more than one strategies, please list all of them.

5. If you used any strategies during the brainwave test aiming to beat it, please rate your confidence level in beating the test:

Not confident at all			Not sure			Extremely confident
0	1	2	3	4	5	6

6. In the behavioural test that follows the brainwave test, did you try any strategies to distort the test? For instance, did you deliberately speed up or slow down your responses times to the sentences? If yes, please list these strategies:

Questions about the alibi

After the mock crime, we asked you to rehearse and imagine an alibi scenario to appear innocent, involving writing your email address. Please answer the following questions about this alibi.

7. In how much detail can you imagine the alibi scenario?

Few details						Many details
0	1	2	3	4	5	6

8. How vividly can you imagine the alibi scenario?

Not vivid at all			Somewhat vivid			Extremely vivid
0	1	2	3	4	5	6

9. How much do you believe that the alibi scenario really occurred in the way you remember it? That is, do you believe you really conducted the act involving writing your email address?

Not at all			Somewhat true			Absolutely True
0	1	2	3	4	5	6

Appendix I: Post-Questionnaire (Experiment 5)

Thank you for participating. Now we want to ask you some questions about your experience of the tasks today. Sometimes people fail to follow instructions for various reasons, for example because they suspect there is something else going on in the study that they are not told about, or they may have misunderstood the instructions. It's very important for us to identify if this happened in the tasks today, because otherwise our experiment results might not be valid. **Please answer these questions as honestly as possible, and please note that your responses will not affect your payment or credits.**

1. When you were trying to imagine the actions today, to what extent were you *intentionally* (on purpose) thinking about the actions you really performed in the first session, even though you were asked not to?

Not at all			Sometimes			Always
0	1	2	3	4	5	6

2. When you were rating the vividness of your imagination, to what extent were you able to make these judgements correctly and report them using the rating scale?

Never			Sometimes			Always
0	1	2	3	4	5	6

3. When you were rating the vividness of your imagination, how often did you *intentionally* (on purpose) rate your imagination as more or less vivid than it actually was?

Never			Sometimes			Always
0	1	2	3	4	5	6

4. If you answered a score higher than 3 to any of the above questions, please explain in more detail here (continue overleaf if necessary):

_____

_____

_____

_____

_____

5. What do you think the study is about (continue overleaf if necessary)?

_____

_____

_____